

Philipps



Universität  
Marburg

# Dissection of Complex Genetic Correlations into Interaction Effects

– DISSERTATION –

zur Erlangung des

**DOKTORGRADES DER NATURWISSENSCHAFTEN**

(Dr. rer. nat.)

dem Fachbereich

**PHYSIK**

der Philipps-Universität Marburg

vorgelegt von

Diplom-Physiker

und Diplom-Informatiker

**MICHAEL GRAU**

aus Marburg.

Marburg (Lahn),

im Mai 2015.

Veröffentlicht in Marburg, 2016.

Vom Fachbereich Physik der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen am 20.07.2015.

Erstgutachter: Prof. Dr. Peter Lenz<sup>1</sup>  
Zweitgutachter: Prof. Dr. med. Georg Lenz<sup>2</sup>  
Drittgutachter: Prof. Dr. Eyke Hüllermeier<sup>3</sup>

Tag der Disputation: 03.08.2015

<sup>1</sup>Philipps-Universität Marburg, Fachbereich Physik, Komplexe Systeme; <sup>2</sup>Westfälische Wilhelms-Universität Münster, Medizinische Fakultät, Translationale Onkologie; <sup>3</sup>Universität Paderborn, Institut für Informatik, Intelligente Systeme.

Published in Marburg, 2016.

Accepted as dissertation by the Department of Physics at the University of Marburg (University ID 1180) on 07/20/2015.

Primary assessor: Prof. Dr. Peter Lenz<sup>1</sup>  
Secondary assessor: Prof. Dr. med. Georg Lenz<sup>2</sup>  
Tertiary assessor: Prof. Dr. Eyke Hüllermeier<sup>3</sup>

Day of thesis defense: 08/03/2015

<sup>1</sup>University of Marburg, Department of Physics, Complex Systems; <sup>2</sup>University of Münster, Faculty of Medicine, Translational Oncology; <sup>3</sup>University of Paderborn, Department of Computer Science, Intelligent Systems.

*I dedicate my dissertation to  
my parents as thank-you for their confidence  
and support throughout the course of my  
studies at all times and for the kind-hearted  
and loving people they are.*



# Abstract

---

Living systems are overwhelmingly complex and consist of many interacting parts. Already the quantitative characterization of a single human cell type on genetic level requires at least the measurement of 20000 gene expressions. It remains a big challenge for theoretical approaches to discover patterns in these signals that represent specific interactions in such systems. A major problem is that available standard procedures summarize gene expressions in a hard-to-interpret way. For example, principal components represent axes of maximal variance in the gene vector space and thus often correspond to a superposition of multiple different gene regulation effects<sup>(e.g. I.1.4)</sup>.

Here, a novel approach to analyze and interpret such complex data is developed<sup>(Chapter II)</sup>. It is based on an extremum principle that identifies an axis in the gene vector space to which as many as possible samples are correlated as highly as possible<sup>(II.3)</sup>. This axis is maximally specific and thus most probably corresponds to exactly one gene regulation effect, making it considerably easier to interpret than principle components. To stabilize and optimize effect discovery, axes in the sample vector space are identified simultaneously. Genes and samples are always handled symmetrically by the algorithm. While sufficient for effect discovery, effect axes can only linearly approximate regulation laws. To represent a broader class of nonlinear regulations, including saturation effects or activity thresholds<sup>(e.g. II.1.1.2)</sup>, a bimonotonic effect model is defined<sup>(II.2.1.2)</sup>. A corresponding regression is realized that is monotonic over projections of samples (or genes) onto discovered gene (or sample) axes. Resulting effect curves can approximate regulation laws precisely<sup>(II.4.1)</sup>. This enables the dissection of exclusively the discovered effect from the signal<sup>(II.4.2)</sup>. Signal parts from other potentially overlapping effects remain untouched. This continues iteratively. In this way, the high-dimensional initial signal<sup>(II.2.1.1)</sup> can be dissected into highly specific effects.

Method validation demonstrates that superposed effects of various size, shape and signal strength can be dissected reliably<sup>(II.6.2)</sup>. Simulated laws of regulation are reconstructed with high correlation. Detection limits, e.g. for signal strength or for missing values, lie above practical requirements<sup>(II.6.4)</sup>. The novel approach is systematically compared with standard procedures such as principal component analysis. Signal dissection is shown to have clear advantages, especially for many overlapping effects of comparable size<sup>(II.6.3)</sup>.

An ideal test field for such approaches is cancer cells, as they may be driven by multiple overlapping gene regulation networks that are largely unknown. Additionally, quantification and classification of cancer cells by their particular set of driving gene regulations is a prerequisite towards precision medicine. To validate the novel method against real biological data, it is applied to gene expressions of over 1000 tumor samples from Diffuse Large B-Cell Lymphoma (DLBCL) patients<sup>(Chapter III)</sup>. Two already known subtypes of this disease<sup>(cf. I.1.2.1)</sup> with significantly different survival following the same chemotherapy were originally also discovered as a gene expression effect. These subtypes can only be precisely determined by this effect on molecular level. Such previous results offer a possibility for method validation and indeed, this effect has been unsupervisedly rediscovered<sup>(III.3.2.2)</sup>.

Several additional biologically relevant effects have been discovered and validated across four patient cohorts. Multivariate analyses<sup>(III.2)</sup> identify combinations of validated effects that can predict significant differences in patient survival. One novel effect possesses an even higher predictive value<sup>(cf. III.2.5.1)</sup> than the rediscovered subtype effect and is genetically more specific<sup>(cf. III.3.3.1)</sup>. A trained and validated Cox survival model<sup>(III.2.5)</sup> can predict significant survival differences *within* known DLBCL subtypes<sup>(III.2.5.6)</sup>, demonstrating that they are genetically heterogeneous as well. Detailed biostatistical evaluations of all survival effects<sup>(III.3.3)</sup> may help to clarify the molecular pathogenesis of DLBCL.

Furthermore, the applicability of signal dissection is not limited to biological data. For instance, dissecting spectral energy distributions of stars observed in astrophysics might be useful to discover laws of light emission.



# Kurzzusammenfassung

---

Lebende Systeme sind überwältigend komplex und bestehen aus vielen interagierenden Teilen. Bereits die quantitative Charakterisierung eines einzelnen menschlichen Zelltyps auf genetischer Ebene bedarf mindestens der Messung von 20000 Genexpressionen. Es ist nach wie vor eine große Herausforderung für theoretische Ansätze, Muster in diesen Signalen zu entdecken, welche spezifische Interaktionen in solchen Systemen repräsentieren. Ein Hauptproblem besteht darin, dass verfügbare Standardmethoden die Genexpressionen in einer schwierig zu interpretierenden Weise zusammenfassen. Hauptkomponenten etwa repräsentieren Achsen maximaler Varianz im Genvektorraum und entsprechen daher häufig einer Überlagerung vieler verschiedener Genregulationseffekte<sup>(e.g. I.1.4)</sup>.

In dieser Arbeit wird ein neuartiger Ansatz zur Analyse und Interpretation derartig komplexer Daten entwickelt<sup>(Chapter II)</sup>. Er basiert auf einem Extremalprinzip, welches eine Achse im Genvektorraum identifiziert, zu der so viele Proben wie möglich so hoch wie möglich korreliert sind<sup>(II.3)</sup>. Diese Achse ist maximal spezifisch und entspricht daher typischerweise genau einem Genregulationseffekt, wodurch sie deutlich einfacher zu interpretieren ist als Hauptkomponenten. Zur Stabilisierung und zur Optimierung der Effekterkennung werden analog und gleichzeitig Achsen im Probenvektorraum identifiziert. Der Algorithmus behandelt generell Gene und Proben symmetrisch. Obwohl sie ausreichend zur Entdeckung von Effekten sind, können Effektachsen Regulationsgesetze nur linear annähern. Um eine breitere Klasse nichtlinearer Regulationen wie Sättigungseffekte oder Aktivitätsschwellen zu repräsentieren, wird ein bimonotonisches Effektmodell definiert<sup>(II.2.1.2)</sup>. Die entsprechende Regression ist monotonisch über die Projektionen von Proben (bzw. Genen) auf entdeckte Genachsen (bzw. Probenachsen). Resultierende Effektkurven können Genregulationsgesetze präzise approximieren<sup>(II.4.1)</sup>. Das ermöglicht die ausschließliche Abtrennung des entdeckten Effekts vom Signal<sup>(II.4.2)</sup>. Signalbestandteile anderer (möglicherweise überlappender) Effekte bleiben unangetastet. Dies wird iterativ fortgesetzt. Auf diese Weise kann das hochdimensionale Ausgangssignal<sup>(II.2.1.1)</sup> in hochspezifische Einzeleffekte zerlegt werden.

Die Methodvalidierung zeigt, dass superponierte Effekte von vielfältiger Größe, Form und Signalstärke zuverlässig zerlegt werden können<sup>(II.6.2)</sup>. Simulierte Regulationsgesetze werden dabei mit hoher Korrelation rekonstruiert. Erkennungsgrenzen bzgl. Signalstärke oder bzgl. der Rate fehlender Messwerte<sup>(II.6.4)</sup> liegen oberhalb praktischer Anforderungen. Der neuartige Ansatz wird mit Standardverfahren wie der Hauptkomponentenanalyse systematisch verglichen. Es wird gezeigt, dass die Signalzerlegung klare Vorteile hat, insbesondere bei vielen überlappenden Effekten mit vergleichbarer Größe<sup>(II.6.3)</sup>.

Ein ideales Testfeld für solche Ansätze sind Krebszellen, da sie von vielen überlappenden Genregulationsnetzwerken gesteuert sein können, welche weitgehend unbekannt sind. Darüber hinaus ist die Quantifizierung und Klassifizierung von Krebszellen durch ihre spezifische Menge antreibender Genregulationen eine Voraussetzung in Richtung Präzisionsmedizin. Um die neuartige Methode gegen reale Daten zu validieren, wird sie auf Genexpressionen von über 1000 Tumorproben von Patienten mit diffus großzelligem B-Zell-Lymphom (DLBCL) angewendet. Zwei bereits bekannte Subtypen dieser Krankheit<sup>(cf. I.1.2.1)</sup> mit signifikant unterschiedlichem Überleben infolge derselben Chemotherapie wurden ursprünglich ebenfalls als Genexpressionseffekt entdeckt. Diese Subtypen können nur mit Hilfe dieses Effekts auf molekularem Level präzise bestimmt werden. Solche vorherigen Ergebnisse erlauben eine Methodvalidierung, und in der Tat wurde dieser Effekt unüberwacht wiederentdeckt<sup>(III.3.2.2)</sup>.

Mehrere weitere biologisch relevante Ergebnisse wurden ermittelt und über vier Patientenkohorten validiert. Multivariate Analysen<sup>(III.2)</sup> identifizieren Kombinationen von validierten Effekten, die signifikante Unterschiede im Patientenüberleben vorhersagen können. Ein neuartiger Effekt besitzt sogar einen höheren Vorhersagewert<sup>(cf. III.2.5.1)</sup> als der wiederentdeckte Subtypeneffekt und ist zudem genetisch spezifischer<sup>(cf. III.3.3.1)</sup>.

Ein angelerntes und validiertes Cox-Überlebensmodell<sup>(III.2.5)</sup> kann signifikante Überlebensunterschiede *innerhalb* bekannter DLBCL Subtypen<sup>(III.2.5.6)</sup> vorhersagen, was zeigt, dass diese ebenfalls genetisch heterogen sind. Detaillierte biostatistische Auswertungen für alle Überlebenseffekte<sup>(III.3.3)</sup> können dazu beitragen, die molekulare Pathogenese von DLBCL zu klären.

Darüber hinaus ist die Anwendbarkeit der Signalzerlegung nicht auf biologische Daten begrenzt. In der Astrophysik könnte z.B. die Zerlegung spektraler Energieverteilungen von Sternen nützlich zur Entdeckung von Lichtemissionsgesetzen sein.



# Contents

|  |           |   |           |
|--|-----------|---|-----------|
| <b>Chapter I From Cells to Knowledge</b>   | <b>1</b>  | <i>II.1.3.2</i> Cohort-independent genomic consensus effects .....                  | 45        |
| <i>I.1</i> Introduction.....   | 3         | <i>II.1.3.3</i> Patient classification in new cohorts by genomic effects .....      | 45        |
| <i>I.1.1</i> What is a complex system?.....  | 3         | <i>II.1.3.4</i> Specific biostatistical evaluation of discovered effects.....       | 45        |
| <i>I.1.2</i> Biological Context.....   | 4         | <b>II.2 Mathematical Framework</b> .....  | <b>46</b> |
| <i>I.1.2.1</i> Diffuse Large B-Cell Lymphoma and cells as complex systems.....                       | 4         | <i>II.2.1</i> Model.....  | 46        |
| <i>I.1.2.2</i> Pathways, normal B cells in the immune system and NF- $\kappa$ B.....                 | 5         | <i>II.2.1.1</i> Signal model .....  | 46        |
| <i>I.1.2.3</i> Molecular causes of B cell malignancy.....  | 6         | <i>II.2.1.2</i> Bimonotonic effect model and effect eigensignals.....               | 47        |
| <i>I.1.2.4</i> Precision medicine .....  | 8         | <i>II.2.2</i> Basic Concepts.....   | 48        |
| <i>I.1.3</i> Common Stages in Systems Science.....   | 8         | <i>II.2.2.1</i> Gene and sample vector spaces.....                                  | 48        |
| <i>I.1.3.1</i> A cascade of abstraction levels for modeling.....                                     | 8         | <i>II.2.2.2</i> Gene and samples axes.....  | 48        |
| <i>I.1.3.2</i> The measurement stage, next-generation RNA sequencing and unexpected complexity ..... | 9         | <i>II.2.2.3</i> Effect curves .....   | 49        |
| <i>I.1.3.3</i> The increasing gap, an obvious but fundamental problem .....                          | 10        | <i>II.2.2.4</i> Effect focus .....  | 49        |
| <i>I.1.3.4</i> Generic footprints of interactions and the summarization stage .....                  | 11        | <i>II.2.3</i> Measures of Interaction .....   | 50        |
| <i>I.1.3.5</i> The association stage and putting it all together .....                               | 13        | <i>II.2.3.1</i> Weighted uncentered correlations aka the cosine distance .....      | 50        |
| <i>I.1.4</i> Motivation and Goals.....   | 14        | <i>II.2.3.2</i> Weighted projections .....  | 51        |
| <i>I.1.4.1</i> Helpful and misleading signal summarizations .....                                    | 14        | <b>II.3 Search Strategy</b> .....   | <b>52</b> |
| <i>I.1.4.2</i> Design goals, the detection task and method preview.....                              | 15        | <i>II.3.1</i> Finding Effects.....  | 52        |
| <i>I.1.4.3</i> Outline.....  | 17        | <i>II.3.1.1</i> Standardization against outliers.....                               | 53        |
| <b>I.2 Standard Analyses</b> .....   | <b>18</b> | <i>II.3.1.2</i> The effect focus and its initial estimation .....                   | 54        |
| <i>I.2.1</i> Supervised Gene Expression Analysis .....   | 18        | <i>II.3.1.3</i> Initial effect axes and symmetrization by twin axes.....            | 55        |
| <i>I.2.1.1</i> Differential expression analyses .....  | 18        | <i>II.3.1.4</i> Correlations and the refined effect focus .....                     | 56        |
| <i>I.2.1.2</i> Application of specific gene signatures .....   | 19        | <i>II.3.1.5</i> Focusing step.....  | 56        |
| <i>I.2.2</i> Unsupervised Gene Expression Analysis .....   | 21        | <i>II.3.1.6</i> Scalar effect score.....  | 57        |
| <i>I.2.2.1</i> Hierarchical clustering .....   | 21        | <i>II.3.1.7</i> Search complexity, presorting and a lookahead scheme.....           | 58        |
| <i>I.2.2.2</i> Principal components analysis (PCA) .....   | 24        | <i>II.3.1.8</i> Qualification of candidates.....                                    | 59        |
| <i>I.2.3</i> Copy Number Analyses .....  | 27        | <i>II.3.2</i> Effect Axes Convergence .....   | 60        |
| <i>I.2.4</i> Viability Curves and the IC50 .....   | 28        | <i>II.3.2.1</i> Iterative selection of representatives .....                        | 60        |
| <b>I.3 Association Methods</b> .....   | <b>29</b> | <i>II.3.2.2</i> Accumulating representatives and the update step.....               | 61        |
| <i>I.3.1</i> Signature Analyses and Gene Set Enrichment .....  | 29        | <i>II.3.2.3</i> Checking for sufficient representatives and for convergence .....   | 61        |
| <i>I.3.1.1</i> Gene set enrichment analysis.....   | 29        | <b>II.4 Regression and Dissection</b> .....   | <b>63</b> |
| <i>I.3.1.2</i> Additional signature statistics and signature heatmaps.....                           | 31        | <i>II.4.1</i> Regression and Effect Curves .....                                    | 63        |
| <i>I.3.2</i> Gene Ontology Analyses.....   | 32        | <i>II.4.1.1</i> Effect strengths for genes and samples.....                         | 64        |
| <i>I.3.3</i> Kaplan Meier Survival and Log Rank Tests .....  | 33        | <i>II.4.1.2</i> The empirical eigenorder .....                                      | 64        |
| <b>Chapter II Signal Dissection</b>  | <b>35</b> | <i>II.4.1.3</i> Bimonotonic regression .....  | 65        |
| <b>II.1 Design Goals</b> .....   | <b>37</b> | <i>II.4.1.4</i> Adaptive smoothing by 2D Fourier transforms.....                    | 67        |
| <i>II.1.1</i> Introductory 3D Example .....  | 37        | <i>II.4.1.5</i> Check for convergence.....  | 70        |
| <i>II.1.1.1</i> A simple linear effect.....  | 37        | <i>II.4.2</i> Effect Dissection.....  | 70        |
| <i>II.1.1.2</i> Supported nonlinear biological effects and three examples.....                       | 38        | <i>II.4.2.1</i> Dissection strengths, final eigensignal and remaining signal.....   | 71        |
| <i>II.1.1.3</i> Merged 3D signal and an exemplary detection task.....                                | 39        | <i>II.4.2.2</i> Effective clustering and limits of projection based methods.....    | 71        |
| <i>II.1.2</i> Method Design Goals .....  | 40        | <i>II.4.2.3</i> Visualization of high-dim. eigensignals: the coordinate view.....   | 73        |
| <i>II.1.2.1</i> Effect focusing should be unsupervised.....  | 40        | <i>II.4.3</i> Remaining Signal and Termination .....                                | 75        |
| <i>II.1.2.2</i> Discovered effects should be specific to true effects .....                          | 40        | <b>II.5 Noise and Significance</b> .....  | <b>76</b> |
| <i>II.1.2.3</i> Partial correlations should be properly resolved.....                                | 41        | <i>II.5.1</i> Significance of Observed Signal Strengths.....                        | 76        |
| <i>II.1.2.4</i> Symmetry of genes and samples.....   | 41        | <i>II.5.1.1</i> Estimating the true noise level .....                               | 76        |
| <i>II.1.2.5</i> Removal of overlapping strong effects.....   | 42        | <i>II.5.1.2</i> Significance of signal strengths.....                               | 78        |
| <i>II.1.2.6</i> Compatibility with gradual effects.....  | 43        | <i>II.5.2</i> Significance of Observed Correlations .....                           | 81        |
| <i>II.1.2.7</i> Number of effects and user-dependency of results.....                                | 43        | <i>II.5.2.1</i> Significance of weighted correlations.....                          | 81        |
| <i>II.1.2.8</i> Completeness of results.....   | 44        | <i>II.5.2.2</i> Significance of all gene and sample correlations for an effect..... | 83        |
| <i>II.1.3</i> Needed Capabilities beyond Detection .....   | 44        |   |           |
| <i>II.1.3.1</i> Comparability and validation of effects across sample cohorts.....                   | 44        |   |           |

## II.6 Method Validation and Comparison .84

|  |     |
|--|-----|
| II.6.1 3D Concept Example.....   | 84  |
| II.6.1.1 Overview of all dissection steps.....                               | 84  |
| II.6.1.2 Comparison with PCA.....  | 85  |
| II.6.1.3 Comparison with hierarchical clustering.....                        | 85  |
| II.6.2 Versatility.....  | 87  |
| II.6.2.1 Scenario definition and 7 distinct effect patterns.....             | 87  |
| II.6.2.2 A detection walkthrough for 1000 dimensions.....                    | 88  |
| II.6.2.3 Comparison of detected and true simulated effects and results....   | 93  |
| II.6.2.4 Comparison with PCA (versatility scenario with 7 effects).....      | 94  |
| II.6.2.5 Comparison with PCA (versatility scenario with 13 effects).....     | 96  |
| II.6.2.6 Comparison with hierarchical clustering.....                        | 100 |
| II.6.3 Superposition Depth.....  | 101 |
| II.6.3.1 Results and comparison with PCA for 1 to 20 times pattern #3... 101 |     |
| II.6.3.2 Results and comparison with PCA for 1 to 20 times pattern #4... 105 |     |
| II.6.3.3 Results and comparison with PCA for 1 to 20 times pattern #6... 108 |     |
| II.6.4 Detection Limits.....   | 110 |
| II.6.4.1 Weak signals.....   | 110 |
| II.6.4.2 Many noise genes.....   | 112 |
| II.6.4.3 Few samples.....  | 118 |
| II.6.4.4 Missing values and their imputation.....                            | 121 |

## Chapter III Dissecting DLBCL Gene Expressions 125

### III.1 Application to DLBCL ..... 127

|  |     |
|--|-----|
| III.1.1 Detection in Single Patient Cohorts.....   | 127 |
| III.1.1.1 Available gene expressions cohorts.....  | 127 |
| III.1.1.2 Dissection overview for single cohorts.....  | 128 |
| III.1.2 Validation of Effects across Cohorts.....  | 129 |
| III.1.2.1 Comparison between two cohorts by corr. of effect gene axes.. 129                          |     |
| III.1.2.2 Validation by independent discovery in several cohorts.....                                | 130 |
| III.1.2.3 Supervised Validation.....   | 131 |
| III.1.3 Genomic Consensus Effects.....   | 132 |
| III.1.3.1 Consensus gene axes.....   | 132 |
| III.1.3.2 Consensus gene scores and their correlation.....   | 133 |
| III.1.4 Application of Genomic Effects.....  | 135 |
| III.1.4.1 Cleaned signal.....  | 135 |
| III.1.4.2 Application of consensus gene axes and sample classification by eigensignal strengths..... | 135 |
| III.1.5 Overview of Scores.....  | 136 |

### III.2 Multivariate Survival Prediction ... 138

|   |     |
|---|-----|
| III.2.1 Survival Model and Effect Selection.....                                  | 138 |
| III.2.1.1 Cox Proportional Hazard Models.....                                     | 138 |
| III.2.1.2 Available survival data and its limited information.....                | 139 |
| III.2.1.3 Choosing sample scores.....   | 140 |
| III.2.1.4 Correcting for survival factors that are not specific for DLBCL.... 141 |     |
| III.2.1.5 Effect selection procedure and likelihood ratio tests.....              | 141 |
| III.2.1.6 Validation techniques.....  | 142 |
| III.2.1.7 Interpreting ambiguities.....   | 142 |
| III.2.1.8 Survival for ABC DLBCL and GCB DLBCL for comparison.....                | 143 |
| III.2.1.9 Revisiting binary subtype classifications and associated cutoffs . 144  |     |
| III.2.2 Bivariate Model for R-CHOP.....   | 145 |
| III.2.2.1 Selection of validated GEP effects as predictor variables.....          | 145 |
| III.2.2.2 Fit results, prediction performance and validation.....                 | 147 |
| III.2.2.3 $\nu = 127$ is a GCB-DLBCL-only survival effect.....                    | 148 |
| III.2.2.4 Predictions within risk partitions of $\nu = 134$ .....                 | 149 |
| III.2.2.5 Predictions within ABC DLBCL and GCB DLBCL subtypes.....                | 151 |

### III.2.3 Bivariate Model for the COO Induced Effect...152

|   |     |
|---|-----|
| III.2.3.1 Selection of validated GEP effects as predictors.....   | 152 |
| III.2.3.2 Fit results, prediction performance and validation..... | 153 |
| III.2.3.3 Subtype-specific analysis of $\nu = 105$ .....          | 154 |

### III.2.4 Bivariate Model for CHOP .....155

|   |     |
|---|-----|
| III.2.4.1 Selection of validated GEP effects as predictors.....   | 155 |
| III.2.4.2 Fit results, prediction performance and validation..... | 156 |
| III.2.4.3 Hierarchical survival analysis of $\nu = 44$ .....      | 157 |
| III.2.4.4 Subtype-specific analysis of $\nu = 44$ .....           | 159 |

### III.2.5 Quinvariate Model for All Samples .....160

|   |     |
|---|-----|
| III.2.5.1 Selection of validated GEP effects as predictors.....                 | 160 |
| III.2.5.2 Fit results.....  | 162 |
| III.2.5.3 Leave-one-out cross-validation and predictor performance.....         | 163 |
| III.2.5.4 Predictions within CHOP and R-CHOP subsets.....                       | 165 |
| III.2.5.5 Prediction performance for FFPE and frozen cell material.....         | 166 |
| III.2.5.6 Predictions within ABC DLBCL and GCB DLBCL.....                       | 167 |
| III.2.5.7 Predictions within risk partitions of $\nu = 134$ .....               | 168 |
| III.2.5.8 Predictions within risk classes by International Prognostic Index 169 |     |

### III.3 Biostatistical Evaluation..... 171

#### III.3.1 Analyses and Statistical Tests .....171

|  |     |
|--|-----|
| III.3.1.1 Association with clinical knowledge..... | 171 |
| III.3.1.2 Association with genomic knowledge.....  | 172 |

#### III.3.2 Effects Identified by Sample Annotations.....173

|   |     |
|---|-----|
| III.3.2.1 $\nu = 2$ : Gender effect and annotation mistakes.....            | 173 |
| III.3.2.2 $\nu = 129$ : Cell of origin induced effect (DLBCL subtypes)..... | 176 |

#### III.3.3 Survival Effects .....183

|   |     |
|---|-----|
| III.3.3.1 $\nu = 134$ : Primary survival effect in DLBCL.....   | 183 |
| III.3.3.2 $\nu \in \{127, 131\}$ : A hierarchical survival effect in GCB DLBCL... 189   |     |
| III.3.3.3 $\nu = 105$ : KIAA1217 (2 <sup>nd</sup> predictor var., COO based model)..... 194   |     |
| III.3.3.4 $\nu = 5$ : A stromal effect (CHOP based model, 1 <sup>st</sup> predictor var.) 195   |     |
| III.3.3.5 $\nu = 44$ : Another stromal effect with a hierarchical survival dependency and revisiting a previous DLBCL survival predictor..... 200 |     |
| III.3.3.6 $\nu = 19$ : A T cell related effect (quinvar. model, 3 <sup>rd</sup> variable).... 204   |     |
| III.3.3.7 $\nu = 75$ : BCL2 (quinvariate model, 4 <sup>th</sup> predictor variable)..... 208  |     |
| III.3.3.8 $\nu = 3$ : A zinc-related effect (quinvar. model, 5 <sup>th</sup> variable)..... 210   |     |

#### III.3.4 Effects without Strong Survival Impact.....215

|   |     |
|---|-----|
| III.3.4.1 $\nu = 20$ : Another perspective on DLBCL subtypes.....         | 215 |
| III.3.4.2 $\nu = 7$ : Presumably the main blood concentration effect..... | 218 |
| III.3.4.3 $\nu = 4$ : A strong immunoglobulin effect.....                 | 220 |

### Conclusion ..... 223

### Research Perspectives ..... 228

### Digital Content..... 231

### Indexes..... 233

|                      |     |
|----------------------|-----|
| Symbols.....         | 233 |
| Named Equations..... | 235 |
| Tables.....          | 235 |
| Figures.....         | 236 |

### Bibliography ..... 239

### Scientific Profile ..... 247

### Acknowledgements..... 249

# Chapter I

---

## From Cells to Knowledge

*After rooting this work in complex systems science, its biological context of Diffuse Large B-Cell Lymphoma is introduced.*

*Summarizations of detailed measurements like human gene expressions are essential to discover novel effects of interactions and to provide an interpretable basis for high-level system modeling.*

*Examples of existing summarizations by generic concepts of interactions are presented and the need for a more compatible concept for gene expressions is demonstrated.*

*Next, selected biostatistical methods for analyses of typical experiments in molecular biology are clarified, including methods that can associate results with existing genomic knowledge. Exemplary analyses performed for several published investigations of DLBCL or of related biological contexts are presented.*



*In general terms, this work is about a complex system, the cancer cell. Theoretical concepts from complex system theory are used to model and detect interactions in this system in order to address biological and medical questions. One specific goal is to help clarifying the molecular pathogenesis of Diffuse-Large B-Cell Lymphoma. This disease and the pathway language utilized to model interactions in cancer cells are introduced. In particular, the NF- $\kappa$ B pathway and its role for normal and malignant B cells is briefly presented.*

*Technologies like microarrays and RNA sequencing can measure more and more parts and details of cancer cells. Hence, methods that can summarize observed signals by interactions become increasingly important in bioscience and in many other fields of science as well. Signal summarization is located above technology-specific signal processing and below system-specific modeling by experts.*

*For biological interactions observed by gene expressions and for similar signals, a more compatible summarization concept is needed, as is motivated by an intuitive 3D example. In particular, it is demonstrated that standard principal components analyses of such signals deliver misleading summaries that may prevent otherwise possible insights into underlying pathways. A preview illustrates the detection task and an outline concludes the introduction.*

## What is a complex system?

---

A typical complex system is fluid flow: While at lower energies laminar flow can be observed, i.e. a smooth flow like that of a calm river without any large local differences in movement direction or speed, at higher energies the same fluid may show turbulence with strong and seemingly random fluctuations over many scales, for example at the end of a waterfall. Between the two extremes, many interesting transitions from laminar to chaotic flow and vice versa can be observed and are studied in the nonlinear dynamics field of theoretical physics. Despite this complexity, the local behavior of systems of fluid flow can be modeled by just two lines of math, the Navier-Stokes equations. The same equations can also be utilized to predict fluid flow, given its initial conditions. For laminar flows, predictions of very high accuracy are possible even over longer time periods. However, turbulent flows can only be predicted with acceptable accuracy for rather short periods and for precise knowledge of initial conditions, because they behave completely differently over time for only minimal deviations in initial conditions. Mathematically, this complexity can be rooted in a nonlinear differential operator in the Navier-Stokes equations. Nonlinearity generally indicates that the superposition principle no longer holds, i.e. the system can no longer be described as a sum of easier parts without considering their *interaction*.

For a more interdisciplinary view on complex systems, the nature of complexity needs to be extracted. From the fluid flow example above two things can be learned: Most importantly, all the complexity of fluid flow is possible, although locally every fluid molecule just follows the same simple and predictable laws of motion described by the Navier-Stokes equations. Hence, complexity is a *result of interaction* and cannot be found nor explained on the level of single elements of the system. Additionally, small changes at one time point can lead

to large qualitative and *unpredictable changes* as the system evolves. A more complete yet concise description of a complex system can be found in the Encyclopedia of Complexity and Systems Science<sup>[1]</sup>:

“Complex systems are systems that comprise *many interacting parts* with the ability to generate a new quality of collective behavior through *self-organization*, e.g. the spontaneous formation of temporal, spatial or functional *structures*. They are therefore *adaptive* as they evolve and may contain self-driving feedback loops. Thus, complex systems are *much more than a sum of their parts*. Complex systems are often characterized as having *extreme sensitivity to initial conditions* as well as emergent behavior that are *not readily predictable* or even completely deterministic.

The conclusion is that a reductionist (bottom-up) approach is often an *incomplete description* of a phenomenon. This recognition, that the collective behavior of the whole system *cannot be simply inferred* from the understanding of the behavior of the individual components, has led to many new concepts and sophisticated mathematical and modeling tools for application to many scientific, engineering, and societal issues that can be adequately described only in terms of complexity and complex systems.”<sup>[1]</sup>

## 1.1.2 Biological Context

---

The medical and biological context of this work is briefly introduced here, including the cancer type in focus, its known subtypes and their original definition. Some functions of the healthy ancestor cells from which this cancer derives are illuminated. In this context, simple pathways and exemplary known molecular causes for the disease are presented. Finally, the goal of precision medicine is explained.

### 1.1.2.1 Diffuse Large B-Cell Lymphoma and cells as complex systems

---

Throughout this work, Diffuse Large B-Cell Lymphoma (DLBCL) serves as the real-world complex system in focus for method development and application. Malignant lymphomas are cancers that develop from cells of the lymphatic system and then proliferate there. There are approximately 422.000 cases of lymphoma per year<sup>[2]</sup> (world-wide estimate from 2008), of which DLBCL is the most common type, accounting for approximately 30-40% of all adult lymphoma cases<sup>[3]</sup>. If untreated, DLBCL ultimately causes death, because the lymphatic system is an integral part of the human immune defense system. DLBCL is known to be a genetically heterogeneous disease with two dominant subtypes that are morphologically hard to distinguish: activated B-cell-like (ABC) and germinal center B-cell-like (GCB) DLBCL. Originally, these two subtypes were detected with and defined via hierarchical clustering of DLBCL gene expressions in 2000<sup>[4]</sup>. They were named after their shared gene expression programs with samples of activated blood B-cells respectively germinal center B-cells from healthy donors<sup>(Figure 1.1.2.1, unterhalb)</sup>; this also suggests that the malignant cells derive from these different normal B-cells.

In principle, the disease is curable with a chemotherapy combining several cytotoxins (small molecule compounds cyclophosphamide, doxorubicin, vincristine and prednisone, in short CHOP), but the 5-years overall survival was only 60% for GCB DLBCL and only 30% for ABC DLBCL patients<sup>[5]</sup>. It was possible to improve survival significantly by 10%-15% via inclusion of an immunotherapy (Rituximab-CHOP, in short R-CHOP)<sup>[5]</sup>. The two subtypes found on gene expression level and validated on survival level suggest that different interactions with the chemotherapy take place in different patients, which encourages the search for subtype-specific therapies for further improvement.

Taking a systems perspective again, there are two natural levels of abstraction here: the single cancer cell and the human body. Both have well-defined geometric borders in form of the cell membrane respectively the

human skin. Neither is a closed system as they obviously interact with their environment. Both levels are valid descriptions and the right choice depends on the question. In molecular biology, the primary focus is to investigate the molecular pathogenesis, i.e. how a single cancer cell works. In particular, *why it has started to proliferate in an uncontrolled way* compared to healthy cells and how it can be *manipulated towards apoptosis* (i.e. programmed cell death). Therefore, the more suitable choice here is to focus on the single cell as the system and abstract its environment by functional molecules that can interact with it, for example by binding to the cell's surface receptors.

The cell fits the above description of a complex system: It is comprised of many functional molecules like proteins and many more interacting parts. In case of a cancer cell, they collectively self-organize a procedure of rapid cellular reproduction. They can adapt and evolve; for example, they might acquire oncogenic mutations to their DNA that are beneficial for their reproduction. It is even assumed that such mutations, either occurring randomly or due to failure in DNA maintenance and repair or induced by exogenous toxins, are also causal

for cancer genesis in the first place<sup>[6]</sup>. This is also a prime example where a tiny change in the cell's DNA (i.e. its "initial conditions") causes the emergence of a qualitatively completely different anti-apoptotic reproduction behavior in the long-term. Trying to understand this complex behavior from the perspective of single proteins or genes is futile. *Their interactions* must be investigated, modeled and understood.

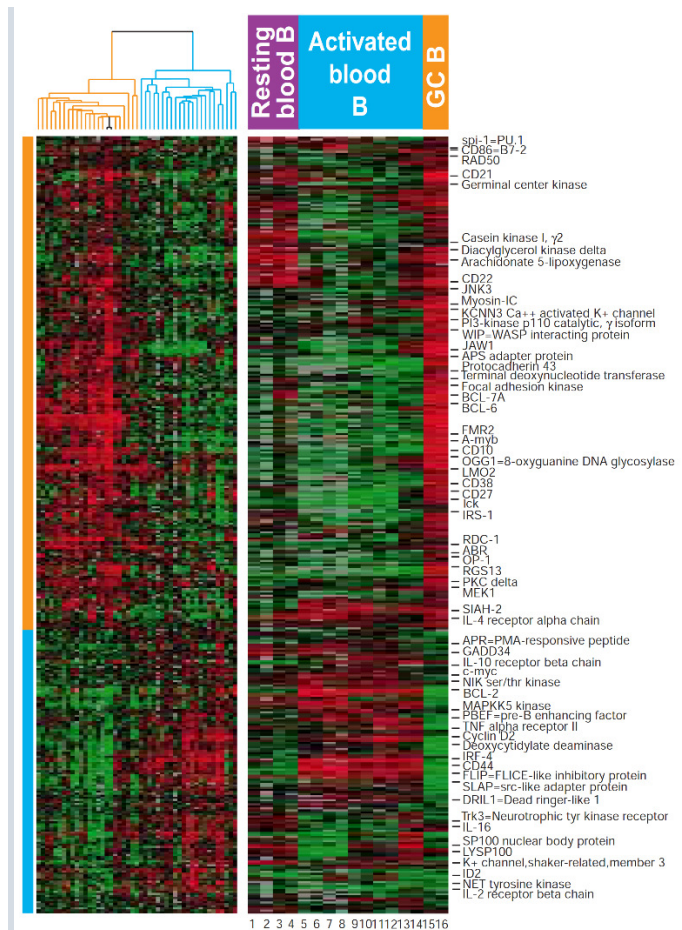


Figure I.1.2.1) Definition of subtypes ABC DLBCL and GCB DLBCL<sup>[4]</sup>

Original detection and definition of the activated B-cell-like (ABC) and germinal center B-cell-like (GCB) subtypes of Diffuse Large B-Cell Lymphoma. On the left, tumor samples from patient and their unsupervised hierarchical clustering by gene expressions is shown. This results in two distinct patient groups, orange and blue. On the right, different normal B cell samples from healthy donors are shown with genes aligned to the left. A similarity of the gene expression programs in germinal center B cells with the orange patient group, and of the activated blood B cells with the blue patient group, suggests a different cellular origin and consequently different pathogenesis of the malignant GCB-like and ABC-like subtypes of DLBCL. Obviously, this similarity is rough, especially with respect to genes upregulated in ABC DLBCL samples.

(Reprinted by permission from Macmillan Publishers Ltd: [Nature](#)<sup>[4]</sup>, copyright 2000)

### I.1.2.2 Pathways, normal B cells in the immune system and NF-κB

Molecular biology uses the language of pathways to model these interactions. One particular important pathway is NF-κB (nuclear factor κB) that stands for a protein family found in several species. These proteins act as transcription factors, i.e. they can enter the cell nucleus, bind to specific DNA sequences, recruit RNA polymerases and thereby initiate transcription of specific target genes. Resulting messenger RNA molecules (mRNAs) then leave the cell nucleus and ribosomes translate them into proteins that finally change the cell's function. Central to the (classical) NF-κB pathway (Figure I.1.2.2, unterhalb) is the p50-RelA heterodimer that is normally bound by an independent inhibitory IκB protein in the cytoplasm and thereby inactivated, as it cannot enter the nucleus in this form. Via various extracellular signals that can trigger cellular responses through surface

receptors, enzymes (more precisely, the IKK complex with I $\kappa$ B kinases  $\alpha$  and  $\beta$ ) are activated and then phosphorylate I $\kappa$ B, thereby activating the p50-RelA transcription factor<sup>[7]</sup>.

Besides many other roles, NF- $\kappa$ B plays an important role in the human adaptive immune system. For example, if a so far unknown antigen has entered the body, e.g. a microbial toxin or another foreign macromolecule. Almost certainly, there are already B-cells in the blood having antibodies (depicted by the large Y-shaped transmembrane protein) with antigen-binding sites that geometrically and chemically match a candidate binding site in the antigen.

This is because of somatic hypermutation in germinal centers of the lymphatic system, which gives rise to a stochastic repertoire of no less than  $10^{11}$  different antigen binding sites<sup>[8]</sup>. Together with another protein on the inside of the cell (CD79), an antibody that is located in the B cell's membrane is called a B-cell-receptor (BCR). The BCR is an interface that allows B cells to react on their environment. Somatic hypermutation may also produce B cells with BCRs that are specific to natural macromolecules of the body (self-antigens). Matching self-antigens cause a strong signaling at these cells' BCRs already during their maturation. Normally, this either reactivates their recombinase machinery to generate another non-autoreactive specificity or sends them into apoptosis before they mature and enter the body's immunocompetent B cell inventory (negative selection), thereby preventing autoimmune diseases<sup>[8]</sup>.

Mature non-autoreactive B cells now react to foreign antigens in the blood. If a matching antigen comes sufficiently close, it chemically binds to one of the B cell's antibodies/receptors on its surface. This activates the B cell: A chain of biochemical interactions inside the cell (signal transduction) is caused that also engages the NF- $\kappa$ B pathway. One possible functional response of this activated B cell is that it starts to proliferate and build a cell population that produces antibodies for this specific antigen, which are secreted into the blood. These free antibodies again bind to matching foreign antigens and thereby inactivate them. In this way, microbial toxins may be blocked from binding to receptors of healthy cells. If the antigen originated from a virus hull, resulting antibodies may also bind to such viruses. This makes it easy for phagocytic cells of the innate immune system to ingest them, thereby destroying these viruses.<sup>[9]</sup>

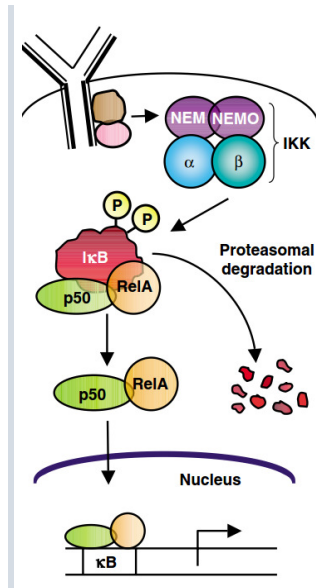


Figure I.1.2.2) Classical NF- $\kappa$ B pathway<sup>[7]</sup>

Extracellular signals trigger an intracellular response through a transmembrane receptor protein, and then activate the IKK complex that subsequently phosphorylates I $\kappa$ B. This allows the p50-RelA complex to enter the nucleus, bind to  $\kappa$ B sites in the DNA and trigger the transcription of downstream genes.

(Reprinted by permission from Macmillan Publishers Ltd: *Oncogene*<sup>[7]</sup>, copyright 2006)

### 1.1.2.3 Molecular causes of B cell malignancy

While proliferation of activated B cells following transient NF- $\kappa$ B activation in response to pathogens is normal for a healthy immune system, lesions in the B cell's DNA like deletions or amplifications may lead to an oncogenic activation of NF- $\kappa$ B. Such constitutive NF- $\kappa$ B activation also underlies the ABC DLBCL subtype<sup>[10]</sup>. In general, activated NF- $\kappa$ B signaling plays a pathogenic role in many types of cancer<sup>[10]</sup>. These tumor cells have a strong selective advantage compared to healthy cells, because the NF- $\kappa$ B pathway also blocks normal cell apoptosis<sup>[10]</sup>. They continue to produce copies of themselves that include the same oncogenic defects in their DNA and proliferate likewise. For patients this causes almost certain death in the long term, if untreated.

A major goal concerning ABC DLBCL thus is to interrupt NF- $\kappa$ B signaling in a way that stops proliferation or even re-enables apoptosis. This is possible, for example, by a small molecule inhibitor for I $\kappa$ B kinase  $\beta$



(IKK $\beta$ )<sup>[10]</sup>, because it attacks downstream of many possible NF- $\kappa$ B signaling pathways: It directly prevents phosphorylation of I $\kappa$ B $\alpha$  by IKK $\beta$  and thereby prevents the p50-RelA complex from entering the nucleus. However, this inhibitor is not appropriate for development of a specific treatment, as it attacks the NF- $\kappa$ B pathway in all cells. Ideally, a genetic “Achilles’ heel” could be found that specifically stops malignant ABC DLBCL cells.

Experiments utilizing RNA interference (RNAi) by small hairpin RNAs (shRNAs) allow the inhibition of specific genes in ABC DLBCL model cell lines. For example, such experiments have shown that inhibiting CARD11, BCL10 or MALT1 (the CBM signaling complex) is toxic for ABC DLBCL cell lines<sup>[10]</sup>. CBM signaling is located upstream to the IKK complex of the NF- $\kappa$ B pathway<sup>(see [10].figure 1A)</sup> and thus is a more specific attack vector, but it is still transiently involved in and required for normal antigen response<sup>[10]</sup>. Approximately 10% of ABC DLBCL patients show a somatic DNA mutation that causes direct oncogenic activation of CARD11<sup>[10,11]</sup> and thereby NF- $\kappa$ B signaling, which may be one of ABC DLBCL’s pathogenic causes. For the majority of ABC DLBCL patients, toxicity after CBM inhibition indicates that their malignant constitutive NF- $\kappa$ B signaling has its source already upstream of the CBM complex. Even for the 10% having CARD11 mutations, there may be other additional upstream signaling sources. One way to find signaling sources is to directly look for genetic aberrations and link this information to gene expressions in the same cells; for instance, this approach allowed identifying SPIB as an upregulated oncogene candidate in ABC DLBCL due to chromosomal gains and amplifications. Indeed, SPIB is also critical for survival of ABC DLBCL cell lines, but not for GCB DLBCL cell lines<sup>[11]</sup>. Discovering and understanding these ABC DLBCL specific interactions is part of ongoing molecular investigations.

The GCB DLBCL subtype on the other hand is not based on constitutive NF- $\kappa$ B signaling and the expression of NF- $\kappa$ B target genes in GCB DLBCL is significantly lower than in ABC DLBCL. Consequently, neither blocking CBM signaling nor treatment with I $\kappa$ B inhibitors is toxic for GCB DLBCL cell lines<sup>[10]</sup>. Here, looking for pathogenic aberrations revealed a loss of the tumor-suppressor gene PTEN<sup>[11]</sup>, and immunohistochemical staining<sup>(e.g. Figure 1.1.2.3)</sup> showed that approximately 55% of GCB DLBCL samples are PTEN negative, but only approximately 14% of non-GCB DLBCL samples<sup>[12]</sup>. While shRNAs allow the experimental inhibition of genes, it is possible to use cDNAs in order to overexpress specific genes in experiments. Overexpressing PTEN in PTEN-deficient GCB DLBCL cell lines killed these cells, confirming PTEN’s role as tumor suppressor. Furthermore, it has been detected that this dependence of GCB DLBCL cells on PTEN loss is because of a constitutive activation of the PI3K signaling pathway<sup>[12]</sup> that is usually inhibited by PTEN: Overexpressing AKT, the main effector of PI3K signaling<sup>[13]</sup>, rescued these cells despite presence of otherwise toxic PTEN cDNA. Additionally, using cDNAs of PTEN mutants that cannot inhibit PI3K signaling were not toxic. Finally, treating PTEN-deficient GCB DLBCL cell lines with a pharmacologic small molecule compound that is a potent inhibitor of PI3K kinases, significantly reduced cell viability, whereas PTEN-positive cells were unaffected<sup>[12]</sup>.

These are just few examples for the genetic heterogeneity of DLBCL.

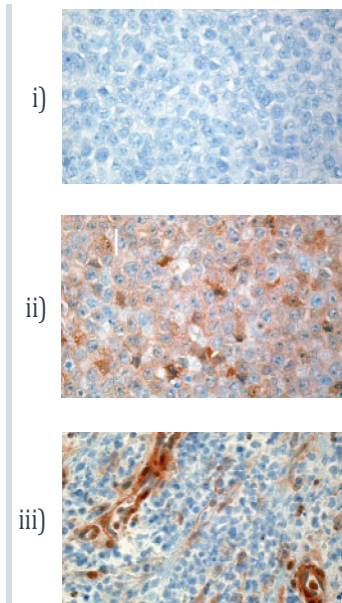


Figure 1.1.2.3) PTEN-stained DLBCL cells<sup>[12]</sup>  
 (i) PTEN-negative GCB DLBCL line HT.  
 (ii) PTEN-positive ABC DLBCL line TMD8.  
 (iii) A PTEN-negative DLBCL patient sample (with blood vessels as internal positive control). (Reprinted and adapted from a co-authored paper<sup>[12]</sup>)

The shared goal in medicine, molecular biology, biophysics, biostatistics, bioinformatics and pharmacy in context of heterogeneous genetic diseases is to enable precision medicine.

To this end, initially oncogenic drivers need to be identified. This requires *genomic measurements* of a sufficient number of cancer samples. With the Cancer Genome Atlas<sup>(see TCGA Research Network, <http://cancergenome.nih.gov>)</sup> a large and public data source already exists to this end and is steadily expanding. Next, candidate oncogenes and tumor suppressors need to be identified by *analyzing and summarizing* these huge and detailed genomic data. Subsequently, promising candidates need to be *biologically validated*. Based on validated results, the pathway language is utilized for *modeling* this atlas of oncogenic drivers; this allows a rich qualitative description of various kinds of discovered interactions in cells and between cells. Pathways also facilitate molecular reasoning on a high level and can help to generate hypotheses or to transfer knowledge from one cancer type to another. On pharmaceutical level, *novel agents* need to be developed that specifically target validated and biologically promising drivers *in-vivo*. Finally, a set of these agents that is specific to the set of drivers detected *in an individual cancer patient* could be applied as therapy; this is called **precision medicine**. This is also the next major breakthrough that is targeted for DLBCL treatment.

The recent<sup>(January 2015)</sup> United States Government Precision Medicine Initiative also demonstrates the priority of this research<sup>[14]</sup>. This initiative has the aim to measure genomic data for *up to one-million samples*<sup>[15]</sup>, an unprecedented amount of genomic data!

From a certain perspective, the complexity of cancer cells may even help towards reaching this goal: The number of genes that can be simultaneously modified by shRNAs or cDNAs in biological validation experiments is limited to just a few. It is not trivial that it is actually possible to send cells into apoptosis by inhibiting or overexpressing just few or even just a single gene. Only by virtue of the complex interaction and signaling chains in tumor cells, it is possible that such Achilles' heels indeed *exist*.

### I.1.3 Common Stages in Systems Science

---

Based on examples from biosciences, similar general processing stages that are shared by many different fields of science are identified and presented here. In particular, the increasing need for an intermediate stage is demonstrated that is able to provide comprehensible summaries of more and more detailed measurements about parts of complex systems.

#### I.1.3.1 A cascade of abstraction levels for modeling

---

The added immunotherapy in form of Rituximab resulted in a DLBCL therapy with significantly more favorable patient outcome; this already indicates that interactions with the human immune system may play a likewise important role in disease progression. In the concrete case, it is understood that the antibody Rituximab binds to the CD20 protein on B cell surfaces, thereby marking them for ingestion and destruction by natural killer cells of the immune system<sup>[16]</sup>. The dynamics of this antibody-triggered cell-cell interaction also presents an interesting biomechanical modeling challenge: It has been observed that Rituximab causes a polarization of B cells by concentrating their CD20 on a single cap of the cell membrane as opposed to a uniform distribution on the surface. By this redistribution natural killer cells of the host are 60% more effective at killing the B cell<sup>[16]</sup>. This selective killing of B cells is a logical complement to the CHOP chemotherapy of DLBCL.

This example implies that the cell-cell interaction level may allow researching additional potential attack vectors, but this also requires a higher level of abstraction for models: So far, interactions within the cancer cell were considered, while the coupling to the cellular environment has been modeled by molecules or antigens binding to the cell's surface receptors. Now the system is no longer the single cell, but consists of cells itself. Again, this is a complex system, as can be easily verified: Cells interact with each other and show self-organization, for example by building organs in the human body. Trying to understand the death of patients from the perspective of a single cancer cell is again futile, even if all its intra-cellular interactions were understood. However, changing the initial conditions slightly, for example by implanting just a single cancerous cell, might cause a tumor to emerge that can interact with the host organism in such a way that it dies, a truly significant change in system evolution over time.

Obviously, both levels of abstraction could be merged, but trying to model the human body by choosing genes as basic elements of interaction is not as useful from a modeling perspective: Models should be of a manageable and comprehensible complexity. Otherwise, no useful and verifiable predictions can be made. It is impractical to gain any novel knowledge with help of overly detailed or too simplistic models. It makes even more sense for many questions to refine the modeling cascade by introducing additional intermediate levels of abstraction, for example organs or cell organelles.

#### 1.1.3.2

### The measurement stage, next-generation RNA sequencing and unexpected complexity

---

Besides the choice of a useful top abstraction level for scientific modeling, it is essential to understand which parts of the system can actually be observed. In molecular biology, a single standard gene expression (GEP) microarray can profile the activity of the human genome at approximately 50000 genomic loci simultaneously, including probes for mRNA sequences from all known human genes. Newer RNA sequencing technologies can even deliver a detailed view of the transcriptome, i.e. of the total active RNA in measured cancer cells, based on millions of reads and not just for a limited number of preselected mRNA sequences probed by a microarray. The possibilities of RNA sequencing in particular have “become increasingly important in cancer research – all at a data scale that was unimagined just several years ago”<sup>[17]</sup> and have already led to “a new appreciation for the *complexity of the transcriptome*, encompassing a multitude of previously unknown coding and non-coding RNA species”<sup>[17]</sup>.

This complexity was somewhat unexpected, because the paradigm of gene transcription into mRNA and subsequent direct translation into proteins by ribosomes now seems to be a too generalized view. The translation into proteins is modulated by an interconnected network of mRNAs as well as short and long noncoding RNA molecules<sup>[18]</sup> for the majority of genes. This post-transcriptional level of regulation probably needs to be investigated with the same effort as the protein level in order to understand its interaction chains. This newfound complexity of post-transcriptional regulation may even require revisiting older interpretations of gene expression measurements: The basic assumption of direct protein level inference from gene expressions should be applied with care and when making gene expression based statements about single genes, independent experimental validations of the protein concentrations are mandatory. For example, gene expressions based on RNA sequencing technologies may show only a very weak (albeit significant) correlation of  $r \approx 0.1$  to  $0.4$  to the gene expressions of the same cells measured by microarray technology<sup>[19, cf. figure 5]</sup>. To complicate things, this may also be in part caused by technological problems or by the computational methods utilized to process the RNA sequencing reads: The estimated expression levels of these methods vary

widely<sup>[20]</sup>, both due to technological uncertainties as well as due to the inherent biological complexity of the transcriptome.

The technological and data processing problems will however eventually be solved as the technology and the methods mature. The sequencing of full-length cDNA molecules (instead of short sequences via shotgun methods) and the direct use of RNA (instead of first converting it into cDNA) are technological steps in that direction<sup>[21]</sup>. Another variant of RNA sequencing that promises an important filtered view is Ribosome profiling: Here only the ribosome:mRNA complexes in cells are measured after halting their translation, rather than the total RNA. This allows to specifically measure only the mRNAs that are indeed being translated into functional proteins<sup>[21]</sup>. Consequently, the basic assumption of protein level inference from measured mRNA concentrations is perfectly valid here.

### 1.1.3.3 The increasing gap, an obvious but fundamental problem

Being able to measure more and more dimensions like RNA concentrations of a complex system is clearly advantageous as it leads to insights into previously underestimated levels of complexity and can provide more possibilities of understanding them. There are some technological, data processing and estimation challenges, but they seem solvable. Assuming that all these challenges have been solved, the ideal result would be a signal comprised of reliably estimated expression levels of all genes and of activity levels for all proteins and for all other functional molecules that may regulate the transcriptional or translational level.

Even then, a fundamental problem would remain: The human working memory for thought and modeling processes has a *limited capacity for simultaneously holding and manipulating independent information*. For example, the visual working memory can only store 3-4 independent items in the short term<sup>[22]</sup>, models for the verbal memory show a maximum of only five or six unrelated words and even with chunking the immediate memory span for sentences is only approximately 15 words<sup>[23]</sup>. This implies that the human brain lacks the capacity to model and understand interesting interactions *directly* from these huge signals that contemporary instruments can measure.

This obvious gap between the amount of observable information from a complex system and the human working memory for modeling is steadily increasing in practically all fields of science, as measurement technologies advance. This is multiplied by the number of measured samples, for example by the up to one-million samples that is prospectively measured in context of the United States Government Precision Medicine Initiative alone<sup>(cf. 1.1.2.4)</sup>. Besides biosciences, especially in astrophysics many improvements had to be developed in the last decade, “attempting to keep up with the vastly increased volume and quality of available data”<sup>[24]</sup>. Here, galaxies or quasars are the complex systems and the primary source of information is their light emissions, more precisely their spectral energy distributions, which are measured with large telescopes and spectrometers. A major, huge and fast growing signal database in this field of science is provided by the Sloan Digital Sky Survey<sup>(SDSS, see <http://www.sdss.org>)</sup> with already approximately 70 terabytes of data in the total SDSS volume III. The current public data release DR10 contains 927,844 galaxy spectra<sup>[25]</sup>, for example. Spectra for the majority of the visible sky have yet to be systematically measured<sup>[26, see figure 2]</sup>, so this data source prospectively also continues to grow (even if there were no further technological advances in spectral coverage).

### 1.1.3.4 Generic footprints of interactions and the summarization stage

Because of this increasing gap, another intermediate stage of research is required. This stage is located above the technology-specific processing and refinement of the growing measurement data about all parts of the system. It is located below the system-specific higher-level modeling of their interactions by experts of the respective field, for example in the pathways language. The goal of this stage is to summarize the detailed signal about the system parts into as few as possible chunks of information that are comprehensible for modeling. Ideally, this summarization is *unbiased*, especially it should not just focus only on some portions of the signal, and it should be *complete* in the sense that all the important classes of interactions and their large-scale effects in the signal are retained and presented to the modeler.

To achieve this, each existing method of the summarization stage applies a generic concept of interaction, either via direct definitions or implicitly through algorithms. These concepts do not try to comprehend the interactions on a system level like in pathways for cancer cells, but rather search for the *basic footprints of these interactions* in the observed signals in order to detect them in the first place. Example concepts are:

| Concept of signal footprints from interactions                | Methods using these concepts                           | Summarization examples   |
|---|--|--|
| Minimal high-dimensional distance (of various metrics)        | Hierarchical clustering <sup>(cf. 1.2.2.1)</sup>       | Generally, all cluster algorithms summarize all measured system parts by relatively few clusters to support subsequent modeling. Clusters contain similar system parts that are distinct from other clusters. For example, the “ABC DLBCL > GCB DLBCL” and “GCB DLBCL > ABC DLBCL” gene signatures summarize gene expressions from many samples by just two flat sets. |
| Orthogonal directions of maximal variance                     | Principal components analysis <sup>(cf. 1.2.2.2)</sup> | Just a few principal components for high-dimensional data may already capture most of the signal’s variability. Many gene expressions may thus be summarized by just a few linear combinations of all genes.   |
| Frequency and periodicity over time                           | Fourier transformation                                 | The first few Fourier terms can already capture large-scale features of a signal. An example is image smoothing by removing noisy high-frequency information.  |
| Assuming a constant neighborhood over a predefined order axis | Circular binary segmentation <sup>[27]</sup>           | Long segments of equal DNA copy number aberrations summarize millions of noisy reads along the genomic sequence that were measured by array comparative genomic hybridization (aCGH); this summary may allow detection and modeling of oncogenes.  |

Table 1.1.3.4) Example concepts of generic summarizations by interaction

All examples can be viewed as a pre-modeling summarization by different effects of interactions on the observable signal; these summaries can then serve as a basis for higher-level system modeling. Other than modeling languages like pathways, these concepts are rather generic and much less specific to a certain field of science. Therefore, corresponding summarization methods and their underlying modeling ideas often turn out to be of *surprisingly interdisciplinary nature*, even if they were developed in the context of only a single field of science. The relative simple method of hierarchical clustering from computer science and its tremendous success in many fields of science, especially in biosciences<sup>(cf. 1.2.2.1)</sup>, is a good example of that.

Another intuitive physical example is a rigid body: When applied to the positional measurements of its parts, principal components analysis (PCA) delivers its three principal axes of rotation, i.e. when applying a torque around such an axis the body begins to rotate exactly around this axis only, i.e. all its atoms that lie on this axis stay in space where they were before rotating. If the same torque was applied around any other non-principal direction this is not the case. Exceptions are perfectly spherical bodies of constant density; here all directions are principal (as long as the body does not already rotate). Therefore, PCA can summarize an insightful effect of the true interactions within a rigid body that gives rise to its rotational behavior, namely that all its atoms

are forced to keep a constant distance to their neighbors and that the motion of each atom is governed by Newton's laws. This illustrates the differences between what can be *measured* (positions of atoms), how these measurements can be *summarized* (principal axes) and the *true nature of interactions* between system parts (forces keeping atomic distances to neighbors approximately constant). This summary might seem trivial, because a rigid body is not a complex system and it is well understood. Additionally, the positional measurements of its parts might still seem comprehensible even at high resolution, if visualized in 3D.

In biology however, neither a final set of laws for cancer cells nor for DLBCL pathogenesis is known. These systems are far more complex, their measurements deliver signals of huge detail and these signals cannot be visualized intuitively in 3D, since they are not only high-resolution but also high dimensional. Modeling such biological systems can therefore greatly profit from signal summarization.

While sometimes effects from the summarization stage may be directly accessible for high-level system modeling, it is often helpful to associate them with existing computable knowledge first, if available. This can also help to prevent redundant discoveries in fields that produce a fast-growing collection of knowledge.

In biosciences, an important computable knowledge base is the *gene ontology* that hierarchically associates genes with their already discovered molecular functions, cellular components or biological processes<sup>(cf. I.3.2)</sup>. Additionally, large *gene signature databases* can help to associate experiments with related discoveries in scientific articles<sup>(cf. I.3.1)</sup>. Bioinformatics and biostatistics methods such as overrepresentation analyses or gene set enrichment analyses facilitate these associations. For example, they could reveal known gene signatures that are significantly enriched for top regulated genes in a cell line experiment for a tested inhibitor. Simple *supervised analyses* like *t*-tests that identify these significantly regulated top genes in the first place are also part of this association stage. Additionally, often system properties of different formats have been measured, for example gene expressions, patient gender and patient outcome following chemotherapy. It is important to associate these different sources of information to get a complete as possible picture of the system and to filter out disease-unspecific information, for example gender-specific gene expression effects. Biostatistical methods like Kaplan-Meier survival analyses and log-rank tests or contingency tables and  $\chi^2$ -tests allow quantifying these associations with *p* values, i.e. with the *probability* to see a particular association or an even stronger one by pure chance.

*Considering everything, first, a complex system in nature is observed using measurement technologies. Ideally, they deliver precise signals for all parts of the system. Because these signals are often too detailed to make directly sense of them, they are summarized next. If computable existing knowledge is available, summarized effects are associated with it. Finally, this yields the basis for modeling interactions between system parts in a matching modeling language.*

*These models ideally provide new insights that lead to predictions and hypotheses about the analyzed system. Via feedback on the experimental design, these hypotheses can be tested. Once a system has been sufficiently understood, it may be possible to manipulate it towards useful goals, for example curing DLBCL.*

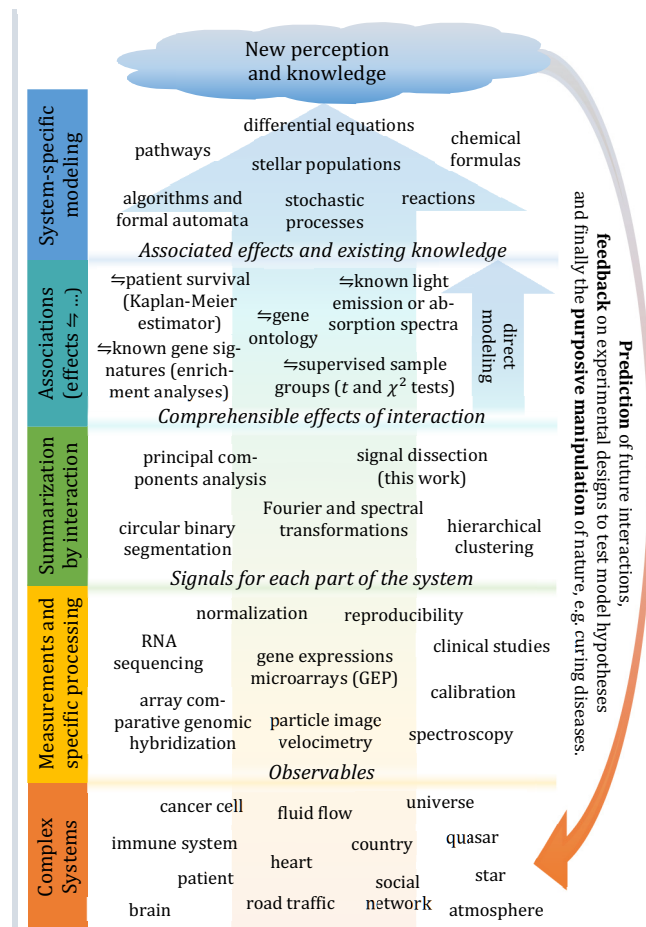


Figure I.1.3.5) Typical stages of systems research with examples, mostly from biosciences

First, it is illustrated why PCA may lead to hard-to-interpret signal summaries. Consequential design goals are presented and a 3D concept example illustrates a useful signal summary contrasting PCA results. Thereby, the detection task is clarified. Additionally, a real-world example for detectable high-dimensional effects of biological origin is provided. Finally, a brief outline concludes the introduction.

### Helpful and misleading signal summarizations

Although all introduced concepts utilized for summarization<sup>(I.1.3.4)</sup> are rather generic and thus have interdisciplinary applications, it is still very important that the concept of the chosen method is *compatible* with the nature of interactions in the underlying system. Otherwise, a summarization into few effects of interaction might not work at all. Alternatively and worse, it might still succeed, but resulting effects are hard or impossible to interpret and thus not helpful or even strongly misleading when modeling the system.

Conceptually, this can already be illustrated<sup>(cf. Figure I.1.4.1)</sup> with just three dimensions: For simulated expressions of three genes driven by two distinct pathways causing linear co-regulation<sup>(red and blue)</sup> the principal components<sup>(yellow)</sup> follow different directions (i.e. describe different laws of gene co-regulation) and thus cannot dissect the two simultaneously measured effects. Even though this signal might be reduced from three to two dimensions by this PCA (as the third principle component explains negligible signal variance on top of the first two), true effects and corresponding groups of simulated patients are still mixed in the new coordinates, making the reduced two-dimensional signal as hard to interpret as the original signal. This is because PCA still treats data points as if they were governed by interactions of a rigid body<sup>(cf. I.1.3.4)</sup>, i.e. the yellow principal components *still are* the physically meaningful principal axes of rotation for an X-shaped body with these points as mass elements. However, this is meaningless and misleading in the gene expression context. A useful summary here would instead separate both groups of patients and deliver one axis, i.e. one linear law of gene co-regulation, for each pathway.

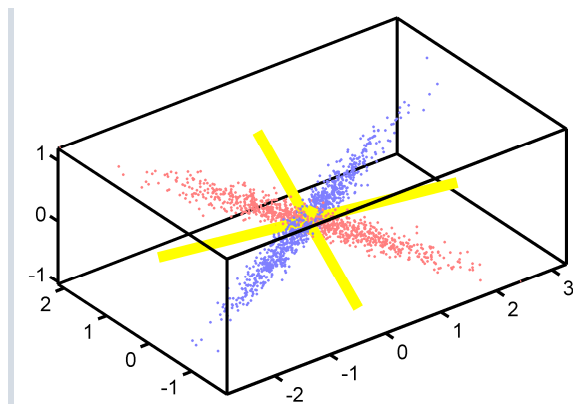


Figure I.1.4.1) Principal components do not point into effect directions

Points for three dimensions simulate two distinct but partly correlated (i.e. not perpendicular) effects for 1000 samples each (red and blue). The principal components returned by PCA are the three yellow perpendicular directions. The two longer components capture nearly all of the signal's variance here; therefore, the third component that protrudes where the two longer cross in the center is relatively short. Further details on how the PCA works follow in I.2.2.2.

Although this seems to be an obvious insight, given this visualization and human intuition in 3D, visualization capabilities and human intuition for high-dimensional data are limited. Therefore and at first sight, principal components of high-dimensional signals like gene expressions may appear to be a perfectly valid and concise summary of that signal, although they actually mix many distinct true effects. Hence, the linear law described by a single principal component usually does not specifically describe the co-regulation mediated by a single pathway, making this signal summary hard or impossible to interpret biologically. This is demonstrated in detail when systematically applying PCA to simulated high-dimensional data for method validation and comparison<sup>(II.6)</sup>.

The same problem is also known in astrophysics. Here, PCA has been utilized to summarize spectral energy distributions measured from stars or from far galaxies. However and consistent, "The main difficulty with PCA



is that the interpretation of the empirically determined PC components in terms of physical properties is complex at best.”<sup>[24]</sup>

Generally, methods of the summarization stage are applied to measurements in order to *get ideas about possible laws governing interactions* between parts of the underlying system. Principal components and similar methods essentially reformulate measured signals by spanning an *alternate coordinate system*. The hope is that some of the new coordinate axes<sup>(yellow)</sup> provide insights into interactions of the analyzed system that were not possible in the original units of measurement, for example by isolating a co-regulation law of a distinct pathway.

This idea of a change in coordinates has also been generalized by spectral methods to spaces of functions, rather than finite-dimensional spaces containing measured samples. For a physical example of such a signal summary, fluid flow can be revisited once more: Local nonlinear interactions of this system are already understood, validated and summarized well in form of the Navier-Stokes equations. But resulting complex large-scale flow behavior is not completely understood and is not clearly described by Navier-Stokes equations. Hence, it makes sense to try to reformulate or re-summarize these known laws about neighboring fluid elements analytically in order to get ideas about how large-scale flow behavior might be generated from local interactions. To this end, an analytic linear expansion of a flow field for low Reynolds numbers has been performed<sup>[28]</sup> by a separation ansatz in a certain geometry that allowed identifying time, radial and angular eigenfunctions of the flow; a linear superposition can then reconstruct the complete flow field. Because resulting linear operators quickly gained in complexity over expansion order, benefits in understanding interactions based on this expansion were unfortunately limited: No obvious summarization of large-scale flow properties was possible in these new spectral coordinates. Finding better analytic summarization concepts that are more compatible with nonlinear interactions in fluid flow is a subject of ongoing research in nonlinear dynamics. Maybe empirical methods summarizing particle image velocimetry measurements could help finding them. Ideally, a simple (maybe statistical) description can be obtained, analogous to the simple summary of large-scale rotational behavior of complex-shaped rigid bodies by principal axes.

In one form or another, the problem of finding useful summaries of high dimensional signals is likely to be known in many more fields of science.

#### 1.1.4.2 Design goals, the detection task and method preview

---

Consequently, it is a major design goal for any novel summarization method that its *summary effects are specific to true effects* in the analyzed system. In particular, summarized effects of gene expressions should not mix signal parts from different pathways that may be active simultaneously in measured cells. Another design goal is the *removal of superposed strong effects* without affecting signals from overlapping weaker, but biologically often more important effects. This is particularly important in the presence strong lab effects caused by measurement technology or protocol. These and several more design goals are presented<sup>(cf. 11.1.2)</sup> in the context of bioscience and gene expressions. In part they can be derived from problems associated with principal components analysis or hierarchical clustering in this context.

Another major conceptual goal for signal dissection is the support for a *broad class of non-linear effects* that may occur in gene expression signals. For example, some genes may reach their saturation expression for lower average activity of the underlying pathway than other genes in the same effect. Some pathways may contain feedback loops that may cause some of its genes to be expressed stronger than linearly (e.g. quadratically) over its average activity. Yet other pathways may show unsteady switch behavior, i.e. some or all of its genes are expressed either at zero or immediately on a plateau of high expression.

A concept example with just three genes works best for illustrating this (Figure 1.1.4.2); it is defined in detail in II.1.1. Briefly, besides a usual linear law of co-regulation (red), pathways in this example simulate a saturation effect (blue), a stronger than linearly regulated gene (green) and a threshold behavior (magenta).

Application of PCA on this signal results in depicted principal components (upper panel, yellow). Again, they are not helpful, both due to PCA's incompatibility with the underlying simulated interactions as explained above (I.1.4.1) and because of a fundamental conceptual limitation shared by all projection methods that try to represent the signal in new coordinates: There are four effects here, but only three dimensions. Whatever directions principal components may point to, whatever eigenvectors may be determined by an alternative spectral separation of the signal, whatever orthonormal rotation of the coordinate system results, after three projections in 3D *only zero remains*. Hence, at least one of the four distinct effects of the example cannot be properly described and dissected by methods based on or equivalent to full projections as a matter of principle. Thus, it is not sufficient to just represent the signal in another coordinate system, if the goal is discover a specific summary *for each effect*.

The task for detection methods is to *recover the laws of gene co-regulation for each simulated pathway empirically from measured points*. (Of course, the color coding is not provided for this task.)

A preview of four monotonic effect gene curves, one for each simulated effect, demonstrates that signal dissection (Chapter II) is able to solve this task (lower panel, yellow).

The 3D example is useful for illustration, but the main goal is to dissect possibly overlapping effects in very *high-dimensional* signals like 50000-dimensional gene expressions. Actually, the developed bimotonic effect model (II.2.1.2) was inspired by ordered heatmaps of real-world gene expression signatures that are *known to be biologically relevant*, like the cell-of-origin induced gene expression effect for distinguishing ABC DLBCL from GCB DLBCL (cf. Figure 1.1.2.1 or Figure 1.1.4.2.b). Only after the method was already operational for such high-dimensional signals, the above 3D example was devised for conceptual illustration

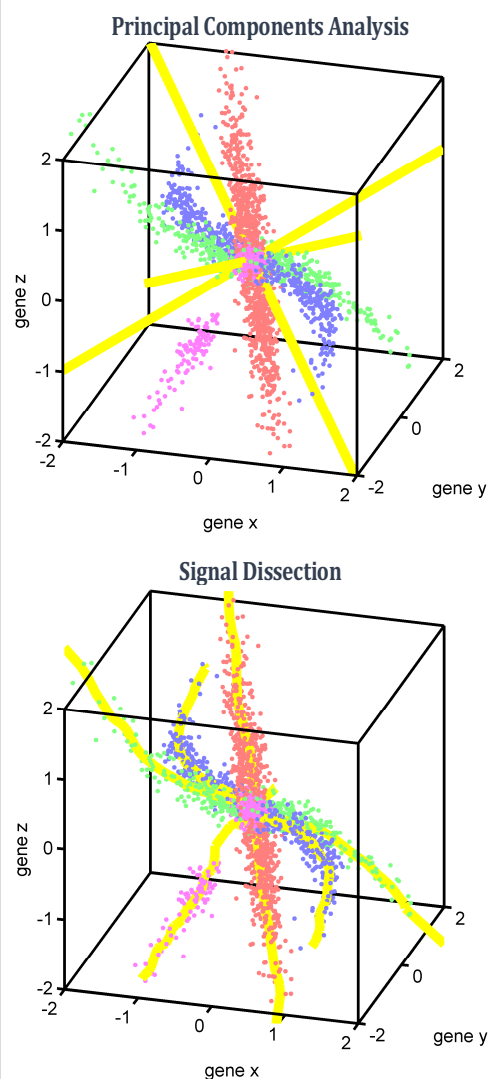


Figure 1.1.4.2.a) 3D concept example with four effects. Misleading principal components and precise effect curves by signal dissection (preview).

Four simulated pathways corresponding to four distinct sample groups have been simulated with different governing laws of regulation for the same three genes. Details on the signal follow in II.1.1.

The upper panel shows all three principal components for this signal; they cannot summarize this signal in an interpretable form.

The lower panel previews all four effect gene curves detected by signal dissection.

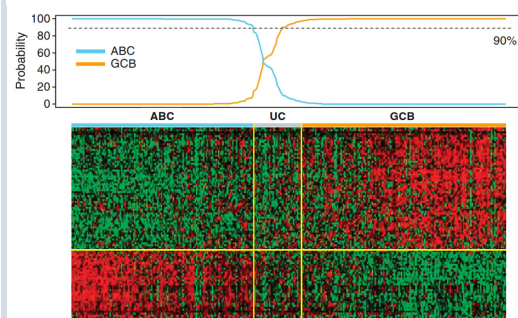


Figure 1.1.4.2.b) Example of a roughly bimotonic real-world effect<sup>[29]</sup>

This heatmap is based on a supervised analysis that sorts samples of cohort GSE31312 based on their differential expressions between predefined gene signatures for ABC-like and GCB-like subtypes of DLBCL. (Adapted by permission from Macmillan Publishers Ltd: *Leukemia*<sup>[29]</sup>, copyright 2012)

purposes. Hence, discovered effects for high-dimensional gene expression signals that follow this effect model should provide an accessible signal summary for further biological modeling of interactions in cells. Other than in the depicted example that is based on a supervised analysis, signal dissection discovers such effects unsupervised and the sum of all dissected effects rebuilds the complete measured signal, except for a noise term.

In the remainder of this chapter, several other methods from the association stage<sup>(I.3)</sup> and for supervised analyses<sup>(I.2.1)</sup> are introduced. They are complementary to unsupervised signal dissection and summarization in the typical research workflow within molecular biology, as illustrated in the stages of science diagram<sup>(Figure I.1.3.5, oben)</sup>. Exemplary analyses from selected co-authored papers are presented. As methods for unsupervised gene expression analysis<sup>(I.2.2)</sup>, hierarchical clustering and PCA are explained.

In Chapter II, the previewed method for signal dissection is presented in detail. After clarifying design goals<sup>(II.1)</sup> and its mathematical framework<sup>(II.2)</sup>, its search strategy<sup>(II.3)</sup> is explained. Subsequently, bimonotonic regression and effect dissection are clarified<sup>(II.4)</sup>. Noise and significance measures are discussed separately<sup>(II.5)</sup>. Among other tests, a versatility test scenario that simulates high-dimensional technical and biological effects of various shapes<sup>(II.6.2)</sup> is utilized to systematically validate the method and to compare it with PCA<sup>(II.6)</sup>.

In Chapter III, signal dissection is applied to gene expression measurements for samples from more than 1000 DLBCL patients<sup>(III.1)</sup>. Resulting gene expression effects are validated across four independent patient cohorts<sup>(III.1.2)</sup>. Validated effects include the rediscovered cell-of-origin effect<sup>(III.3.2.2)</sup> and several genetically novel effects that show significant association with patient survival. Multivariate survival analyses allow construction of a Cox survival predictor that reveals significant survival differences *within* standard DLBCL subtypes<sup>(III.2.5)</sup>. All top survival effects are biostatistically presented and evaluated in detail<sup>(III.3.3)</sup>; they may lead to insights into the molecular pathogenesis of DLBCL.

*A common challenge in many fields of biology is to identify essential differences between measured samples and to relate them with their genetic origin.*

*Generally, two classes of analysis methods are distinguished. Supervised analyses are applied in contexts with some prior knowledge about measured data. They are intended to answer specific questions or to test a particular hypothesis. In contrast, unsupervised analyses follow the aim to reveal previously unknown patterns in measured signals that may help to gain new insights or to infer new hypotheses.*

*Methods of both types are utilized recurrently in typical research workflows in molecular biology. Selected examples that have been analyzed for several published studies are presented here.*

### 1.2.1 Supervised Gene Expression Analysis ---

In supervised cases, it is *already known* what samples need to be compared to identify genes of a specific interest. An example is the analysis of gene expressions of cancer cells after application of a drug versus untreated controls. Alternatively, it may be already known what genes need to be interrogated to answer a specific biological question about samples. Here, several examples of such supervised analyses are presented.

#### 1.2.1.1 Differential expression analyses ---

To quantify the significance of differential expressions of genes between two known settings, various  $t$ -tests of  $\log_2(\text{gene expression ratios})$  can be computed. Expression ratios are typically computed relative to the average expression of all samples in the current context in order to neutralize potential offset effects and to focus on the differences. One has to distinguish between paired scenarios and unpaired scenarios. In the unpaired case, for instance, samples from cancer patients versus (an arbitrary number of) samples from healthy controls is compared with two-sample  $t$ -tests<sup>(e.g. [79], pages 276-279)</sup>. In the paired case, for example, the same cell line has been measured either following treatment with an inhibitor or in untreated form for several time points. In this case, paired  $t$ -tests<sup>(e.g. [79], pages 522-528)</sup> are utilized to focus on the differences induced by the inhibitor. Pairing has the advantage to ignore additional differences that might have biologically occurred or technically incurred between the time points.

As an example of a typical supervised gene expression analysis, the examination of peripheral blood mononuclear cell samples (PMBCs) from renal cell cancer patients (RCC patients) is briefly presented. Genes are depicted<sup>(cf. Figure 1.2.1.1)</sup> that are significantly differentially expressed between these PMBC samples compared to independent control samples from healthy donors. This analysis has been performed for a study to investigate the immunosuppression characteristics of RCC. More precisely, a previously administered vaccine failed to induce clinically relevant immune responses and the aim was to elucidate the molecular mechanisms responsible for that. The biological analysis of differentially expressed genes revealed that already before treatment with the vaccine several genes associated with immune functions are expressed at far lower levels in RCC samples compared to healthy controls<sup>[30]</sup>. Further association analyses and their biological interpretation confirmed this immunological impairment<sup>(e.g. Figure 1.3.1.2)</sup>.

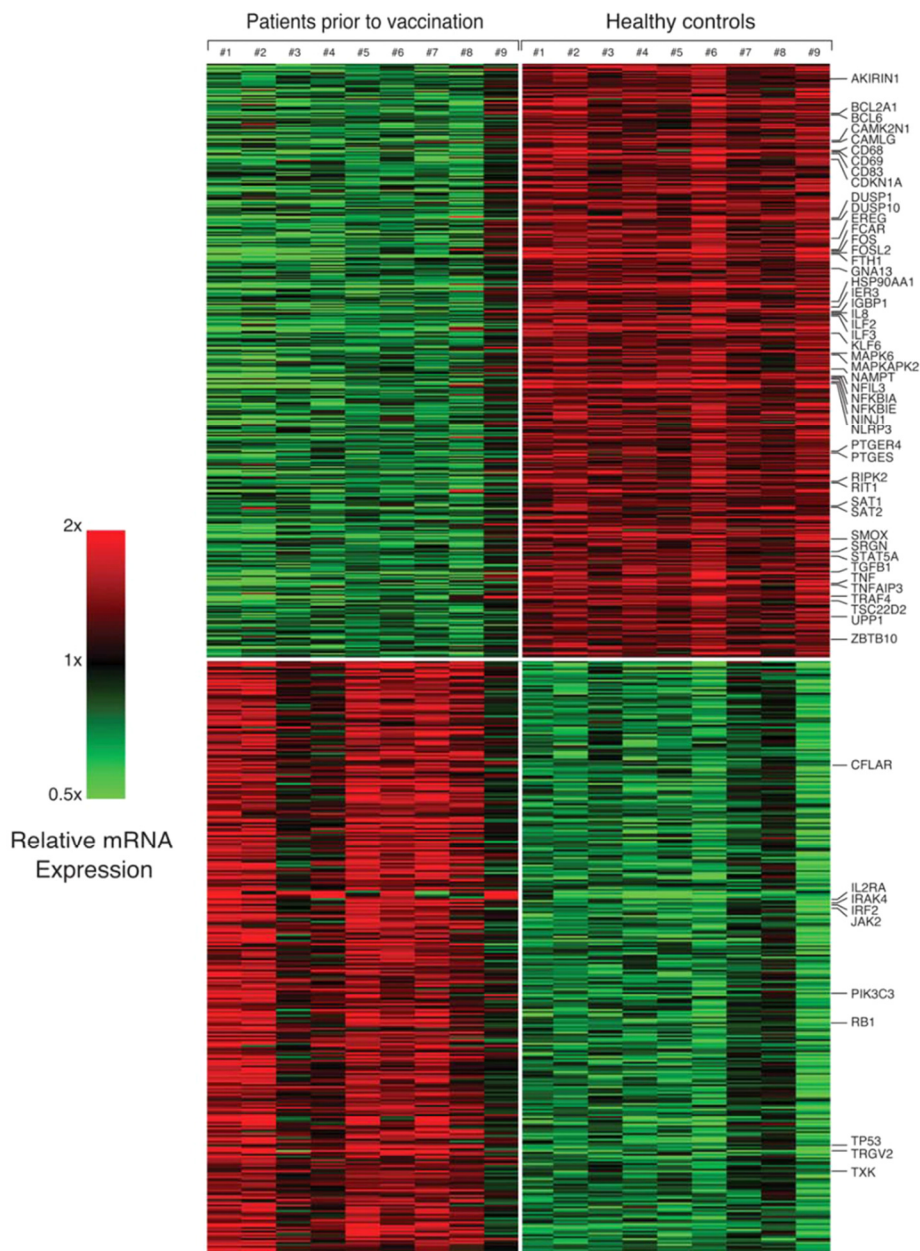


Figure 1.2.1.1) Example for a supervised gene expression analysis that determines significantly differentially expressed genes between two known settings

Depicted are supervisedly determined top upregulated and top downregulated genes in 9 peripheral blood mononuclear cell samples (PBMCs) from renal cell cancer (RCC) patients prior to vaccination (left) compared to 9 healthy control samples (right). All genes with a  $p$  value  $< 0.001$  are depicted (two-sample  $t$ -tests). Labelled genes are involved in immunological processes. (Analysis performed for and reprinted from [30].)

### 1.2.1.2

## Application of specific gene signatures

Once experiment-specific signatures of top-upregulated or top-downregulated genes have been obtained (cf. 1.2.1.1), they may be used to interrogate other experiments for the same genes or to test biological hypothesis.

For instance, most genes that are top-downregulated following the IKK $\beta$  inhibitor MLN120b can be interpreted as NF- $\kappa$ B target genes, as NF- $\kappa$ B signaling is halted by this inhibitor before the phosphorylation of I $\kappa$ B in the classical NF- $\kappa$ B pathway (cf. Figure 1.1.2.2). If another inhibitor candidate is assumed to inhibit NF- $\kappa$ B target genes as well, it should cause downregulation of largely the same genes. To test this, the gene signature for the IKK $\beta$  inhibitor can be applied to experiments with this other inhibitor.

In the example (cf. Figure 1.2.1.2), a PI3K inhibitor is tested, and indeed it significantly decreases expression of NF- $\kappa$ B target genes in two cell lines for four time points. Further experiments in the same study identify a critical role for PI3K (and the downstream kinase PDK1) for viability of a subset of ABC DLBCL cell lines that are characterized by CD79B mutations<sup>[31]</sup>.

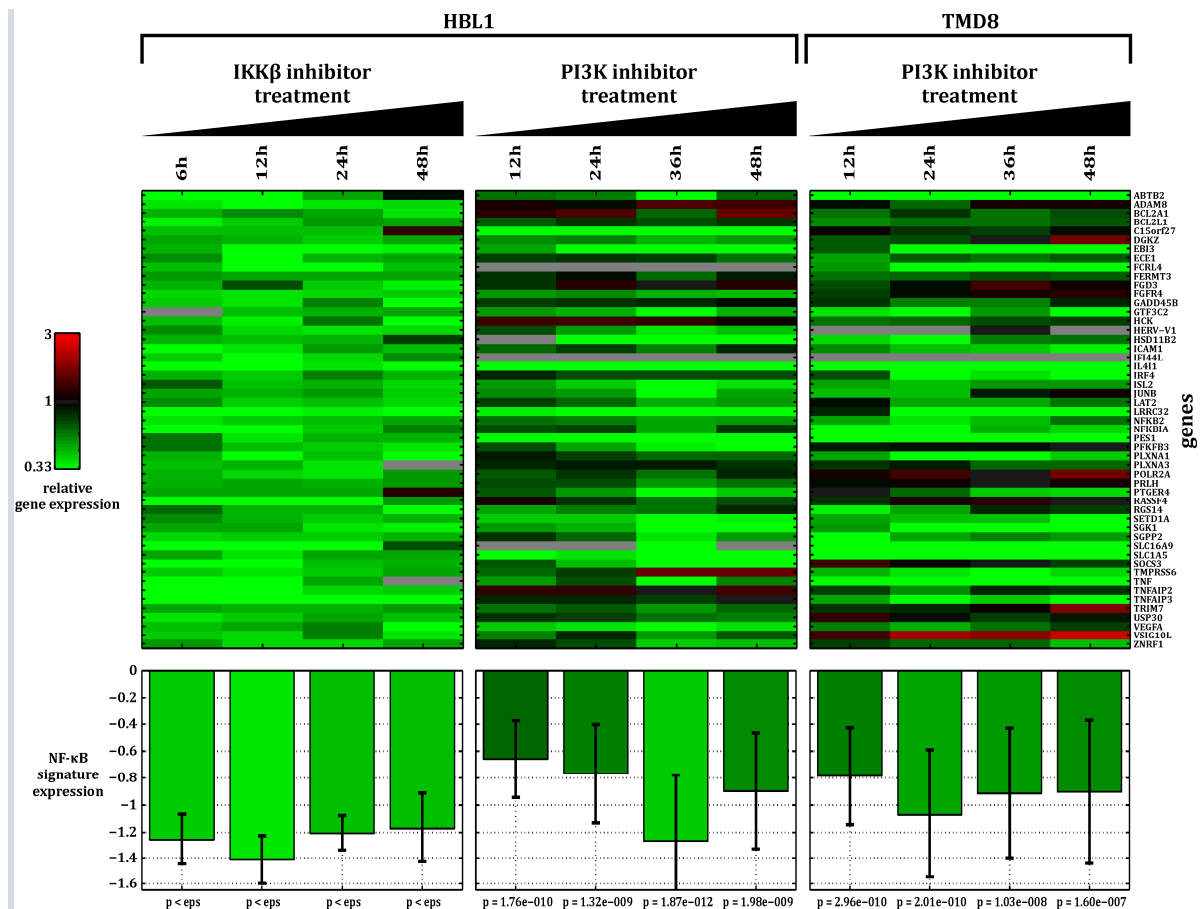


Figure 1.2.1.2) Application of an identified NF- $\kappa$ B signature (left) to gene expressions following treatment with a PI3K inhibitor

Top-downregulated genes for the IKK $\beta$  inhibitor MLN120b (left) have been supervisedly identified based on gene expression profiling for four time points and for the ABC DLBCL cell line HBL1 (selection criteria: at least 50% downregulation for at least three time points). Identical genes are interrogated in an identical experimental setup for a PI3K inhibitor<sup>[31]</sup>. Both the HBL1 cell line in the center panel and the ABC DLBCL cell line TMD8 have been tested. Signature averages are significantly downregulated for both cell lines in all four interrogated time points ( $t$ -tests versus zero regulation). (eps denotes a numeric resolution limit of a previously utilized  $t$ -distribution integration function at  $2.22 \cdot 10^{-16}$ . Gray pixels indicate missing values due to the spot filter.) (Analyzed for <sup>[31]</sup>.)

Similar analyses have been performed for another study<sup>[32]</sup>. Analogous to the IKK $\beta$  inhibitor above, the MALT1 inhibitor Z-VRPR-FMK was already known to interfere with survival of ABC DLBCL cells *in vitro*. However, Z-VRPR-FMK is not adequate for clinical applications, as it needs to be administered in very high concentrations. To identify small molecule inhibitors with more favorable pharmacological properties, top downregulated genes of Z-VRPR-FMK have been determined first (cf. 1.2.1.1). The resulting MALT1 specific gene signature has been subsequently applied to gene expressions following treatment with the phenothiazine derivative Mepazine. Indeed, the MALT1 signature was downregulated significantly four time points as above, but with weaker average folding. Chemically, this might originate from the fact that Z-VRPR-FMK inhibits MALT1 by covalent binding, while Mepazine binds it in a non-covalent and reversible manner<sup>[32]</sup>.

Such targeted applications of biologically selected gene signatures in order to ask specific questions do not need any multiple hypothesis corrections. On the other hand, in case of likewise applications of signatures that were selected e.g. by gene set enrichment analyses based on large signature databases (cf. 1.3.1), corrections for multiple hypothesis tests should be performed (cf. 1.3.1.1).

## 1.2.2 Unsupervised Gene Expression Analysis

In the more general unsupervised scenario, the goal is to discover all yet unknown sample groups (or gene subsets) that show consistent differences in measured signals. In context of a disease for example, gene expressions of samples from many patients might be explored with the aim to identify all disease-characteristic molecular effects of gene regulation.

Unsupervised analyses require completely different methods and are considerably harder to perform than supervised analyses. Mathematically, this difficulty stems from the huge number of theoretically possible subsets of patients and genes that *might* represent biological effects. This number grows like the power set with  $2^m$  respectively  $2^n$ , where  $m$  is the number of genes and  $n$  the number patients. Bulk methods that try to enumerate all possibilities and try to compute some score for each candidate subset are therefore futile for typical application sizes with  $m \approx 20000$  genes and  $n \approx 100$  to  $1000$  samples. Hence, every unsupervised method *needs a kind of search strategy or model for the interactions* it tries to unveil.

A standard method for exploring gene expressions in the search of novel effects is hierarchical clustering. It is utilized frequently in molecular biology and medicine for unsupervised analyses. Another standard method in the unsupervised context is principal components analysis (PCA). As a blind source separation method it reaches conceptually beyond hierarchical clustering, since it does not only reorganize genes and samples but also re-expresses measured gene expressions in new coordinates called principle components. In the ideal case, a principal component represents a biologically specific effect. For example, it may be interpreted as gene regulations caused by a specific pathway. Though utilized relatively seldom in biosciences, PCA has a strong mathematical foundation and many applications, for example in astrophysics. Both standard methods are briefly presented here and problems of both methods are illustrated.

### 1.2.2.1 Hierarchical clustering

This method<sup>[33]</sup> originates back to the 1950s<sup>[34,35]</sup>, i.e. long before the advent of microarray technology. In life sciences it was established more than 15 years ago to analyze correlations in gene expressions<sup>[36]</sup>. Since then it has become a standard method to discover gene signatures or clusters of samples. It has helped to gain many important insights into molecular processes for many organisms ranging from bacteria<sup>[37,38]</sup> and plants<sup>[39,40]</sup> to higher animals like mice<sup>[41,42]</sup>, chimpanzees<sup>[43]</sup> and, of course, humans<sup>[4,44-58]</sup>. Besides the predominant gene expressions<sup>[4,39-48,50-52,54-62]</sup> it has been successfully applied to other measures such as DNA methylation<sup>[53]</sup>, micro RNA expressions<sup>[42,56]</sup>, DNA copy numbers<sup>[49]</sup> and protein concentrations<sup>[37]</sup>. Consequently, it has contributed to a broad spectrum of objectives, e.g. understanding viral or bacterial infections<sup>[43,58,59]</sup>, effects of cigarette smoke<sup>[46,50]</sup> or environmental toxins<sup>[41]</sup>, aging research<sup>[62]</sup>, molecular psychiatry<sup>[56]</sup> or industrial aims like new bioenergy crops<sup>[40]</sup>.

For DLBCL<sup>(1.1.2.1)</sup>, hierarchical clustering has been particularly useful, as it revealed two distinct and previously unknown subtypes, i.e. ABC DLBCL and GCB DLBCL<sup>(cf. Figure 1.1.2.1)</sup>. These subtypes coincide with significantly different patient outcome for the same chemotherapy. Subsequently, this discovery enabled researching distinct pathogenic causes<sup>(1.1.2.3)</sup> for these subtypes.

Conceptually, hierarchical clustering takes a  $m \times n$  data matrix and first computes the distances between each pair of genes (matrix rows) *or* each pair of samples (matrix columns). The clusterings of genes and samples are *independent of each other*. Sometimes only one of the two order dimensions is clustered. Agglomerative hierarchical clustering initially considers all measured points as clusters of size one. Based on their pairwise distances and a linkage method, the nearest clusters are merged to form a larger cluster. Then

distances between centers of all current clusters are computed again, and again the nearest two clusters are merged. This continues iteratively until only a single cluster remains that contains all genes respectively all samples. The remaining cluster is depicted as the root of the dendrogram, i.e. the resulting binary distance tree (cf. Figure 1.2.2.1, unterhalb). The height of the dendrogram depicts distances between connected clusters. A method to *cut* the dendrogram is required to obtain clusters for further analysis and discussion. For example, a manually prescribed distance threshold could be used (i.e. cutting at a constant dendrogram height).

Important functional parameters of hierarchical clustering are the *metric for computing distances* between genes respectively samples and the sub-method of *linkage* that determines how distances between clusters are computed, if they contain more than a single gene respectively sample. One possibility is to use the distance of the two nearest members of two clusters; this method is called single-linkage. Complete linkage compares the farthest members of two clusters. Average linkage, for another example, compares the geometric centers of two clusters with the distance metric.

The default distance metric is the *Euclidean distance* that measures the geometric distance between points in  $\mathbb{R}^m$  (sample columns) respectively  $\mathbb{R}^n$  (gene rows). Another way to measure distances is to utilize *correlations* between points. Compared to the Euclidean distance, distances based on correlations have the advantage of being independent of absolute expression strengths. If for example two genes show the same direction of regulation for all measured samples, but one gene has a much stronger folding than the other, their Euclidean distance would be high, but their correlation-based distance would still be low. Since co-regulation for all samples may already indicate that both genes are controlled by the same pathway, the correlation-based distance is often advantageous in biosciences. Hierarchical clustering with a correlation based distance metric was established for gene expression analyses in 1998<sup>[36]</sup>.

The following example (cf. Figure 1.2.2.1; based on my own implementation) is instructive to explain the interpretation of a typical hierarchical clustering result and to demonstrate potential problems. The heatmap depicts gene expressions for samples from a CHOP-treated DLBCL cohort (data accessible at NCBI GEO database<sup>[63]</sup>, accession GSE10846). An existing signature has been utilized to select an initial subset of all measured genes for this analysis. Hence, this analysis is not completely unsupervised (cf. II.1.2.1).

While orange and green gene clusters show strong expressions, the blue gene cluster shows relatively weak expressions overall and only contains a few strongly expressed genes at the top. But these genes are not aligned in a monotonic way, i.e. they are incompatible to the depicted sample clustering and may be regulated by a distinct pathway. The depicted sample clustering has been mainly determined by the first two gene clusters. The biological specificity and coherence of the analyzed gene signature could possibly be increased by filtering out the blue gene cluster. This is an example of manual focusing in the gene space. If one focused on the orange gene cluster only, a different sample clustering would result, as upregulated samples with respect to orange genes are currently split over two different sample clusters. This demonstrates the element of arbitrariness associated with manual focusing. Hence, it is one design goal for signal dissection (Chapter II) to realize a completely unsupervised effect focusing (see also II.1.2.2).



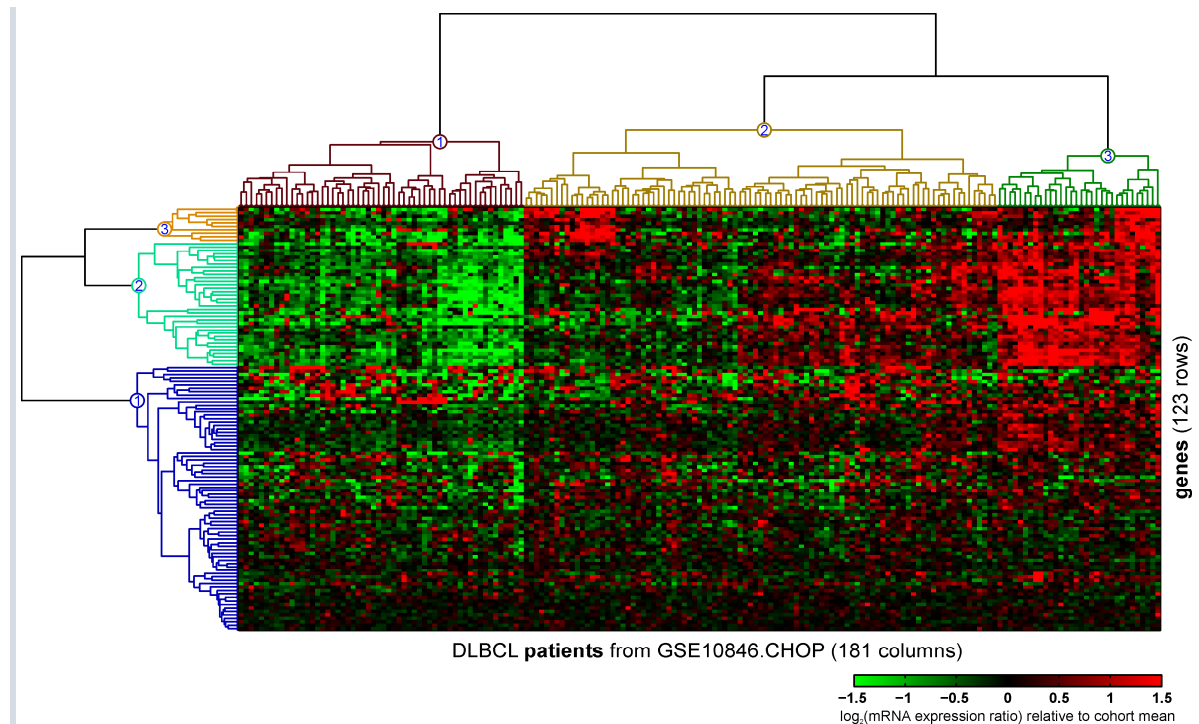


Figure 1.2.2.1.a) Example result from bi-hierarchical clustering

Sample clusters at the top are color-coded by average patient outcome. While for the red cluster 35 deaths were observed (with 21 censored events), in the green cluster only 12 deaths were observed (with 20 censored events). This indicates that underlying genes might be biologically relevant for DLBCL. However, given the visible heterogeneity, these genes may not belong to the same biological function. In particular, non-monotonic signals (like at the top left of the sample cluster in the center) indicate the existence of at least one overlapping gene regulation effect that might give rise to an alternate sample clustering. As described above, it could be revealed by manual focusing on exclusively the orange gene cluster and by another subsequent hierarchical clustering analysis.

However, more complex overlapping structures cannot be represented well by hierarchical clustering, even with manual focusing. The hierarchical clustering illustrated below<sup>(Figure 1.2.2.1.b)</sup> arranges an exemplary subset of gene expressions from cohort GSE31312 (data accessible at NCBI GEO database<sup>[63]</sup>). While clear sample clusters can still be obtained, many substructures are visible that are inconsistent with the overall gene expression trend. Again, manual focusing would result in different sample clusterings, but likewise manual focusing on sample subsets would result in different gene clusters. These ambiguities are all the more present if not just a tiny subset, but the whole signal matrix with over 50000 measured probesets has to be analyzed unsupervisedly. Hence, a concept that does not just rearrange genes and samples, but that can also *dissect* overlapping effects *by modifying the signal itself* seems to be indicated in order to analyze such complex data consistently.

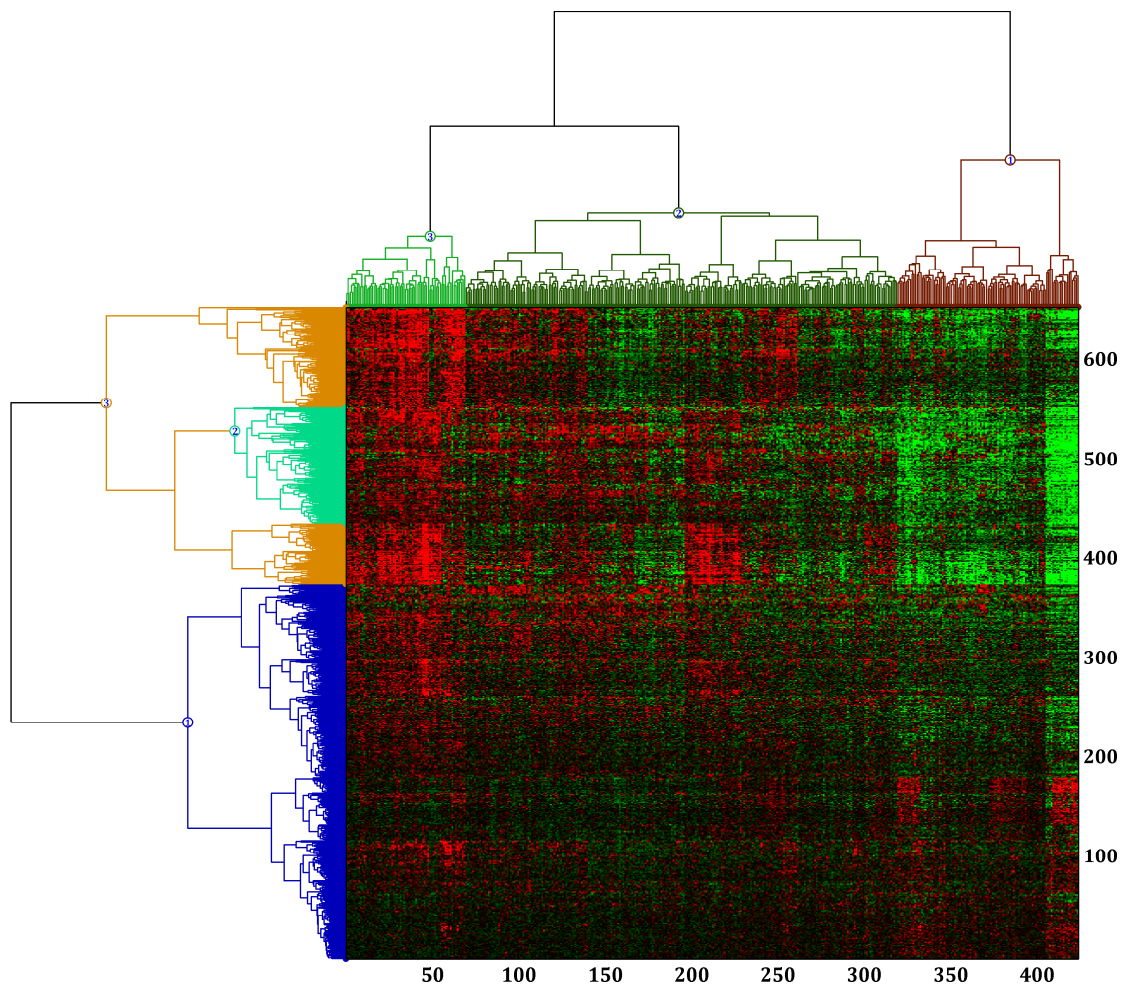


Figure 1.2.2.1.b) Example result from bi-hierarchical clustering for a more complex gene expression signal

### 1.2.2.2

## Principal components analysis (PCA)

Historically, principal components analysis was invented by Karl Pearson in 1901<sup>[64]</sup>. Today, PCA is one of the most successful analysis methods in many areas of science, including e.g. chemometrics<sup>[65]</sup>. For genomic data, PCA has been successfully utilized e.g. for dimension reduction and for visualization of complex data in 3D<sup>[66]</sup>.

Conceptually, principal components analysis goes beyond hierarchical clustering, as it does not merely reorganize genes and samples into distance trees, but models gene expressions themselves as a superposition of expressions along principal components. Essentially, a principal component is a linear combination of genes (i.e. a direction in  $\mathbb{R}^m$ ) that shows maximal variance, i.e. samples have maximally varying expressions along these directions. For biological interpretation, it may be assumed that each such direction corresponds to a pathway that causes co-regulation of its genes along that direction. Additionally, the method reduces co-variance between the principal components to zero, i.e. no two directions are allowed to explain expressions of the same genes in a co-regulated or correlated way. Geometrically this is perceptible by the fact that all principal components are *perpendicular* to each other. This restriction implicitly assumes that biologically distinct pathways cannot contribute to gene expressions in a partially correlated way, and thus may be well separable by minimizing their co-variance. (This assumption is shown to be problematic when trying to

dissect partially correlated effects<sup>(cf. II.6.1.)</sup> Often just a few principal components suffice to explain most of the variance in the gene expression signal; in this way PCA allows summarizing microarray measurements by fewer dimensions (also called dimension reduction).

For demonstration, all three principal components<sup>(yellow)</sup> for a simple 3D signal are depicted<sup>(Figure 1.2.2.2.a)</sup>. Line lengths depict the relative signal variance explained by each principal component.

If simulated points were positional measurements of all mass elements in a *rigid body*, these principal components would have the physical characteristics of *principle axes of inertia*: Applying a torque around either of these axes would cause the body to begin rotating *only* about the respective axis. I.e. when rotating around the axis with maximal variance, all points on that axis (and especially the tips of the rigid body) would *stay where they are*. This is generally not the case when applying a torque around an arbitrary direction. As one principal component already explains the dominant part of the variance in this simulated signal, it can be utilized to summarize the signal for all points. If the same points were reinterpreted as expressions of three genes in 2000 patient samples, the direction of this dominant principal component might represent the law of co-regulation mediated by an underlying pathway. This law (in form of a linear combination) could subsequently be utilized for further high-level modeling of that pathway. This would not be feasible, if the modeler would only have a table with raw data values for these 2000 points, again demonstrating the advantage of summarization for modeling and interpretation.

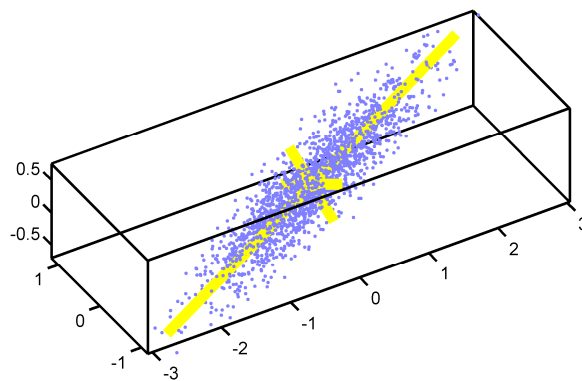


Figure 1.2.2.2.a) Simple 3D illustration of PCA

Points for three dimensions and 2000 samples measure a single simulated effect. Principal components returned by PCA are depicted as three yellow perpendicular lines. Line lengths are proportional to the square roots of the variances, i.e. they depict the standard deviations along these directions.

Mathematically, principal components are found by diagonalizing the covariance matrix of the signal. Every gene can be seen as a random variable  $X_i$  that corresponds to the distribution of that gene's expressions for all samples. (In the above example, each gene  $X_1, X_2$  and  $X_3$  was measured for 2000 samples.) The covariance between two genes  $cov(X_i, X_j)$  is defined as  $E[(X_i - E[X_i]) \cdot (X_j - E[X_j])]$ , where  $E$  is the expectancy operator.  $E[X_i]$  can be empirically estimated by the mean of all sample expressions of gene  $X_i$ . Similarly, empirical covariances  $C_{i,j} \equiv cov(X_i, X_j)$  for every pair of genes can be estimated numerically. As covariances are symmetric, the resulting matrix  $C \in \mathbb{R}^{n \times n}$  is also symmetric (and real-valued). The principal axes theorem of linear algebra<sup>[67, theorem 7.6.3]</sup> states that for every such matrix an orthogonal matrix  $P \in \mathbb{R}^{n \times n}$  exists such that  $D \equiv P^{-1}CP$  is a real-valued diagonal matrix. I.e.  $P$  rotates points in such a way that their covariance vanishes in the new coordinate system (zero off-diagonal elements in  $D$ ). Columns of  $P$  are eigenvectors of  $C$  in original coordinates; they define the principal components (and their directions as depicted in yellow) that span the new coordinate system. Their eigenvalues  $diag(D)$  finally equal the signal's variance along these new axes. Diagonalization of  $C$  is the computationally difficult part of PCA, but can be readily performed by numeric algorithms, for example by singular value decomposition (e.g. implemented by the `svd` function in MATLAB®).

Principal component analyses have already been successfully applied to gene expression signals. For example, gene expressions of a synchronized yeast culture have been measured at different times in the cell cycle relative to an unsynchronized control. It was possible to summarize this signal by just two eigenarrays<sup>[68]</sup> (i.e. by principal components in the samples space). Correlation of samples to these eigenarrays correctly

reproduced the experimental cell cycle setup. Additionally, approximately 40% of the signal's variance could be explained by just two eigengenes (i.e. principal components in the gene space). Correlations to these two eigengenes allowed estimating the role of all genes in the cell cycle<sup>[68]</sup>. In another application, differentially expressed genes between drug-sensitive and drug-resistant cases have been supervisedly identified. The first three principal components for this signal were have been demonstrated to be already sufficient to predict drug sensitivity of most samples correctly<sup>[45]</sup>.

However, compared to the ubiquitous hierarchical clustering, PCA seems to be utilized relatively seldom in the context of unsupervised gene expression analyses. In part, this may be caused by PCA's underlying *concept of interaction*, i.e. to maximize variance per direction and to zero correlation between directions. This concept does not always fit well to effects of gene interactions that are mediated by pathways and observed by gene expressions. The problem can be illustrated (Figure 1.2.2.2.b) by simulating two subgroups of patient samples (red and blue), each driven by a distinct pathway that regulates the expressions of the same three genes. A useful summary here would separate these groups of patients and deliver one axis (i.e. one principal component or one linear law) per simulated linear pathway. Principal components (yellow) however do not reflect directions of the blue and red pathways. PCA cannot find the correct pathway directions here, in part because they are not perpendicular to each other. I.e. they are partly correlated to each other, which is incompatible to PCA's concept of interactions. Hence, principal components cannot dissect the two patient groups, but instead summarize the signal by the two longer yellow directions, i.e. by new coordinates that *mix* both groups. This mixing of distinct pathways is hard to interpret, especially for real world signals that have a much higher number of dimensions (i.e. genes).

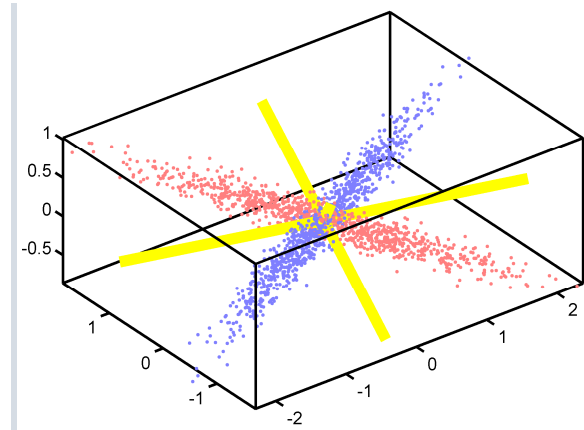


Figure 1.2.2.2.b) Principal components do not point into effect directions  
 Points for three dimensions simulate two distinct but partly correlated (i.e. not perpendicular) effects for 1000 samples each (red and blue). Principal components returned by PCA are depicted as three yellow perpendicular lines. The two longer principal components capture nearly all of the signal's variance here; therefore, the third component that protrudes where the two longer cross in the center is relatively short.

In contrast, if depicted red and blue points would again interact like mass elements in a common rigid body, these principal components would *still* point in the directions of the principal axes of rotation for this X-shaped body, with the above-explained physical meaning. In this context, the red and blue pathway directions would not be as useful for summary for rotational properties, because if a torque would be applied around them, the X-shaped body would inevitably rotate in a way that also moves points near the red respectively blue effect axes, as they are not principal axes of rotation. This demonstrates that the model of interactions that underlies a summarization method should be *compatible* with the signal context for interpretability.

Besides gene expression microarrays, another important source of genomic information is array comparative genomic hybridization (aCGH). Instead of interrogating concentrations of transcribed mRNA, here the relative amount of DNA is quantified. Normally, two copies of each gene are available in the human genome (in case of *X* and *Y* chromosomes in males, the reference is one copy). Segments of DNA may exist that have a lower or higher number of copies. To unsupervisedly detect these segments and to summarize aCGH signals, a neighborhood model can be employed<sup>(cf. Table I.1.3.4)</sup>; for example, circular binary segmentation can detect these segments<sup>(cf. [27], see also `cghcbs.m` of the MATLAB® Bioinformatics Toolbox)</sup>.

These analyses may reveal, for instance, that a cell line has a double-deletion of a specific DNA segment. In this case, genes encoded in this segment can no longer participate in the gene regulation network in affected cells. Such DNA defects may be one possible origin or driver of cancer genesis. An already presented example<sup>(cf. I.1.2.3)</sup> is the loss of *PTEN* in a subset of GCB DLBCL samples<sup>[12]</sup>.

Another exemplary result from aCGH analyses shows<sup>(Figure I.2.3)</sup> amplifications of segments that include the *MCL1* locus:

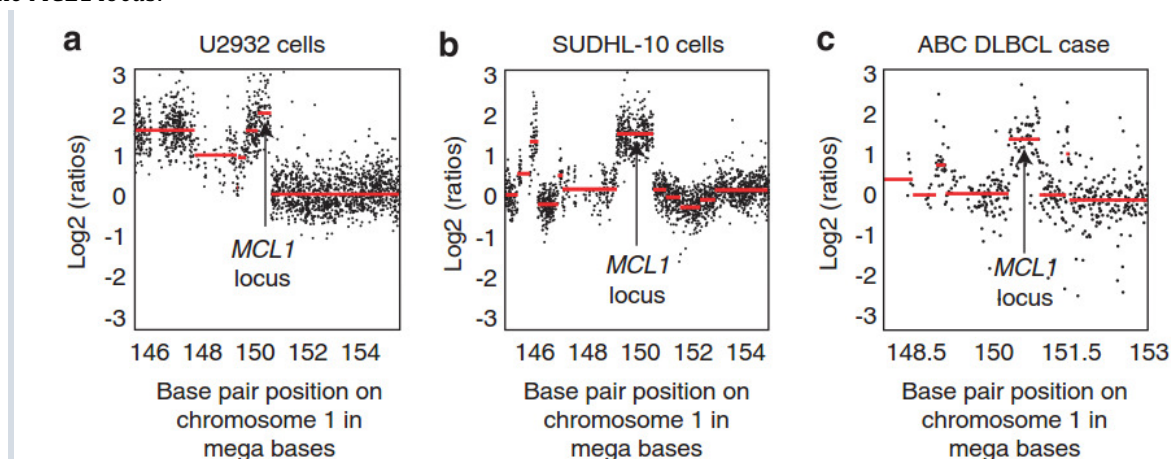


Figure I.2.3) Exemplary aCGH analyses

Two DLBCL cell lines (panels a and b) and an ABC DLBCL case (panel c) are characterized by high-level *MCL1* amplification.

(Analyses performed for and reprinted from a co-authored paper<sup>[3]</sup>.)

Chromosomal gains or amplifications of this locus occur in approximately 26% of ABC DLBCL samples. For this subset, *MCL1* deregulation has anti-apoptotic characteristics and contributes to therapy resistance<sup>[3]</sup>.

Besides whole genome analyses, several other experiments are performed in molecular biology that are associated with other types of analyses. For instance, viability measurements over titrated concentrations of a drug candidate can be utilized to quantify the sensitivity of cell lines.

The drug is typically titrated over a logarithmic range of concentrations for the same cell line several times. The IC50 is defined as the inhibitory concentration where the viability has decreased to 50%. Usually, none of the tested concentrations in the logarithmic range matches the IC50 exactly. Hence, an interpolation between data points is needed to determine it.

Such an interpolation can be determined with a sigmoidal model  $\frac{1}{1+\exp(s \cdot (x-m))}$ , where  $x$  is the logarithmic concentration,  $m$  is the resulting logarithmic IC50 and  $s$  determines the steepness of the viability decrease. To respect that for some cell lines an offset viability may remain after the strong decrease in viability, the model can be added this degree of freedom by  $\frac{1-o}{1+\exp(s \cdot (x-m))} + o$ . This model has been fitted to measured titration curves (with the `fit.m` function in MATLAB®). The IC50 concentration is obtained by the fitted  $m$ . Visually (cf. Figure 1.2.4), the IC50 corresponds to the intersection of the interpolated titration curve (gray) at 50% viability (blue).

The example demonstrates IC50 analyses that have been performed for DLBCL cell lines OCI-Ly3 (upper panel) and TMD-8 (lower panel). Both cell lines have been treated with an antibody drug conjugate (ADC) that targets CD79B. CD79B is physiologically expressed in the vast majority of B cells and thus represents a promising target for DLBCL<sup>[69]</sup>. The cytotoxic agent in this ADC is the microtubule-disrupting agent monomethyl auristatin E (MMAE). It induced cell death in the majority of DLBCL model cell lines, as depicted exemplary for OCI-Ly3 and TMD-8. These ADCs are also clinically relevant<sup>[69]</sup>.

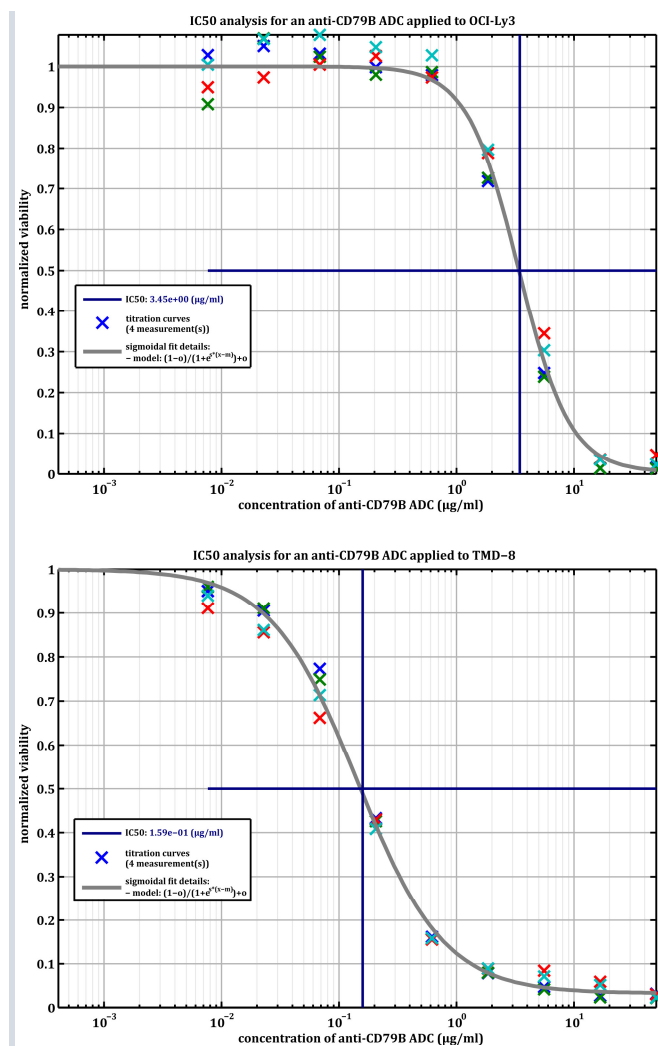


Figure 1.2.4) Examples of IC50 fits for an anti-CD79B ADC

Sigmoidal fits are utilized to determine IC50s for an antibody drug conjugate (ADC) targeting CD79B as described in the text. The same drug is applied to two DLBCL cell lines, OCI-Ly3 and TMD-8. (Analyses performed for <sup>[69]</sup>)

## 1.3 Association Methods

---

*An important step to gain new knowledge from experiments is to associate measurement results with already existing knowledge<sup>(cf. I.1.3.5)</sup>.*

*In molecular biology, important sources of computable genomic knowledge are public gene signature databases. They contain sets of genes that were discovered to function alike. A similar source is the gene ontology. It groups genes hierarchically by known molecular functions, by cellular components or by known biological processes.*

*In context of patient samples, typically clinical information is available in addition to genomic information sources. These information can be of categorical type (like gender) or of metric type (like blood values). They can also be of a time-ordered censored type (like patient survival measured by follow-up studies). Different statistical methods exist for the association of each type of information with gene expressions.*

*Selected examples analyzed for published studies are presented in this subchapter to illustrate some of these association methods. The same methods are applied systematically for biostatistical evaluation of discovered GEP effects<sup>(cf. III.3)</sup>. Results for all 135 validated GEP effects are digitally provided in both tabular and graphical form<sup>(cf. page 231)</sup>.*

### 1.3.1 Signature Analyses and Gene Set Enrichment

---

Usually, genomic discoveries in bioscience have been published in the form of sets of genes that have shown similar behavior in analyzed biological settings. For example, they may have been identified by similar differential expressions between cell types or by co-regulation in response to a drug, etc. These gene signatures have been collected in several large public databases. For a comprehensive association with all existing biological knowledge encoded by these gene signatures, I have imported and combined the following public databases:

- MolSigDB<sup>[70]</sup> (9479 signatures, v4.0, downloaded in May 2014)
- GeneSigDB<sup>[71]</sup> (3138 signatures, v4.0, downloaded in September 2011)
- Staudt lab signature DB<sup>[72]</sup> (253 lymphoma specific signatures, downloaded in November 2012)
- HGNC gene families<sup>[73]</sup> (285 signatures, downloaded in May 2014)

The resulting combined signatures database enables a systematic association of genomic experiments with 13155 known gene signatures from various fields of biology.

#### 1.3.1.1 Gene set enrichment analysis

---

Several statistical analyses can be utilized to test the association of a given biological experiment to these signatures. The probably most relevant statistical method to this end is gene set enrichment analysis<sup>[74]</sup>.

First, genes are ranked by the experiment in focus. Typically a supervised analysis<sup>(cf. I.2.1.1)</sup> determines these gene ranks, for example genes are sorted by their folding from the most upregulated to the most downregulated gene. For a given known gene signature, ranks of genes that are contained in this signature are marked next. Then it is tested, if these marked genes are distributed in a significantly non-random form.

Higher enrichment scores are obtained if all signature genes are located either at the downregulated side or all at the upregulated side of the analyzed experiment. More precisely, if all genes in a signature of size  $x$  are identical with the set of genes that have top ranks  $1 \dots x$ , then the enrichment score equals  $+1$ . If they are identical to the set of bottom-most genes, the enrichment score equals  $-1$ . Enrichments on either side may be biologically interesting, while a signature is typically relatively uninteresting, if its genes are half upregulated and half downregulated, or if they are all weakly regulated (corresponding to middle ranks). In these cases, the enrichment score is lower. For completely random (i.e. uniform) distribution of ranks, the score approaches zero. Gene weights may be utilized, for example, to prevent weakly or insignificantly regulated genes from having overdue impact on the score. The enrichment score is computed as a cumulative rank statistic that can be visualized as an enrichment mountain (e.g. Figure 1.3.1.1). From left to right, the curve increases for every gene in the signature (blue lines) and decreases for every other gene. Both sets of genes are normalized and thus both contributions add to zero. The enrichment score is defined as the extremum of this curve (see [74] for details).

To test the significance of enrichment scores, permutation tests are applied. These tests are a major factor for the computational cost of gene set enrichment analyses. As usually many signatures are tested simultaneously, a false discovery rate (FDR) is additionally computed as control for multiple hypothesis testing. The FDR estimates the ratio of signatures with likewise or stronger statistics that might exist by pure chance due to testing many signatures. Signatures can be categorized biologically. Often only signatures from specific categories are of interest. Hence, FDRs are typically computed separately for each category (rather than for the full database) to respect this external knowledge.

In the study for which the depicted example (Figure 1.3.1.1) has been analyzed, a murine *Eμ-Tcl1* model for the human chronic lymphocytic leukemia (CLL) has been utilized to compare *Eμ-Tcl1* mice with *Cxcr5<sup>-/-</sup> Eμ-Tcl1* mice [75]. The spontaneous tumor development has been followed for both genotypes of mice. Mice without CXCR5 showed a substantially delayed onset of the disease. For analysis, first, differentially expressed genes between six samples from *Eμ-Tcl1* mice and five samples from *Cxcr5<sup>-/-</sup> Eμ-Tcl1* mice have been supervisedly determined (cf. 1.2.1.1). Ranking all genes by their  $p$  values for their differential expressions, the combined signatures database (cf. 1.3.1) has been screened by gene set enrichment analyses. The depicted significant enrichment has been detected for the cell division cycle 2 (CDC2) signature (from the GNF2 expression compendium obtained via the MolSigDB [70]). Genes are significantly downregulated in *Cxcr5<sup>-/-</sup> Eμ-Tcl1* mice. Hence, this signature predicts a proliferative advantage in *Eμ-Tcl1* tumor cells. In total, seven different proliferation related signatures have been associated with likewise significant enrichment. Together with other experiments and analyses, the study clarifies steps of CXCR5-dependent tumor cell lodging and resulting proliferative stimuli to leukemia B cells [75].

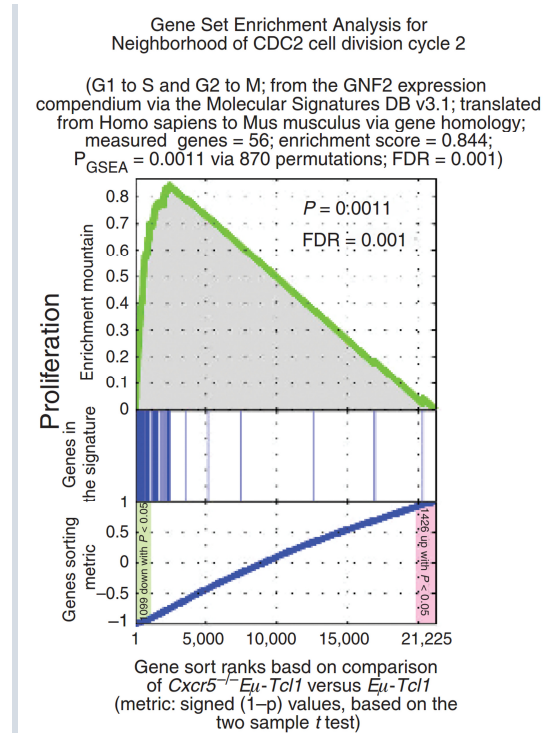


Figure 1.3.1.1) Significant enrichment of a proliferation related signature

The signature CDC2 (from the GNF2 expression compendium via the Molecular Signatures DB [70]) is significantly enriched (see text for details). (Analyzed for and reprinted from a co-authored paper [75].)



A similar analysis has been performed for an investigation of anaplastic large cell lymphoma (ALCL)<sup>[76]</sup>. First, genes that were significantly differentially expressed in cell lines K299, FE-PD and DEL following IRF4 knockdown by RNA interference have been supervisedly determined (cf. I.2.1.1). For the resulting gene ranks, the combined signature database (cf. I.3.1) has been screened and a previously described MYC gene signature has been identified by significant enrichment. Together with other analyses and experiments, MYC has been revealed as a primary target of IRF4 in this study<sup>[76]</sup>.

### I.3.1.2 Additional signature statistics and signature heatmaps

My analysis pipeline routinely computes additional statistics and heatmaps for enriched signatures that depict the actual gene expressions of samples in the current experiment. For each sample, the signature expression is aggregated and tested for significance (via *t*-tests relative to baseline). If these statistics are not significant or if average foldings are relatively weak compared to top genes in the same experiment, results from enrichment analyses should be interpreted with care in my experience (even if they are significant).

For the already briefly presented study in tumor immunology about RCC (cf. I.2.1.1 and [30]), several significantly enriched signatures confirm an immunological impairment. For example, the immune signature “immediate early” (from the Staudt lab signature DB<sup>[72]</sup>) has been identified with an enrichment score of 0.961 ( $p \leq 0.0007$  by permutation test,  $FDR \leq 0.1\%$ ). The depicted heatmap (cf. Figure I.3.1.2) shows significant downregulation of signature genes in the majority of RCC patient samples prior to vaccination relative to healthy control samples. In four RCC samples this effect is particularly strong, leading to significant sample-wise average expressions<sup>(bottom)</sup>, despite the low number of genes in this signature.

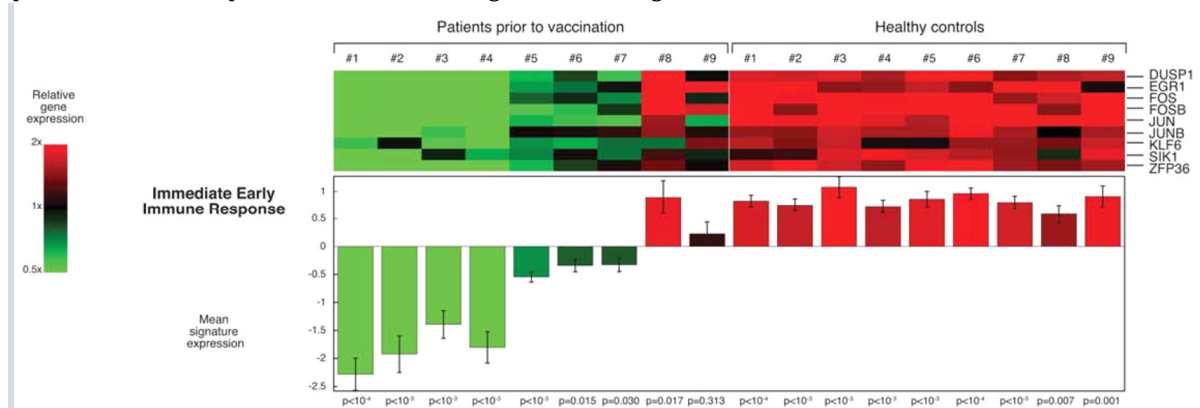


Figure I.3.1.2) A immediate early immune response signature identified by gene set enrichment analyses

The immune signature “immediate early” (from the Staudt lab signature DB<sup>[72]</sup>) has been identified by significant enrichment (enrichment score 0.961 and  $p \leq 0.0007$  by permutation test). The depicted heatmap shows significant downregulation of signature genes in the majority of RCC patient samples (left) relative to healthy control samples. Additional signature statistics are a FDR for the enrichment of  $< 0.1\%$  and a strong average  $\log_2$  (ratio) over all samples of  $-1.67$ .

(Analysis performed for and reprinted from [30].)

These heatmap information complement significant enrichment. Together, such statistics can provide a reliable basis for further biological interpretation of respective signatures.

Gene ontology (GO) terms are similar to gene signatures, but they are organized hierarchically. There are three hierarchy roots: biological processes, molecular functions and cellular components<sup>([77], downloaded in June 2011)</sup>. The deeper a term is in a hierarchy, the more specific is its biological meaning. The online source QuickGO<sup>[78]</sup> provides an overview of the definition of every GO term.

Overrepresentation analyses can associate gene ontology terms with experiment-specific signatures or with discovered genomic effects. Typically, every term is tested for overrepresentation by a hypergeometric test<sup>(cf. [79], pages 369-371)</sup>. Improvements to the statistical analysis that incorporate the parent-child relationship of terms have been suggested<sup>[80]</sup>. For all gene ontology analyses for this work<sup>(presented in III.3)</sup>, I use only direct term annotations of genes in order to focus on the biologically most specific terms.

A disadvantage compared to gene set enrichment analyses is that overrepresentation analyses require a definition of “top genes” for the experiment or effect being analyzed. I.e. a cutoff has to be specified that introduces an element of user-dependency. In contrast, enrichment analyses work with the ranks of all genes and do not require any cutoff.

GO analyses may be useful for the identification of (re)discovered effects of already known biological functions. For example, a relative large GEP effect in DLBCL has been unsupervisedly discovered is significantly associated with gene ontology terms from all three term trees<sup>(cf. III.3.3.4)</sup>. In context of new effects, especially for smaller effects induced e.g. by RNA interference experiments, GO analyses often do not result in significantly overrepresented terms in my experience.

### 1.3.3 Kaplan Meier Survival and Log Rank Tests

Kaplan-Meier survival curves<sup>(cf. [79], pages 760-767)</sup> estimate survival over time from right-censored event data. In case of analyzing overall patient survival, observed deaths are the events and censored events correspond to patients that were lost to follow-up before their death occurred. Not only observed events but also censored events carry important information, as censored patients lived *at least* until the last recorded follow-up time. Resulting Kaplan-Meier survival estimates respect both types of information. Survival curves e.g. for two different subsets of patients, can be compared and tested for significant differences by log rank tests<sup>(cf. [79], pages 767-773)</sup>.

For multivariate survival analysis of combinations of GEP effects, Cox survival models<sup>(cf. III.2.1.1)</sup> are employed. Kaplan Meier survival estimates are then utilized extensively<sup>(in III.2)</sup> to visualize survival differences<sup>(e.g. Figure III.2.5.3.b)</sup> that are predicted by unsupervisedly discovered gene expression effects or by multivariate combinations thereof.

The same analyses can be utilized in supervised contexts to test specific biological hypothesis. For example, a cross-species investigation based on mouse lymphoma models<sup>[81]</sup> has identified a subgroup of GCB DLBCL patients defined by high BCL2 expression. This subgroup is associated with significantly more favorable outcome for high NF- $\kappa$ B expression<sup>(cf. Figure 1.3.3)</sup>. In contrast, constitutively active NF- $\kappa$ B signaling also characterizes the ABC DLBCL subtype<sup>(cf. I.1.2.3, [10])</sup> that is associated with significantly adverse outcome<sup>(cf. Figure III.2.1.8)</sup> relative to GCB DLBCL. Hence, the NF- $\kappa$ B pathway plays opposing roles, depending on the cellular context<sup>[81]</sup>.

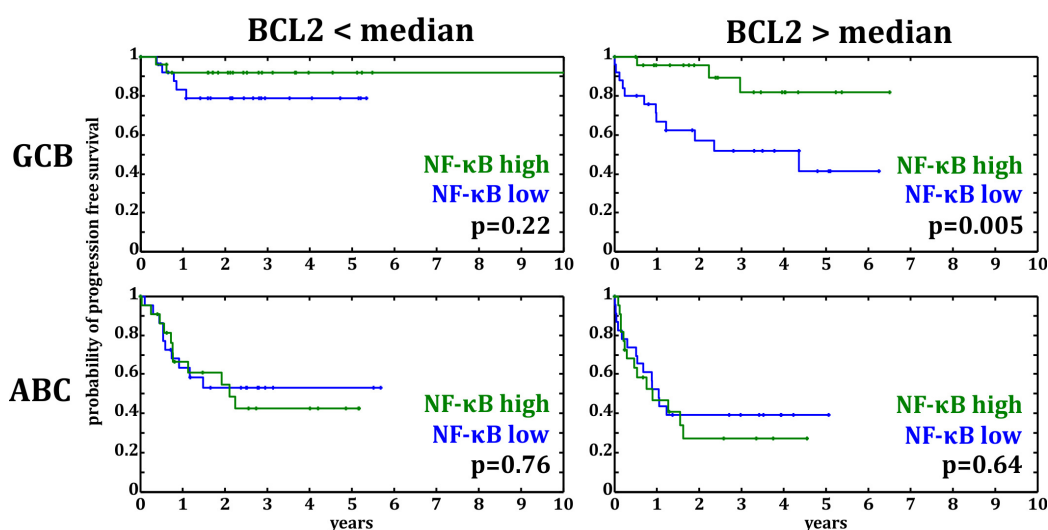


Figure 1.3.3) Kaplan-Meier survival analyses show significantly more favorable outcome for high NF- $\kappa$ B expressions in a GCB DLBCL subset with high BCL2 expression Kaplan-Meier survival analyses for samples in GSE10846.R-CHOP<sup>(cf. III.1.1.1, [5])</sup> are depicted for eight sample subsets. First, samples have been stratified by their BCL2 gene expression and by their previously assigned DLBCL subtypes<sup>[5]</sup>. For each of the four resulting patient subsets, patients have been grouped further by their NF- $\kappa$ B expression (as quantified by the average expression of the signature “NFkB\_Up\_all\_OCILy3\_Ly10” from the Staudt lab signature database<sup>[72]</sup>). For GCB DLBCL patients with high BCL2 expression, a significantly more favorable survival was identified for high NF- $\kappa$ B expressions ( $p=0.005$ , log rank test, 49 samples). (Analyzed for <sup>[81]</sup>)

Interestingly, BCL2 has also been *unsupervisedly* (re)discovered based on signal dissection results and multivariate survival analysis as one of the top DLBCL survival effects<sup>(effect validation index  $v=75$ , cf. III.2.5)</sup>. Biostatistical analyses for this effect<sup>(cf. III.3.3.7)</sup> help to determine a subset of patients with *low* BCL2 expressions and high expressions of a T cell co-stimulation related effect<sup>(cf. III.3.3.6)</sup>. This discovered subset is associated with significantly different survival for different expressions of a zinc related effect<sup>(cf. III.3.3.8)</sup>.



# Chapter II

---

## Signal Dissection

*Modern technologies can measure many parts of a system simultaneously. For example, human whole genome microarrays provide approximately 50000 gene expressions per tumor sample. To bridge the gap between information about all system parts and the high-level modeling of their interactions, there is an increasing need to summarize these measurements.*

*In principle, a signal can be separated in many ways like a sum into summands, but a specific dissection is required to identify summands that represent distinct true effects of interactions in the underlying system. An ideal summary represents the complete high-dimensional multi-sample signal by as few as possible, yet easily interpretable and statistically significant effects.*

*This chapter presents a novel signal dissection method that utilizes a correlation maximization principle and bi-monotonic regression to this end. Various biological effects are simulated to validate the method and to systematically compare it with principal component analysis and hierarchical clustering.*



To exemplify a basic detection task, first, a simulated 3D signal is presented. It serves throughout this chapter for geometric illustrations of method concepts like effect curves that are difficult to imagine in high-dimensional signal spaces.

When developing a novel algorithm it is instructive to reflect existing problems that ideally should be solved. To this end, characteristics of gene expression signals or requirements by the analytical workflow in life sciences are reflected, together with present problems of hierarchical clustering or principal components analysis. Corresponding conceptual improvements that are implemented by signal dissection are briefly previewed.

### II.1.1 Introductory 3D Example

Although the method is designed and optimized for high-dimensional signals, an example with just three dimensions can be more illustrative. It provides a geometrical perception of how the method works, based on visualizations in the familiar 3D space. Furthermore, the simulated signal mimics four basic biological regulation concepts. High-dimensional effects that resemble real-world gene expression signatures with many genes are simulated later for method validation and systematic comparison<sup>(II.6.2.1)</sup>.

#### II.1.1.1

#### A simple linear effect

In the easiest case, genes controlled by a pathway are co-regulated in a linear way, i.e. samples follow a line in gene space. The red effect on the right simulates this using a gene axis of  $\langle e_i^g | a_{\text{red}}^g \rangle \equiv (-0.2, 0.2, 1)_i$  for  $i = x, y, z$ , i.e. gene  $z$  is the top gene of that pathway with gene  $y$  being five-fold weaker co-regulated and gene  $x$  being also five-fold weaker, but regulated with opposite sign. If  $s$  denotes the average pathway strength or activity in samples, then sample points follow the effect via  $s |a_{\text{red}}^g\rangle^0$ , where  $|a_{\text{red}}^g\rangle^0 \equiv |a_{\text{red}}^g\rangle / \|a_{\text{red}}^g\|$  is its normalized *gene axis*. To construct a typical gradual effect for which extreme expressions are less probable than weaker regulations, simulated pathway strengths  $s$  for all depicted 1000 red samples<sup>(Figure II.1.1.1)</sup> follow a normal distribution  $\sim \mathcal{N}(0,1)$ ; this also explains the solitary outlier at the bottom. To simulate measurement noise, each sample point was added normally distributed expressions of standard deviation 0.1 for each gene.

Real-world gene expressions are usually analyzed as  $\log_2(\text{ratio})$ s relative to control samples or relative to the average of all measured patients. Hence, zero indicates no pathway activity *relative to this baseline*, positive values represent *up-regulated* genes and negative values represent *downregulated* genes relative to baseline. In general, a pathway may simultaneously modulate gene expressions towards upregulation for a set of genes (here  $\{y, z\}$  with  $s > 0$ ) and towards downregulation for another set of genes (here only  $\{x\}$  with  $s < 0$ ). For samples with negative pathway activity  $s < 0$ , “up” and “down” regulation change roles. Whether  $|a_{\text{red}}^g\rangle$  or  $-|a_{\text{red}}^g\rangle$  is the positive direction of a

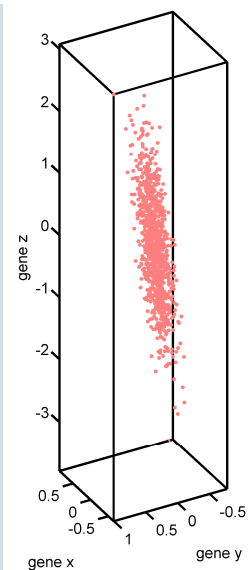


Figure II.1.1.1) 3D example; red effect: a linear law of gene regulation

All three genes are regulated linearly. Gene  $z$  dominates with five-fold higher regulation strength as the other two genes. (1000 samples were simulated as explained in the text.)

pathway is a matter of convention and cannot be decided by the method. In any case, both sets of genes are anti-correlated to each other by virtue of the underlying connecting regulation law.

### II.1.1.2 Supported nonlinear biological effects and three examples

Although the assumption of linearity is a standard first attempt for modeling in physics and biosciences, nature is not always linear. For example, feedback loops in pathways that may lead to nonlinearities and maybe even oscillations over time.

In case of gene expressions, a snapshot of average mRNA expressions in a cell ensemble is measured, i.e. no dynamics over time are recorded. To see nonlinearities like oscillations I such signals would likely correspond to an unnatural effect. For example, consider a subset of samples that shows co-regulation for a set of pathway genes, while one half of these genes are regulated stronger than the other half. Additionally, there are other samples with stronger average activity of the same pathway, but for them the initially weaker half of genes is now regulated stronger than the other half. For samples with even stronger pathway activity, the relative expression of the gene halves is flipped again and thus equals the initial set of samples with relatively low average pathway. Such oscillatory behavior seems biologically implausible in the gene expressions context. Even if such an effect was measured, it would probably be more constructive for system modeling to detect both halves of genes separately as two partially correlated effects. Besides, no general-purpose method that is able to detect and properly dissect superposed effects of all imaginable nonlinear forms can exist, because this task is highly ambiguous and has no unique solution.

Still, there are non-linear biological effects in typical gene expression signals. One example are pathways including genes that reach their *saturation* expression before other genes regulated by the same pathway (i.e. they reach their maximum expression for lower average pathway activity). Other pathways may regulate some genes *sub-proportionally* or *over-proportionally* (e.g. quadratically) relative to their overall pathway activity. Yet others might show a switch behavior, e.g. some of their genes are either switched off or immediately expressed on a constant high level for some *activity threshold*. All these nonlinearities have one thing in common: They are *monotonic* over the average pathway activity. And they extend along a guiding *linear* gene axis, over which they are monotonic. Projections of sample points on this gene axis quantify their average pathway activities. Three basic examples illustrate such monotonic effects:

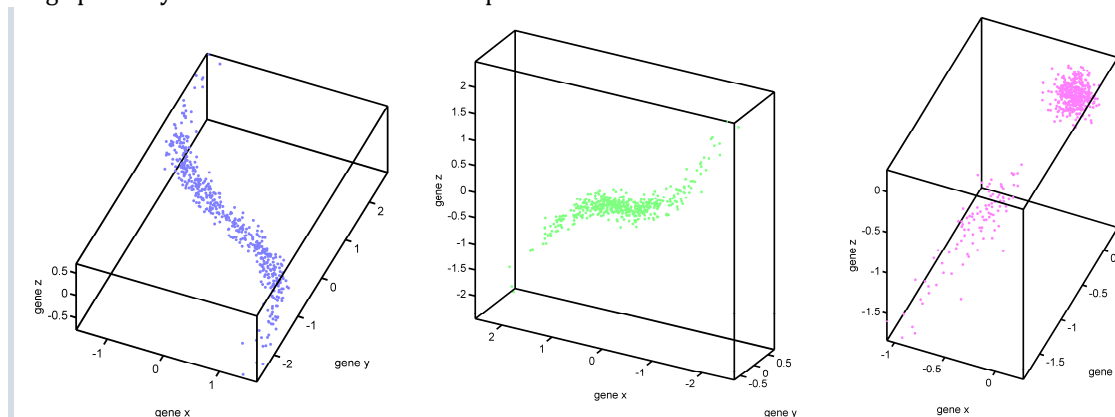


Figure II.1.1.2) 3D example; blue, green and magenta effects: monotonic non-linear laws of gene regulation

Blue effect) Gene *x* saturates over the average pathway activity proportional to the logistic function, while genes *y* and *z* follow linear laws (500 samples).

Green effect) Gene *z* shows quadratic increase respectively decrease, while genes *x* and *y* follow linear laws (500 samples); top point densities are relatively low here.

Magenta effect) A one-sided effect with an offset that simulates a pathway with a threshold activity (500 samples, but only the top 100 have nonzero signal before adding simulated noise).



Precise definitions are as follows:

- The blue saturation effect uses  $\langle e_i^g | a_{\text{blue}}^g \rangle \equiv (-0.5, 0.75, 0.2)_i$  as its dominant linear gene axis, simulates 500 activities  $s \sim \mathcal{N}(0,1)$  and uses a centered logistic function  $x \equiv 3 \cdot \left(-0.5 + \frac{1}{1+\exp(-4s)}\right)$  to simulate a saturation of gene  $x$ .
- The green effect extends along  $\langle e_i^g | a_{\text{green}}^g \rangle \equiv (-0.8, -0.2, 0.5)_i$ , again simulates 500 samples with activities  $s \sim \mathcal{N}(0,1)$  and regulates gene  $z$  quadratically via  $z \equiv \text{sign}(s) \cdot s^2$ .
- Finally, the magenta effect extends linearly along gene axis  $\langle e_i^g | a_{\text{magenta}}^g \rangle \equiv (-0.4, -0.7, -0.7)_i$ , but only the top 20% of all 500 simulated pathway activities  $s \sim \mathcal{N}(0,1)$  are kept; all others are set to zero in order to mimic a one-sided threshold behavior. To simulate measurement uncertainties, all simulated sample points are added normal noise with standard deviation 0.1 for each gene.

### II.1.1.3 Merged 3D signal and an exemplary detection task

In this example, all four simulated effects take place in different sample subsets that could represent disease subtypes. Initially, it is unknown, which sample belongs to which subtype. Neither is known how many effects are contained in the signal. Like in a real measurement setting, only the merged signal (without color-coding) is available as input for detection (Figure II.1.1.3). The task is to detect and dissect all four effects correctly. Furthermore, empirically discovered *laws of gene regulation* should be provided for each one in an interpretable form. More precisely, the method should yield an effect curve in the space spanned by genes  $x, y$  and  $z$  for each effect, together with correlations and effect weights that assign samples to subtypes.

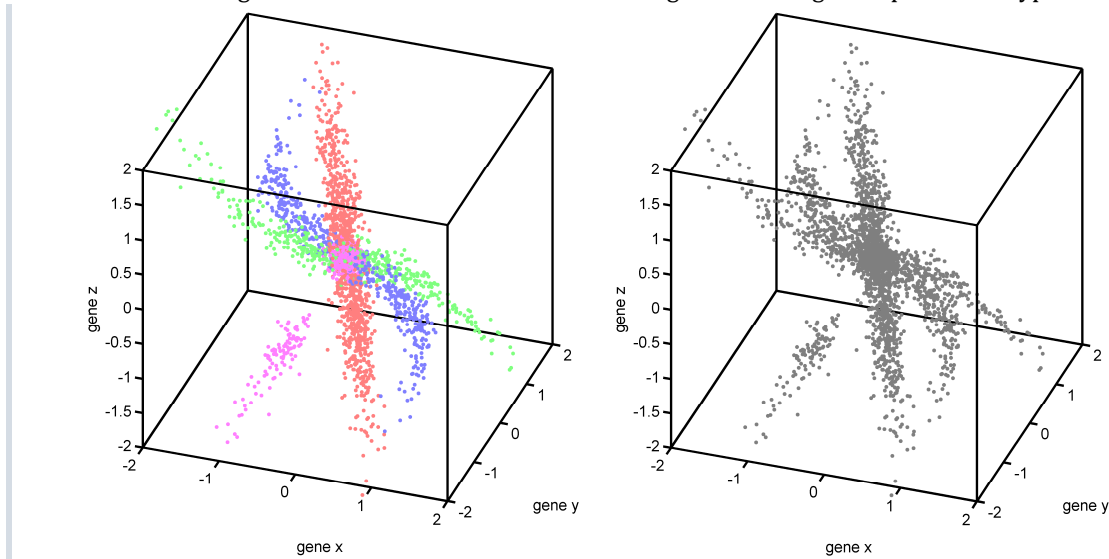


Figure II.1.1.3) Merged 3D example with four effects

Four simulated basic pathways correspond to four distinct sample subsets. They are regulated by different laws for the same three genes, as defined above. (Sample points are plotted in a perspective that has been chosen to show all effects with as few as possible overlap.)

The task for detection methods is to recover all four simulated pathway laws empirically from the points. (The color-coding is not part of the input for detection methods, as illustrated on the right.)

Although based on only three dimensions, this example is difficult to dissect from a certain conceptual perspective: The signal contains four effects, but its gene space has only three dimensions. Mathematically consequently and visually obvious, these effects are *partially correlated* to each other. This also implies that they cannot be dissected by usual projections, because after three projections in 3D only zero remains. In high-dimensional gene expression measurements, pathways usually only regulate some shared genes, but not all of their genes are shared. This makes their dissection easier. Still, visually the 3D effects are clearly separated by their point density. In addition, their guiding linear gene axes point in clearly distinct directions, despite their partial correlation. This can be utilized to dissect them as is illustrated when explaining the search strategy (II.3).

---

Several characteristics of signals in life sciences are listed in this subchapter. Associated problems from methodological perspective are presented. Primarily, examples based on hierarchical clustering are provided, as this method is the quasi-standard for unsupervised analyses in biosciences<sup>(cf. I.2.2.1)</sup>. Additionally, problems are presented in the context of principal components analysis<sup>(I.2.2.2)</sup>, as PCA is conceptually more comparable to signal dissection. Corresponding method design goals are clarified and method concepts reaching these goals are briefly previewed.

---

Whole genome microarrays measure many different effects, including obvious differences like gender or race. Hence, a single sample clustering cannot summarize all measured sample characteristics simultaneously, as most effects are independent and their signals may overlap.

Per design, hierarchical clustering tries to derive just a single sample clustering based on a simultaneous comparison of all genes in its input signal. This usually leads to noisy and unreliable results when applied to the complete measurement signal. To some extent this is still the case for too loosely focused signals<sup>[e.g. 55.figure 1]</sup>. To overcome this problem, usually the analysis is focused on a tight subset of measured genes only<sup>[44,47,48,51,54,57,60,61]</sup>. For example, it might be focused on the most variable genes in a given context or with help of a previously defined gene signature. However, these interventions make resulting clusterings *dependent on the chosen focus* and thus are an incomplete summary with respect to the complete signal.

In brief, a direct application of the method to the full measurement signal should be possible and should not require any supervised restrictions of the gene or sample space, as they can make results incomplete and maybe even biased.

Signal dissection can be applied to the full signal, as its search strategy<sup>(cf. II.3)</sup> realizes an automatic, unbiased and adaptive focusing of effects.

---

Even after an initial (either supervised or unsupervised) general focusing, typically several yet unknown disease-specific effects remain in the focused signal. Often their underlying biological pathways still overlap each other independently like ordinary human features such as gender, skin color and blood groups, i.e. they do not stand in any hierarchical context to each other. To dissect these biological effects properly, one ideally needs an *individual* sample clustering *for each* independent true effect.

Conceptually, hierarchical clustering tries to organize samples in a tree structure that is simultaneously based on *all* remaining effects in the focused signal. The resulting sample clustering then may constitute an automated mixture of biologically yet unidentified and potentially independent effects, which makes resulting dendrograms<sup>[as in 44,47,60]</sup> harder to interpret and hard to compare. One might extract sub clusters from a larger dendrogram to finally arrive at single effects<sup>[e.g. 47]</sup>, but this is only possible if those effects were not overlapped and split by other effects in the first place. Furthermore, these results are dependent on yet another supervised focusing decision. It is legitimate and often statistically significant to just use the dominant effect, i.e. to cut the dendrogram at a top level and to define just few sample clusters<sup>[e.g. 61]</sup>, but this is again systematically incomplete. In the extreme case of manual focusing, one might use external information to focus directly on the context in question, for example, on genes that are differentially expressed in tumor samples that have different drug sensitivity<sup>[45.fig.2]</sup>. This may be useful to construct predictive models, but it is only available in

*supervised* scenarios and thus cannot reveal any previously unknown biological groups of samples or genes as a matter of principle.

PCA does not try to represent overlapping effects as a hierarchical distance tree, but it assumes that effects correspond to directions of maximal variance. This can also lead to a mixture of multiple true effects by each principal component as illustrated in I.2.2.2. This is also confirmed in method validation<sup>(II.6.3)</sup> and again makes principal components hard or even impossible to interpret. This problem is also known in astrophysics, a field where PCA has been applied more frequently compared to bioscience: “The main difficulty with PCA is that the interpretation of the empirically determined PC components in terms of physical properties is complex at best.”<sup>[24]</sup>

Rather than coping with signal summaries that mix real effects, it is usually more constructive to first unsupervisedly detect and identify *each single effect separately* and then let an expert combine them later to generate an interpretable compound model, for example by multivariate survival analyses<sup>(III.2)</sup> based on discovered gene expression effects.

In brief, biologically independent effects should not be mixed automatically in any form as this makes results hard to interpret and difficult to compare.

Signal dissection uses a general superposition model for the signal<sup>(cf. II.2.1.1)</sup> that supports effect hierarchies, but is not limited to them and also supports independently overlapping effects. A correlation maximization principle is utilized by the search strategy<sup>(II.3)</sup> to detect and robustly identify effect axes that are specific to distinct true effects.

### II.1.2.3 Partial correlations should be properly resolved

---

Special cases of overlapping effects are partially correlated yet distinct effects, e.g. the four effects in the 3D concept example<sup>(Figure II.1.1.3)</sup>. Methods like PCA that require orthogonality of resulting effect axes (i.e. principal components) cannot properly resolve such partially correlated effects<sup>(cf. II.6.1)</sup>. The same is true for all methods that are based on (or can be equivalently formulated as) a series of projections of the signal.

Especially in context of genetically heterogeneous diseases like DLBCL, it is important that the method is able to dissect overlapping and partially yet not fully correlated sets of genes. Underlying distinct biological pathways can then be inferred from resulting subsets of highly correlated genes. This inference is considerably harder, if a single and relatively large set of only moderately correlated genes is returned that corresponds to a mixture of these partially correlated effects.

In brief, methods should not be restricted detecting only 100% uncorrelated effects (i.e. to orthogonal effect axes), because this leads to hard to interpret summaries of signals containing partially correlated yet distinct true effects.

Dissection strengths<sup>(cf. II.4.2.1)</sup> are utilized to precisely dissect the signal parts that are most correlated with the respective discovered effect, thereby keeping distinct or only partially correlated signal parts for later discovery as separate effects. These separate effects are usually correlated higher to these signal parts and thus represent them with more specificity, which makes their interpretation easier.

### II.1.2.4 Symmetry of genes and samples

---

In general, a biological effect causes expressions in an initially unknown subset of genes and in an initially unknown subset of samples. As outlined above<sup>(II.1.2.1, II.1.2.2)</sup>, it is useful to focus on a gene subset in order to determine affected samples. Analogously, it increases precision to focus only on the affected samples when

trying to determine participating genes in the first place. Both focusing needs are interdependent and therefore should be handled *simultaneously*.

Hierarchical clustering can either cluster samples based on their expressions of all presented genes or it can cluster genes based on their expressions in all presented samples. Often both resulting dendrograms are displayed simultaneously, but they *can only be computed independently from each other*, as hierarchical clustering can only compare objects (i.e. genes or samples) of the same type at a time (and only with a predetermined fixed number of features). It may be possible to first compute gene clusters based on all samples (ignoring that effects may only affect a subset of them) and then utilize the resulting dendrogram to focus within the gene space, i.e. to compute sample clusterings separately for each identified gene cluster. However, this iterative procedure assigns different roles to genes and samples and makes results additionally dependent on the sequence of those roles.

In brief, genes and samples should be treated symmetrically and simultaneously by the method in order to optimally focus on an effect.

Signal dissection realizes this symmetrization immediately after selecting an initial representative gene or sample<sup>(cf. II.3.1.3)</sup> for a potential effect. Henceforth, roles of genes and samples are interchangeable during detection to avoid any bias. Effect eigensignals are modeled<sup>(cf. II.2.1.2)</sup> and determined<sup>(cf. II.4)</sup> as two-dimensional functions, thereby detecting and representing correlations between genes and between samples simultaneously.

#### II.1.2.5 Removal of overlapping strong effects

---

Gene expression experiments can only measure the sum of all superposed effects, and stronger effects may shine out weaker but more important ones. Strong effects may be of biological origin, for example racial differences<sup>[57]</sup>. Often they are unintended cohort-specific lab effects of unknown cause; some of which may be identified retrospectively, for example differences in experimental labeling protocols<sup>[5]</sup>.

Hierarchical clustering effectively splits the cohort by the strongest effect, for example by cell subtype<sup>(e.g. B- or T-cell based ALL; cf. [45.sup.fig.1])</sup>. Consequently, parts of independently overlapping weaker effects are only detectable *as sub-clusters* within each cluster of the strong effect, making true causes of the weaker effect harder to infer, as demonstrated in II.6.2.6. Furthermore, in such a case the effective cohort size is reduced, eventually preventing detection of biologically more meaningful yet smaller and weaker effects, e.g. gene expression effects with prognostic value<sup>[45.fig.2]</sup>. PCA on the other hand, usually can handle a single overlapping strong and linear effect well, as such effects define a clear direction of maximal variance due to their strength<sup>(cf. II.6.2.5)</sup>. The information along this direction is then projected away by the first principal component, which makes initially overshadowed effects detectable by further principal components. However, detection precision breaks down in presence of more than one strong effect of similar size<sup>(cf. II.6.3)</sup>.

In brief, unimportant strong effects should be prevented from dominating results and they should be removed without affecting information from weaker yet potentially more important effects.

Signal dissection detects the empirical eigenorder<sup>(cf. II.4.1.2)</sup> of the strong effect and utilizes bimonotonic regression<sup>(cf. II.4.1.3)</sup> to estimate signal parts caused by it. This enables its later dissection without losing information about overlapped weaker effects. Later detection iterations can thus detect these weaker effects based on all samples and without any artificial reduction in the sample size due to the strong effect. As large overlapping effects often are just uninteresting lab effects, their removal can also be seen as a type of advanced signal normalization that is complementary to global signal preprocessing (e.g. by quantile normalization).

Most biological effects are gradual in nature, i.e. some samples are affected stronger than others, and their gene expressions effectively sort them in a gap-free continuous way. The same holds for genes participating in an effect.

Usual gene signatures or sample clusters are represented as *flat sets* with hard borders that were determined by e.g. partitioning methods or were inferred from dendrograms of hierarchical clustering. This representation is a consequence from the underlying cluster concept that assumes high similarity within clusters and large gaps between clusters. Gradual effects cannot be represented well by this concept and cut points thus become unreliable and hard to reproduce.

PCA results are more precise relative to flat gene signatures, as principal components can describe directions along which linear gradual effects extend. The coordinates of samples along this principal component then reflect gradual effect strengths, i.e. average pathway activities. However, gradual biological effects like saturations<sup>(cf. basic examples in II.1.1.2)</sup> that deviate from a linear axis cannot be properly described by a single principle component. Instead, the effect's non-linear signal part is represented by one or several perpendicular principal components. Again, this makes it hard to infer the effect's underlying biological law of gene regulation from PCA's signal summary.

In brief, flat sets and hard cuts are an unsuitable model for gradual effects and should be replaced by more flexible forms of effect representation that can also describe non-linear gradual effects without artificially splitting them.

Signal dissection uses a generic bimonotonic effect model<sup>(cf. II.2.1.2)</sup> that supports gradual effects. In the special case that the effect is truly of binary nature (and thus could be represented by a traditional hard cut as well), this information is retained in form of a steep signal change in the regressed effect curve<sup>(cf. II.2.2.3 and II.4.1)</sup>.

It is a common yet difficult question, where one effect ends and another one starts and thus how many effects make up a signal in total. The same question occurs when deciding if a given signal part still makes up an effect or is merely noise.

Hierarchical clustering just returns a dendrogram (i.e. a distance tree) and leaves it to the user to read clusters from it, i.e. to define the number of effects and their borders. This can strongly influence results<sup>[82]</sup>, making sample clusterings harder to reproduce and harder to compare. This problem is amplified in the context of gradual effects<sup>(II.1.2.6)</sup> or when hierarchical clustering is utilized iteratively to realize a manual effect focusing as explained above<sup>(II.1.2.4)</sup>.

PCA returns a full orthonormal rotation of the coordinate system underlying the signal space, i.e. it always returns as many principal components as there are signal dimensions. Usually there are much more dimensions than true effects. (In case of partial correlations in subspaces of the signal it may also be less than required for proper dissection of true effects<sup>(cf. II.1.2.3)</sup>.) Usually principal components are sorted descending by the amount of signal variance explained by them. Then only those are retained that explain more than a selected threshold. In this case, the number of effects again is determined manually and retrospectively.

Preferably, user parameters should all be defined before effect detection starts and should be kept to a minimum. In particular, the number of effects in the signal should be determined unsupervisedly and not by retrospective user action.

Signal dissection estimates the true noise level<sup>(cf. II.5.1.1)</sup>. Based on this estimation, the significance of the signal strength in an effect's focus can be computed<sup>(cf. II.5.1.2)</sup>. Additionally, the significance of correlations between genes and samples in an effect is computed<sup>(cf. II.5.2)</sup>. Based on prescribed significance thresholds, this can be utilized for effect qualification<sup>(cf. II.3.1.8)</sup>. Only few parameters are defined for this effect qualification and all are prescribed. As soon as no more gene or sample qualifies, the method terminates, thereby determining the number of effects in the signal unsupervisedly.

#### II.1.2.8

### Completeness of results

---

Ideally, an analysis method can detect all consistent interactions that were measured in a given signal and represent them 1:1 by interpretable effects.

Hierarchical clustering can only return a single sample clustering for presented data. The only way to get towards completeness is to use it multiple times via sequential focusing as explained<sup>(II.1.2.4)</sup>, with all associated problems outlined above. In case of principal components, they represent 100% of the measured signal in new coordinates per construction. But as explained above, principal components do not necessarily match true effects and in case of high-dimensional signals, most of them usually explain just different aspects of signal noise.

Preferably, effects that explain noise (i.e. false positives) should be prevented, but simultaneously a strong statement about the remaining unexplained signal should be made with respect to the non-existence of further true effects in it (i.e. no false negatives, either).

The signal model<sup>(cf. II.2.1.1)</sup> and the generic bimonotonic effect model<sup>(cf. II.2.1.2)</sup> can represent a broad class of effects and signals comprised of such effects. With respect to preventing false positives while simultaneously discovering true positives, again effect qualification<sup>(cf. II.3.1.8)</sup> is utilized based on the significance of an effect candidate's signal strengths and correlations. Method validation demonstrates that this works reliably with a high rate of discovered true positives while only accepting occasional false positives<sup>(cf. II.6.2.5)</sup>.

#### II.1.3

### Needed Capabilities beyond Detection

---

Additional to solving discussed conceptual problems of previous detection methods<sup>(II.1.2)</sup>, the typical workflow in molecular biology requires several capabilities beyond unsupervised effect discovery, for example their cross-cohort validation. These capabilities are out of scope for detection methods like hierarchical clustering or PCA, but they should be included as part of a comprehensive solution. They are presented and applied in Chapter III, but ideally, the detection stage is already designed with these subsequent requirements in mind, as briefly introduced below.

##### II.1.3.1

#### Comparability and validation of effects across sample cohorts

---

Often, gene expression signals for a single patient cohort contain cohort-specific lab effects, e.g. systematic properties of the utilized microarray, experimental setup or protocol. Resulting effects on the signal are statistical true positives, but biological false positives. Therefore, it is mandatory in life sciences to validate any discovered gene expression effects, however significant they may be. This is also helpful to filter out false positives that might have been detected due to imperfect dissections or due to underestimated noise.

To realize this validation, a comparison method for detected effects is required. For the detection stage this implies that effects should already be discovered and stored in a *form that is suitable for later comparison*.

Effects discovered by signal dissection can be compared directly by weighted correlations<sup>(cf. II.2.3.1)</sup> of their axes. This cutoff-free comparison does not only test whether flat sets of top genes are compatible, but precisely quantifies the consistency of the *relative order* and of all regulation strengths of genes respectively samples. These correlations can be associated with a  $p$  value for statistical assessment<sup>(cf. II.5.2)</sup>. Based on these comparisons, effects can be validated<sup>(cf. III.1.2)</sup> by looking for multiple unsupervised discoveries of the same genomic effect in independent cohorts. Alternatively, effects may be validated by supervisedly testing their existence in signals from other cohorts (similar to effect application for classification; cf. II.1.3.3).

### II.1.3.2 Cohort-independent genomic consensus effects

If the same biological phenomenon has been discovered in multiple patient cohorts, then the information from all available sample cohorts should be used to represent this genomic effect in its most general way. To this end, a procedure should combine all individual discoveries to a consensus gene effect that is as cohort-independent as possible.

Consensus gene effects can be readily computed from multiple discoveries of the same biological effect by averaging their gene axes, gene correlations and gene weights returned by signal dissection<sup>(cf. III.1.3)</sup>.

### II.1.3.3 Patient classification in new cohorts by genomic consensus effects

Once consensus gene effects have been constructed, a typical application is to sort or classify new samples of new cohorts by them, thereby predicting their association with the discovered genomic effect. For example, patients could be sorted by an effect that describes the gene expression differences of particular disease subtypes. Resulting correlations of samples with subtypes and associated  $p$  values might be useful for determining the optimal therapeutic strategy for these patients in precision medicine contexts.

All algorithmic parts of signal dissection except for the search strategy can be identically reused to compute consensus sample effects<sup>(cf. III.1.4)</sup>, thereby quantifying and predicting effect strengths in samples from new cohorts. To this end, the search strategy is replaced by the respective consensus gene effect that should be used for sample classification.

### II.1.3.4 Specific biostatistical evaluation of discovered effects

To help experts characterize and identify consensus effects biologically, ideally all available computable biological knowledge that is significantly associated with an effect should be summarized in an easily retrievable form. To this end, effects should retain as much as possible information. Only then, specific associations can be found and misleading false-positive associations can be prevented.

For example, in case of clustering and partitioning methods, statistical association methods are limited to flat sets, again. This essentially restricts to categorical statistical tests. Sample clusters may be associated with e.g. clinical covariates and gene clusters could be analyzed in context of known gene signatures by overrepresentation analyses. However, more sophisticated statistical association methods like gene set enrichment analyses<sup>(I.3.1)</sup> require more information like a continuous quantification of the involvement of *every single gene* in a particular effect.

All effects discovered by signal dissection and constructed consensus effects can be tested directly by gene set enrichment analyses. More precisely, several scores for gene ranking are available to this end<sup>(cf. III.1.3.2)</sup>. Additionally, effects can be biostatistically associated with several other sources of computable genomic respectively clinical knowledge<sup>(cf. III.3)</sup>.

## II.2 Mathematical Framework

Signal dissection is based on a superposition model for the signal and an unrestrictive bimonotonic effect model. After clarifying and motivating this approach, basic concepts to formalize the algorithm like gene and sample vector spaces and effect curves are defined.

The concept utilized for quantification and later detection of interactions<sup>(cf. I.1.3.4)</sup> is motivated and corresponding functions like uncentered weighted correlations are defined. For illustration, an equivalent geometric interpretation is provided as well.

### II.2.1 Model

Like principal components analyses<sup>(cf. I.2.2.2)</sup> that search for maximal variance in the signal or hierarchical clustering<sup>(I.2.2.1)</sup> that is based on a distance metric and a linkage method, every unsupervised method needs a compatible model for the specific type of interactions that it tries to unveil<sup>(cf. I.1.3.4)</sup>. Ideally, this model leads to an *unambiguous* dissection of the signal into interpretable effects of interaction, while simultaneously making as few as possible assumptions and demanding as few as possible constraints.

#### II.2.1.1 Signal model

The complete measured signal for  $m$  genes (or features) and  $n$  samples can be represented as a matrix  $M_0 \in \mathbb{R}^{m \times n}$ , where  $\mathbb{R}^{m \times n} \equiv \left\{ (X_{ij})_{i=1 \dots m, j=1 \dots n} \mid X_{ij} \in \mathbb{R} \right\}$  is the **signal space**. The lower index zero indicates the initial signal before the first detection and dissection iteration.

To dissect the signal into its generating effects, I assume that the measured signal  $M_0$  is the result of a *superposition of different effects*  $E_k$ . Like  $M_0$ , all  $E_k$  are matrices  $\in \mathbb{R}^{m \times n}$ . Together with a random matrix  $\eta$  of the same size and with normally distributed components, the **signal model** as matrix equation is a simple sum<sup>(Eqn. II.2.1.1)</sup>.

As effects  $E_k$  are dissected iteratively by the method,  $M_k \equiv M_0 - \sum_{k'=1}^k E_{k'}$  denotes the remaining signal at the end of detection and dissection iteration  $k$ . Hence,  $M_{k-1}$  is the initial signal for the following iteration and  $M_0$  is the initial signal for the first iteration.

In molecular biology, the method is applied to  $\log_2$ -transformed gene expression ratios. Hence, the model is multiplicative instead of additive for raw signal intensities. This is intended, because folding is the biologically relevant measure and e.g. 8-fold upregulation versus healthy controls is about as important as downregulation to  $\frac{1}{8}$ , whereas the addition of a constant amount of mRNA expression by itself has no clear biological meaning without knowing the base concentration. For example, adding  $x$  mRNA molecules to a reservoir of already  $100x$  mRNAs of the same sequence in the same cell usually has *biologically insignificant* functional consequences compared to adding the same  $x$  mRNA molecules to a cell that did not express this sequence at all so far. With help of  $\log_2$ -transformations, absolute numerical values become the same for the same biological importance. For the above example,  $\log_2 8 = 3$  and  $\log_2 \frac{1}{8} = -3$ . Consequently, the method should search for additive effects in this  $\log_2$ -scale, rather than searching for additive effects on signal intensity level.

$$M_0 = \sum_{k=1}^{\hat{k}} E_k + \eta$$

Eqn. II.2.1.1) Signal model

The initial signal  $M_0$  is modeled as a sum of effects  $E_k$  plus normally distributed noise  $\eta$ . This matrix equation is valid for all genes  $i$  and all samples  $j$ . The total number of effects  $\hat{k}$  is initially unknown, as are all  $E_k$ .



Dissecting a sum into summands is ambiguous without further information or constraints. Hence, an **effect model** is required. While effects are visually clear in the 3D example<sup>(Figure II.1.1.3)</sup>, this needs to be generalized in a formal way to guide detection and regression of high-dimensional effects.

Let  $I_0^g \equiv (1, 2, \dots, m)$  be the measurement order of genes and  $J_0^s \equiv (1, 2, \dots, n)$  be the measurement order of samples. These external **reference orders** are needed for formal reasons when resorting a signal matrix according to an effect. They usually do not correspond themselves to any effect of interaction. For sort vectors  $(I, J)$ , i.e. for a permutation  $I$  of  $I_0^g$  and a permutation  $J$  of  $J_0^s$ , let  $X(I, J) \equiv \left( X(I(i), J(j)) \right)_{i=1 \dots m, j=1 \dots n}$  denote the correspondingly resorted matrix.

The effect model was inspired by the looks of typical real-world heatmaps that have been *sorted* by differential gene expression<sup>(examples: Figure II.2.1.2, Figure I.1.2.1, [83].Figure 1A, [45].Figure 2)</sup>. The model should be able to represent nonlinear laws of gene regulation that are *monotonic* over the average pathway activity, as illustrated for basic 3D effects<sup>(II.1.1.2)</sup>.

For generalization and for definition, I demand that each effect  $E_k$  has an **eigenorder**  $(I_k, J_k)$  of gene and sample permutations such that the resorted effect signal  $\tilde{E}_k \equiv E_k(I_k, J_k)$  is **bimonotonic**, i.e. monotonic for both genes and samples. This definition is utilized later<sup>(II.4)</sup> to empirically regress an effect's own contribution to the signal and to dissect it from any superposed foreign effects or noise. Hence,  $\tilde{E}_k$  is named the **eigensignal** of effect  $k$ . Writing bimonotonicity out, the effect model reads:

$$\begin{aligned} \forall k: \exists I_k \in \text{perm}(I_0) \exists J_k \in \text{perm}(J_0): \\ \forall i \in [1, m]: \left( \begin{array}{l} \forall j \in [1, n-1]: E_k(I_k(i), J_k(j+1)) \geq E_k(I_k(i), J_k(j)) \\ \vee \forall j \in [1, n-1]: E_k(I_k(i), J_k(j+1)) \leq E_k(I_k(i), J_k(j)) \end{array} \right) \\ \wedge \forall j \in [1, n]: \left( \begin{array}{l} \forall i \in [1, m-1]: E_k(I_k(i+1), J_k(j)) \geq E_k(I_k(i), J_k(j)) \\ \vee \forall i \in [1, m-1]: E_k(I_k(i+1), J_k(j)) \leq E_k(I_k(i), J_k(j)) \end{array} \right) \end{aligned}$$

Eqn. II.2.1.2) Effect Model

$\text{perm}(X)$  Full permutations of a finite set  $X$ , i.e.  $\text{perm}(X) \equiv \{\pi(X) | \pi: X \rightarrow X \text{ bijective}\}$

Other than a supervised sorting of gene expressions<sup>(e.g. Figure II.2.1.2)</sup>, signal dissection utilizes this bimonotonic effect model to *unsupervisedly discover* effects and to *determine* their co-regulated and anti-regulated genes in the first place. More precisely, *correlations* (or anti-correlations) between all genes and between all samples of an effect are a consequence of bimonotonicity (except for unregulated genes or samples at the effect's zero transition). This is utilized for effect detection by searching for high correlations in the signal<sup>(cf. II.3)</sup>. Empirical sample and gene eigenorders are determined for each discovered effect. These empirical eigenorders presort expressions approximately bimonotonic for the effect's top genes and samples. Then the bimonotonic model is applied to regress the effect's eigensignal<sup>(cf. II.4.1)</sup>. Finally, the eigensignal is dissected<sup>(cf. II.4.2)</sup>, which potentially reveals previously overlapped effects for discovery in subsequent detection iterations.

Although bimonotonicity is a sufficient constraint to be able to regress the signal of detected effects<sup>(II.4)</sup>, it is a rather *generic* effect model compared to models that involve specific functional forms or explicit parameters.

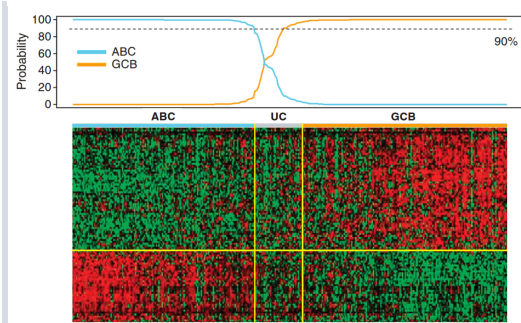


Figure II.2.1.2) Example of a roughly bimonotonic real-world effect<sup>[29]</sup>

This heatmap is based on a supervised analysis that sorts samples of cohort GSE31312 based on their differential expressions between predefined gene signatures for ABC-like and GCB-like subtypes of DLBCL. Originally, this subtype prediction has been developed and applied to an older DLBCL cohort<sup>[83]</sup>, which also resulted in a roughly bimonotonic signature. (Adapted by permission from Macmillan Publishers Ltd: *Leukemia*<sup>[29]</sup>, copyright 2012)

This allows fitting a wide range of regulation laws, including gradual effects like saturations, sub-proportional (e.g. square root like) or super-proportional (e.g. quadratic) regulations as well as unsteady step functions (e.g. between gender groups).

Conceptually, bimonotonicity links the signal of both genes and samples. This type of intrinsic two-dimensionality is an advantage over methods or distance measures that compare either genes or samples<sup>(cf. II.1.2.4)</sup>, especially for dissection tasks in the context of strong<sup>(cf. II.1.2.5)</sup> or many<sup>(cf. II.6.3)</sup> overlapping effects.

In summary, I model the initial signal  $M_0$  as a superposition of effects  $E_k$ . Each effect has an eigenorder  $(I_k, J_k)$  in which its eigensignal  $\tilde{E}_k$  is perfectly bimonotonic. Eigensignals represent and idealize distinct laws of gene regulation or other interactions in the underlying system.

## II.2.2 Basic Concepts

The mathematical framework is set up here to prepare the formalization of signal dissection. Basic descriptive concepts and their notation are defined and motivated.

### II.2.2.1 Gene and sample vector spaces

The **gene space**  $V^g$  is a vector space spanned by  $m$  genes  $\{|e_i^g\rangle | i = 1 \dots m\}$ . Every  $|e_i^g\rangle$  has coordinates  $(\delta_{\mu i})_{\mu=1 \dots m} \in \mathbb{R}^m$  in the gene reference order  $I_0^g$  (where  $\delta_{\mu i}$  is the Kronecker delta, i.e.  $\delta_{\mu i} = 1$ , if  $\mu = i$  and zero otherwise). For clarity, the upper index  $g$  is used to indicate elements of this space or information tokens about all genes. For each sample, all genes have been measured and hence **samples** are points in this vector space spanned by all genes. Let  $|s_j\rangle = \sum_{i=1}^m \langle e_i^g | s_j \rangle |e_i^g\rangle$  denote the  $j^{\text{th}}$  sample vector with expressions  $\langle e_i^g | s_j \rangle \equiv M_0(i, j)$  for gene indices  $i \in \{1, \dots, m\}$ .

The **sample space**  $V^s$  is a vector space spanned by  $n$  samples  $\{|e_j^s\rangle | j = 1 \dots n\}$ . Every  $|e_j^s\rangle$  has coordinates  $(\delta_{vj})_{v=1 \dots n} \in \mathbb{R}^n$  in the sample reference order  $J_0^s$ . The upper index  $s$  indicates elements of this space or information tokens about all samples. For each gene, all samples have been measured and hence **genes** are points in this vector space spanned by all samples. Let  $|g_i\rangle = \sum_{j=1}^n \langle e_j^s | g_i \rangle |e_j^s\rangle$  denote the  $i^{\text{th}}$  gene vector with expressions  $\langle e_j^s | g_i \rangle \equiv M_0(i, j)$  for sample indices  $j \in \{1, \dots, n\}$ .

Generally, a gene vector can be expressed in terms of all contextual sample vectors via  $|g_i\rangle = \sum_{j=1}^n \langle e_j^s | s_j \rangle |e_j^s\rangle$  and vice versa via  $|s_j\rangle = \sum_{i=1}^m \langle e_i^g | g_i \rangle |e_i^g\rangle$ , if they share the same underlying signal matrix. Later, additional lower indices may be used to clarify the underlying signal matrix or the patient cohort.

If the signal matrix and vector components are not clear from context or for definition purposes, I use abbreviations like  $|M_k(I_0, j)\rangle$  or  $|M_k(i, J_0)\rangle$ . They define vectors by directly specifying their coordinates in the respective canonical base, i.e. formally  $\forall i: \langle e_i^g | M_k(I_0, j) \rangle \equiv M_k(I_0(i), j) = M_k(i, j)$  and  $\forall j: \langle e_j^s | M_k(i, J_0) \rangle \equiv M_k(i, J_0(j)) = M_k(i, j)$ .

### II.2.2.2 Gene and samples axes

Let  $|a^g\rangle \in V^g$  be an arbitrary nonzero vector. Its components  $\langle e_i^g | a^g \rangle$  for all genes define a specific *direction* or *axis* in the gene space. Every direction can be interpreted as a *linear approximation of a specific gene regulation law*. If a pathway mediating this type of interactions exists in the underlying system, affected samples form a point cloud around this axis.

The search strategy<sup>(cf. II.3.1)</sup> screens and scores many candidate directions to discover such dominant linear directions, i.e. to discover **gene axes** pointing to effects<sup>(see II.1.1.2 for 3D examples)</sup>. Every effect compatible with the

effect model has a gene axis, as monotonicity is always associated with an axis, over which the function is monotonic.

Likewise, let  $|a^s\rangle \in V^s$  denote the **sample axis** of an effect, i.e. its dominant linear direction in sample space. Here,  $\langle e_j^s | a^s \rangle$  can quantify involvement of samples in an effect.

### II.2.2.3

### Effect curves

While a pair of a gene and a sample axis can already pinpoint an effect, they are only linear approximations of its law of regulation. After regression of an effect's eigensignal  $\tilde{E}_k$ , this approximation can be improved by replacing axes with effect curves that are monotonic over these axes. (More precisely, if using projections on effect axes as scalar curve parameter, then coordinate functions of effect curves are monotonic for all dimensions.)

An effect's gene curve completely describes the empirically regressed law of gene regulation for each sample and its sample curve completely describes regulation differences between samples for each gene. Hence, both are different descriptions of the same information. Indeed, both are just re-parameterizations of the effect's eigensignal.

An effect's **gene curve** is a vector-valued function over sample indices  $j$  with components  $\langle e_i^g | c_k^g(j) \rangle \equiv E_k(i, j)$ . In eigenorder  $J_k$  of effect  $k$ , each component of  $|c_k^g\rangle$  is monotonic, i.e.  $\forall i: (\forall j \in [1, n-1]: \langle e_i^g | c_k^g(J_k(j+1)) \rangle \geq \langle e_i^g | c_k^g(J_k(j)) \rangle \vee \forall j \in [1, n-1]: \langle e_i^g | c_k^g(J_k(j+1)) \rangle \leq \langle e_i^g | c_k^g(J_k(j)) \rangle)$ . Gene curves for effects in the 3D example are regressed and illustrated in II.4.

Analogously, an effect's **sample curve** is a vector-valued function over gene indices  $i$  with components  $\langle e_j^s | c_k^s(i) \rangle \equiv E_k(i, j)$ . In eigenorder  $I_k$  of effect  $k$ , each component of  $|c_k^s\rangle$  is monotonic, i.e.  $\forall j: (\forall i \in [1, m-1]: \langle e_j^s | c_k^s(I_k(i+1)) \rangle \geq \langle e_j^s | c_k^s(I_k(i)) \rangle \vee \forall i \in [1, m-1]: \langle e_j^s | c_k^s(I_k(i+1)) \rangle \leq \langle e_j^s | c_k^s(I_k(i)) \rangle)$ .

In brief, effect curves re-parameterize an effect's eigensignal, run through the effect's point clouds in gene respectively sample space and generalize linear approximations by gene or sample axes to more precise monotonic approximations.

### II.2.2.4

### Effect focus

Normally, an effect neither regulates all measured genes nor does it necessarily exist in all measured samples. When using scalar measures based on *all* genes and *all* samples, small effects are hard to detect<sup>(cf. II.1.2.1)</sup>. If, for example, 20000 genes were measured for each sample, but only 10 genes are truly and strongly correlated, then corresponding correlations between samples are typically heavily diluted and not significant when using all 20000 measured expressions for their calculation. Masses of noise genes would dominate the computation of correlation coefficients. However, if gene weights were utilized to focus on only a small subset of genes that contains all 10 true positives, resulting correlation coefficients would become strong and significant, as noisy information from unimportant genes would be excluded. Hence, *I focus* on gene and sample *subspaces* in which an effect exists in order to detect it<sup>(cf. II.3.1)</sup>. Additionally, discovered effect foci are utilized for effect dissection<sup>(II.4.2.1)</sup>. They help to prevent the modification of signal parts from foreign effects, i.e. from effects that represent gene regulations outside of the current effect's focus.

Collecting weights as vectors, the **effect focus** consists of **gene and sample weights**  $|w^g\rangle$  and  $|w^s\rangle$  with components  $\in [0,1] \subset \mathbb{R}$ . Weights are initially estimated based on the standardized signal<sup>(cf. II.3.1.3)</sup> and are iteratively refined based on correlations<sup>(cf. Eqn. II.3.1.4.b)</sup> during search strategy.

To pinpoint bimonotonic effects, the search strategy needs a *measure of interaction* that can be computed from the perspective of single genes  $|g_i\rangle$  or single samples  $|s_j\rangle$ . Fortunately, (nonzero) bimonotonic effect eigensignals have high correlations between genes and between samples as a consequence. Hence, a type of correlation can be utilized to quantify interactions and to search for bimonotonic effects. This measure is compatible to previous methods in biosciences, including the common form of hierarchical clustering for microarray data<sup>[36]</sup>.

An “effect” in molecular biology typically is observed as the co-regulation of many genes in many samples. From co-regulated gene expressions, it might be concluded that participating genes take part in the same biological function of measured cells. While co-regulation is to my knowledge just a descriptive term without any formal definition, correlation is formally well defined and can quantify co-regulation. It goes beyond just demanding regulation “in the same direction” (i.e. either all upregulated or all downregulated relative to a reference) and also tests *whether differences in regulation strength are related*. Hence, correlation is a specific form of co-regulation and can be used to quantify the consistency of interactions that cause correlations between measured gene expressions in the first place. To detect biologically specific effects, I additionally combine correlations with the effect focus<sup>(cf. II.2.2.4)</sup>. To this end, a weighted form of correlations is needed.

### II.2.3.1 Weighted uncentered correlations aka the cosine distance

To quantify the interaction of a sample respectively a gene  $|x\rangle$  with other samples respectively genes in an effect, I utilize weighted uncentered Pearson correlations to the effect’s representative gene respectively sample axis  $|a\rangle$ , i.e. to its linearized law of regulation in the respective signal space<sup>(cf. II.2.2.1)</sup>.

As already explained<sup>(cf. II.2.2.4)</sup>, using *weights* is necessary to focus on an effect and to prevent computed correlations from being washed out by inclusion of many non-effect dimensions. Additionally, using *uncentered* correlations respects that a  $\log_2(\text{ratio})$  of zero already indicates no regulation and thus defines the global baseline. (In this context, “pre-centered” may be a more intuitive term than the usual “uncentered” is.) Pre-centering has the advantage that the score can also identify points as highly correlated if they are located in the same direction far off from baseline, even if these points show no correlation relative to the center of their common offset. The usual centered Pearson correlation would instead ignore this offset in a common direction and thus would drop important information.

Let  $|w\rangle$  denote gene respectively sample weights in the current effect focus. Then the weighted uncentered correlation of  $|x\rangle$  with  $|a\rangle$  in focus  $|w\rangle$  can be easily defined with scalar products<sup>(Eqn. II.2.3.1)</sup>.

$$[x|a]_{|w\rangle} \equiv \frac{\langle w, x|w, a \rangle}{\|w, x\| \|w, a\|}$$

Eqn. II.2.3.1) Weighted uncentered correlation

Dots denote component-wise multiplication (Hadamard product).  
This measure  $\in [-1, 1] \subset \mathbb{R}$  is utilized by signal dissection for quantification of interactions between genes or between samples.

In the unweighted case ( $|w\rangle=|1\rangle$ ), i.e.  $\forall i: \langle e_i^g | w \rangle = 1$  and if components of  $|x\rangle$  and  $|a\rangle$  have zero means, this definition is identical to the usual Pearson correlation coefficient; hence the name. (In this case,  $\langle w, x|w, a \rangle = \langle x|a \rangle = \sum_i x_i a_i$  equals the uncentered covariance of components from  $|x\rangle$  and from  $|a\rangle$ , times the number of components. Likewise,  $\|w, x\| \|w, a\| = \|x\| \|a\| = \sqrt{\langle x|x \rangle} \sqrt{\langle a|a \rangle}$  equals the product of the uncentered standard deviations of components in  $|x\rangle$  and in  $|a\rangle$ , times the number of components. The number of components cancels and uncentered equals centered for zero means, resulting in the usual Pearson correlation coefficient that is defined as covariance over the product of standard deviations.)

This measure also has an illustrative geometric interpretation: The cosine in a right-angled triangle is defined as the quotient of its adjacent side to its hypotenuse. In the unweighted case (i.e. again  $|w\rangle=|1\rangle$ ),

$\|x\|$  can be seen as the length of the hypotenuse, while the length of its projection on  $|a\rangle$ , i.e.  $\langle x|a\rangle/\|a\|$ , is the length of the adjacent side. Therefore, uncentered correlation is *equivalent to the cosine of the angle* between the two compared vectors:  $[x|a]_{|1\rangle} = \cos(\angle(x, |a\rangle))$ . Hence,  $[x|a]_{|w\rangle}$  can be thought of a generalized cosine between two vectors in the weighted subspace cut by  $|w\rangle$ .

Like usual correlations, weighted correlations assume values  $\in [-1,1]$ . Together with the above geometric interpretation,  $+1$  can be interpreted as 100% parallelism of  $|x\rangle$  and  $|a\rangle$ , while  $-1$  indicates perfect anti-parallelism. Zero designates perpendicular vectors, i.e. uncorrelated effects. Consequently, if  $|a\rangle$  describes an effect axis, perpendicularity implies that no parts of the signal of  $|x\rangle$  can be explained by the linear law of gene regulation encoded by  $|a\rangle$ .

Again, vectors may be specified directly by their coordinates from a signal matrix; abbreviations like  $[M_k(I_0, j)|a^g]_{|w^g\rangle}$  or  $[M_k(i, J_0)|a^s]_{|w^s\rangle}$  denote correlations with  $|M_k(I_0, j)\rangle$  respectively  $|M_k(i, J_0)\rangle$  (cf. II.2.2.1).

### II.2.3.2

## Weighted projections

Let  $|a\rangle$  denote a projection target,  $|x\rangle$  a vector from the same space and  $|w\rangle$  weights for all dimensions of this space. The projection target typically is a gene or a sample axis of an effect again, i.e. it linearly encodes a specific law of regulation. Already defined correlations (Eqn. II.2.3.1) are neither proportional to  $\|x\|$  nor to  $\|a\|$ . Hence, correlations can only compare directions. Sometimes the absolute signal strength of  $|x\rangle$  that is explained by a specific law of regulation  $|a\rangle$  is of interest, i.e. the *signal strength of  $|x\rangle$  in direction of  $|a\rangle$* .

To this end, weighted projections of  $|x\rangle$  in direction of  $|a\rangle$  can be defined (Eqn. II.2.3.2). They are identical to weighted correlations, except for the factor for the weighted norm of  $|x\rangle$ . (The upper index zero is used to avoid confusion with the usual scalar product that does not normalize by the norm of  $|a\rangle$ .)

$$\langle x|a\rangle_{|w\rangle}^0 \equiv \frac{\langle w \cdot x | w \cdot a \rangle}{\|w \cdot a\|}$$

Eqn. II.2.3.2) Weighted projections

Dots denote component-wise multiplication (Hadamard product). This measure is similar to weighted correlations (cf. Eqn. II.2.3.1), but it scales with the signal strength of  $|x\rangle$ .

*At the beginning of each detection iteration, the signal is screened for effect candidates that qualify relative to significance thresholds for their signal strength and correlation.*

*If an effect exists, it affects some genes and some samples. This is utilized to discover an initial representative that roughly points in an effect's direction. Once an initial gene or sample has been identified, additional genes or samples of the same effect are searched and incorporated into its definition. A maximum principle guides this search towards effect axes to which as many as possible genes and samples are correlated as high as possible. This continues iteratively until the estimated effect axes have converged.*

*After convergence, effect axes can be considered independent of individual features from single representatives and hence are representative for the effect as a whole. They linearly approximate its laws of gene regulation and serve as starting point for the precise regression of its eigensignal<sup>(II.4)</sup>.*

*As soon as no gene and no sample qualify any longer, the method declares that no significant effects remain in the signal and terminates.*

### II.3.1

## Finding Effects

It is hard to directly search for bimonotonic effects: In principle, one could enumerate all possible pairs of gene subsets and sample subsets from the joint power set  $\mathcal{P}(m) \times \mathcal{P}(n)$ , sort them by their average expressions and directly score them. However, this is not viable for typical  $m$  and  $n$ , since  $|\mathcal{P}(m) \times \mathcal{P}(n)| = 2^{m+n}$ . Hence, a search strategy is required. My search strategy is based on measures for interactions<sup>(II.2.3)</sup> that can be computed from the perspective of a *single* gene or *single* sample and that can be summarized by an effect score<sup>(II.3.1.6)</sup>. In brief, the more genes and samples participate in an effect and the higher correlated they are to this effect, the larger is this score.

For performance reasons, a deterministic lookahead scheme based on a presorting<sup>(II.3.1.7)</sup> is utilized to efficiently screen genes and samples for a good initial representative, i.e. for a new effect  $k$  with locally maximal score. The initial representative is either a gene with index  $i_{k,1}$  and gene expressions  $|g_{i_{k,1}}| \equiv |M_{k-1}(i_{k,1}, J_0)|$  or a sample with index  $j_{k,1}$  and gene expressions  $|s_{j_{k,1}}| \equiv |M_{k-1}(I_0, j_{k,1})|$ . To immediately symmetrize the situation<sup>(II.3.1.3)</sup>, gene and sample axes are computed for either type of initial representative.

Initial axes already point in the direction of the newly discovered effect  $k$ , but usually can approximate its law of regulation only roughly. For example, the gene axis<sup>(yellow)</sup> based on the marked initial sample for the blue effect<sup>(Figure II.3.1)</sup> is only a good approximation for its outmost tips.

Next, all steps leading to a discovered effect's initial axes are presented in processing order, thereby explaining the

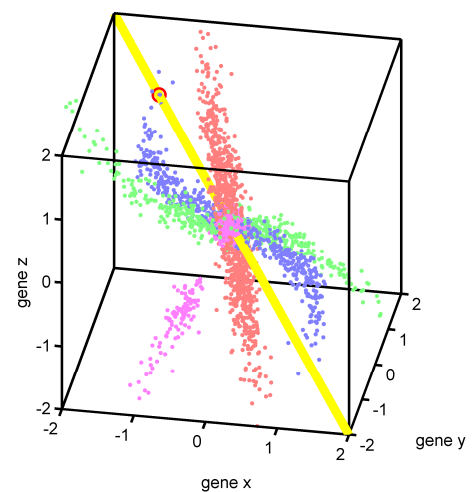


Figure II.3.1) 3D concept example, first detection iteration, initial representative for the blue effect and its associated initial gene axis

algorithm's structure. Again, components of genes and samples are always implicitly taken from the contextual signal matrix<sup>(cf. II.2.2.1)</sup>. Here, this is  $M_0$  for the first detection iteration and in general it is the remaining signal<sup>(cf. II.2.1.1)</sup>, i.e.  $M_{k-1} = M_0 - \sum_{k'=1}^{k-1} E_{k'}$ . For example,  $\langle e_j^s | g_{i_{k,1}} \rangle = M_{k-1}(i_{k,1}, j)$ .

### II.3.1.1

## Standardization against outliers

One conceptual goal of the method is to be robust against outliers. To this end and also as a basis for initial weights (as defined in the next section), the standardized signal  $M_{k-1}^S$  is computed from the remaining signal  $M_{k-1}$  at the beginning of every detection iteration  $k$ . These standardizations are associated with the loss of one signal dimension (for 3D examples, the signal is embedded into a 2D surface). Hence they should not be utilized as surrogate for the unstandardized signal, but they can be used for scoring or weighting.

An iterative standardization is utilized that converges uncentered variances  $E[(X - 0)^2]$  for all genes respectively all samples to one, where  $X$  denotes the random variable sampled by the components of a particular gene or sample vector. The empirical estimation  $\hat{E}$  of this variance for a sample  $|s_j\rangle$  with components  $\langle e_i^g | s_j \rangle$  equals  $\sum_{i=1}^m \langle e_i^g | s_j \rangle^2 / m$  and hence is proportional to its squared Euclidean norm  $\|s\|^2 = \langle s_j | s_j \rangle = \sum_{i=1}^m \langle s_j | e_i^g \rangle \langle e_i^g | s_j \rangle = \sum_{i=1}^m \langle e_i^g | s_j \rangle^2$ , but is normalized such that it is independent of the number of genes  $m$ . A result of this proportionality is that for any standardized sample  $|s_j^S\rangle \equiv |M_{k-1}^S(I_0, j)\rangle$ , the Euclidean norm equals  $\|s_j^S\| = (m \hat{E}[X^2])^{1/2} = \sqrt{m}$ , as  $\hat{E}[X^2] = 1$  after standardization.

For the 3D example this implies that all standardized sample vectors lie on the  $S^2$  sphere<sup>(Figure II.3.1.1.a)</sup> in gene space with radius  $\sqrt{3}$ . Similarly, all three standardized gene vectors have norm  $\sqrt{n}$  and thus lie on  $S^{2499}$  with radius  $\sqrt{2500} = 50$  in the associated sample space. As uncentered variance equals one for all columns (and all rows) of  $M_{k-1}^S$ , it follows  $\sum_{j=1}^n \sum_{i=1}^m M_{k-1}^S(i, j)^2 / (mn) = 1$  for the whole matrix. If the input signal matrix was centered (which is a usual preprocessing step for  $\log_2$ -transformed gene expressions), this implies that the standard deviation from one pixel to another in  $M_{k-1}^S$  is  $\pm 1$ , too. Hence, all possible signal strength fluctuations have been equalized.

The implementation of this standardization is straightforward: Uncentered variances are empirically estimated for every gene row and every sample column as above, resulting in gene variances  $v^g \in \mathbb{R}^{m \times 1}$  and the sample variances  $v^s \in \mathbb{R}^{1 \times n}$ . The signal matrix is then divided component-wise by the component-wise square roots of the matrix product  $v^g v^s \in \mathbb{R}^{m \times n}$ . This is iterated until uncentered variances of all genes and all samples equal one within an epsilon that is determined as  $1/1000$  of the current noise level<sup>(cf. II.5.1.1)</sup>. Convergence is usually reached within few tens of iterations.

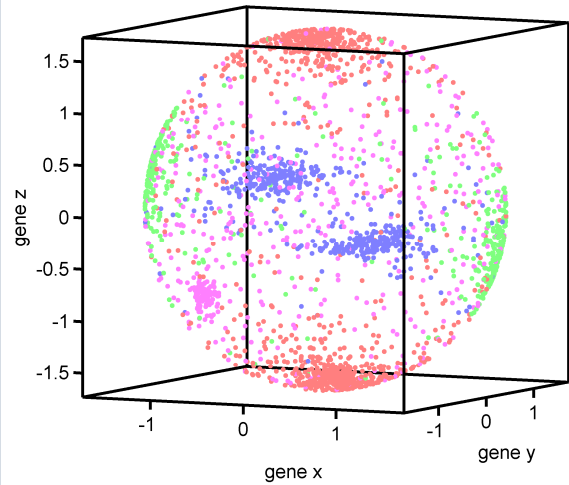


Figure II.3.1.1.a) Standardization results  $M_k^S$  in gene space for the 3D example via equalization of uncentered variances

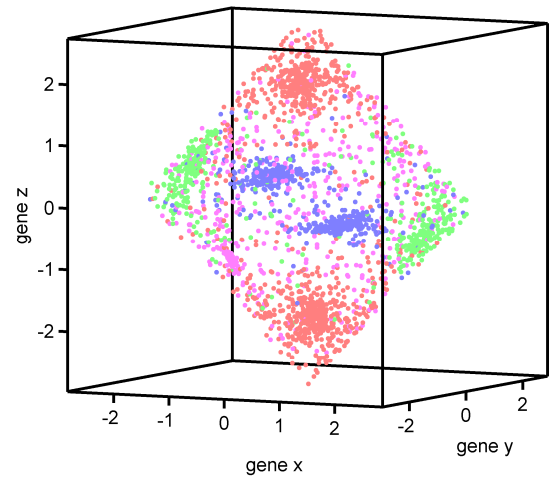


Figure II.3.1.1.b) Standardization results in gene space for the 3D example when equalizing absolute norms

Notably, this procedure equalizes norms in both the gene and the sample space, while dividing genes by their norms (or samples by theirs) only equalizes norms in either space, but not in both spaces simultaneously.

The same iterative procedure can be utilized to equalize by other measures instead of uncentered variance, for example by the absolute norm  $\sum_i |x_i|$ . For the 3D example this would result in a projection onto diamonds<sup>(cf. Figure II.3.1.1.b)</sup> in both the gene and the sample space. However, to eliminate signal strength fluctuations and for initial weights, the sphere seems to be the more natural choice for standardization.

### II.3.1.2

## The effect focus and its initial estimation

---

Finding a formula for an adaptive effect focus that is as general-purpose as possible has been the second-hardest task after realizing bimonotonic regression<sup>(cf. II.4.1)</sup>, because every processing step depends on the effect focus and they do so in an iterative way with self-feedback.

For example, a too narrow effect focus may seem sharp and well-defined locally, but may lead to seeing only a fraction of the true effect when computing correlations and projections. This could iteratively grow into far too narrow “stripe-like” false positive effects. These stripes look like true positives in their narrow focus, but are essentially ordered noise. For combinatorial reasons, sort orders that arrange a small fraction of samples in such a way that hundreds of measured probesets seem to be co-regulated always exist in large signal matrices with  $\sim 50000$  probesets<sup>(e.g. Figure II.6.4.2.d)</sup>. On the other hand, a too broadly defined gene focus includes too many noise genes and thus correlations or projections of samples based on this focus are washed out<sup>(cf. II.1.2.1)</sup>. Iteratively, this can grow into narrow false positives as well, because weights are dependent on correlations of previous convergence iterations for effect axes<sup>(cf. II.3.2.2)</sup>. Consequently, sample weights may be washed out towards zero for all but those few samples, whose genes can be ordered in such a way that a non-vanishing signal remains despite the over-broad gene focus. This is similar to defining the sample focus too sharp or over-optimized in the first place.

Furthermore, effect focus formulas working perfectly in one scenario, for example in the versatility scenario<sup>(II.6.2)</sup>, might produce artefacts in other scenarios, for example for the 3D example. Therefore, the aim should be to find a common and balanced formula for the effect focus that *simultaneously* works in all method validation scenarios including real world cohorts in order to advance towards a general-purpose method.

A working though recursive formula for focusing weights has been found based on correlations with the effect and based on the significance of these correlations. For an initial gene or sample candidate, neither correlations nor their  $p$  values are available. The effect focus is therefore defined and refined in several steps during a detection iteration. Here, only initial weights are defined; the correlation based formula for all following refinements is provided in II.3.1.4.

### ■ Initial weights estimation

Initially, no information other than the signal strength (in original or standardized units) is available to define weights. For an initial (and rough) approximation of the effect focus, I use weights based on the standardized signal<sup>(II.3.1.1)</sup>. Compared to the signal in original units this has the advantage of being relatively outlier-robust. More importantly, very strong non-standardized signals often are indicative for multiple constructively overlapping effects. However, with an interpretable dissection in mind, genes or samples that are exclusively affected by a single effect are preferred as initial effect representatives. Standardized gene expressions are larger if the original signal of a sample *is concentrated in fewer genes*, and lowest, if all genes are expressed with approximately the same strength. This can be utilized to tighten the effect focus for representatives with expressions pointing to more specific effects. Still, the effect focus should neither become



too narrow for reasons explained above. Otherwise and in the extreme case, effect candidates would be dominated by only one top gene or top sample; this would be neither informative nor could it achieve a qualifying effect score.

I therefore utilize the absolute standardized signal as initial weights, but cap it at 80%:

$$\text{For an initial gene } |g_{i_{k,1}}\rangle, \text{ initial sample weights are: } \langle e_j^s | w_{\text{initial}}^s \rangle \equiv \min \left( 1, \frac{|M_{k-1}^s(i, j)|}{0.8 \cdot \max_j (|M_{k-1}^s(i, j')|)} \right)$$

$$\text{For an initial sample } |s_{j_{k,1}}\rangle, \text{ initial gene weights are: } \langle e_i^g | w_{\text{initial}}^g \rangle \equiv \min \left( 1, \frac{|M_{k-1}^g(i, j_{k,1})|}{0.8 \cdot \max_j (|M_{k-1}^g(i', j_{k,1})|)} \right)$$

Eqn. II.3.1.2) Initial effect focus based on the standardized signal

The main goal of these initial weights is just to prevent too many noise dimensions in context of high-dimensional spaces, as they could wash out scores, wash out the twin axis computed below and thus prevent effect detection.

Practically all dimensions that could be true effect dimensions still have full weights due to the 80% cap. When a single dimension dominates the original signal, this dimension has maximal value in the standard vector; let  $c$  denote this value. The minimal possible value of a component in a standardized vector is obtained for the hyperdiagonal, i.e. when all dimensions equally contribute to the original signal. Via Pythagoras for  $m$  dimensions it holds  $\sum_{i=1}^m d^2 = c^2$ . Thus, the dimension-dependent minimal ratio of components in a standardized vector equals  $\frac{|d|}{|c|} = \frac{1}{\sqrt{m}}$ . Hence and in general, the more dimensions the standardized signal has, the more possibilities exist to get below the 80% cap. Hence, initial weights based on standardized signals prevent washed-out scores in a way that is adaptive to signal dimensionality.

More precise weights for the effect focus are defined as soon as correlations have been computed<sup>(cf. II.3.1.4)</sup>.

### II.3.1.3 Initial effect axes and symmetrization by twin axes

One design goal is to symmetrize roles of genes and samples<sup>(cf. II.1.2.4)</sup>. To symmetrize as early as possible, gene and sample axes need to be defined for either type of initial candidate.

Naturally, an initial gene  $|g_{i_{k,1}}\rangle$  defines its own sample axis  $|a^s\rangle$  and an initial sample  $|s_{j_{k,1}}\rangle$  defines its own gene axis  $|a^g\rangle$ .

To compute the respective twin axis in the other vector space, weighed projections<sup>(cf. II.2.3.2)</sup> of all other genes on  $|g_i\rangle = |M_{k-1}(i, J_0^s)\rangle$  respectively of all other samples on  $|s_j\rangle = |M_{k-1}(I_0^g, j)\rangle$  can be computed, using respective initial weights<sup>(cf. Eqn. II.3.1.2)</sup>:

$$\text{The twin gene axis } |a^g\rangle \in V^g \text{ for an initial gene } |g_{i_{k,1}}\rangle \text{ is defined by } \langle e_i^g | a^g \rangle \equiv \langle g_i | g_{i_{k,1}} \rangle_{w_{\text{initial}}^s} / \|w_{\text{initial}}^s\|.$$

$$\text{The twin sample axis } |a^s\rangle \in V^s \text{ for an initial sample } |s_{j_{k,1}}\rangle \text{ is defined by } \langle e_j^s | a^s \rangle \equiv \langle s_j | s_{j_{k,1}} \rangle_{w_{\text{initial}}^g} / \|w_{\text{initial}}^g\|.$$

Eqn. II.3.1.3) Initial twin axes for symmetrization

This results in a pair of initial axes  $(|a^g\rangle, |a^s\rangle)$  for either type of initial candidate. Initial weights for components of the twin axis are computed by repeating twin axis computation for the standardized signal  $M_{k-1}^S$  and then capping at 80% again<sup>(cf. Eqn. II.3.1.2)</sup>. Together with initial weights for the initial candidate, they complete the initial effect focus  $(|w_{\text{initial}}^g\rangle, |w_{\text{initial}}^s\rangle)$ . Initial axes and the initial focus provide a symmetric basis for all following computations and scores.

Analogous projections are also utilized when updating twin axes based on refined weights<sup>(cf. Eqn. II.3.1.4.b)</sup> during effect axes convergence<sup>(II.3.2)</sup>.

Initial axes ( $|a^g\rangle, |a^s\rangle$ ) are already a rough linear approximation of an effect candidate. In order to score this effect, correlations of all genes and all samples with its initial axes are computed next. Using the initial effect focus as weights, uncentered correlations<sup>(cf. II.2.3.1)</sup> can be utilized to this end. All resulting correlations can be collected as two vectors:

**Gene correlations**  $|r^g\rangle \in V^g$  are defined by  $\langle e_i^g | r^g \rangle \equiv [g_i | a^s]_{|w_{\text{initial}}^s\rangle} = [M(i, j_0) | a^s]_{|w_{\text{initial}}^s\rangle}$ .

**Sample correlations**  $|r^s\rangle \in V^s$  are defined by  $\langle e_j^s | r^s \rangle \equiv [s_j | a^g]_{|w_{\text{initial}}^g\rangle} = [M(i_0, j) | a^g]_{|w_{\text{initial}}^g\rangle}$ .

Eqn. II.3.1.4.a) Initial correlations

For each correlation, a  $p$  value can be calculated<sup>(cf. II.5.2.1)</sup> to quantify its significance. Resulting  $|p^g\rangle$  and  $|p^s\rangle$  are utilized together with computed correlations to focus on the effect more precisely than it was initially possible based on the standardized signal<sup>(Eqn. II.3.1.2)</sup>:

Refined **sample weights**: Let  $x_j \equiv |\langle e_j^s | r^s \rangle| \cdot (1 - \langle e_j^s | p^s \rangle)^2$  and  $y_j \equiv \min\left(1, \frac{x_j}{0.5 \cdot \max_j(x_j)}\right)$ , then

$\langle e_j^s | w^s \rangle \equiv \begin{cases} y_j, & \text{if } y_j \geq \frac{0.67}{n} |\{y_{j'} | y_{j'} \leq y_j\}| \\ 0 & \text{otherwise} \end{cases}$ . Refined **gene weights**: Let  $x_i \equiv |\langle e_i^g | r^g \rangle| \cdot (1 - \langle e_i^g | p^g \rangle)^2$  and

$y_i \equiv \min\left(1, \frac{x_i}{0.5 \cdot \max_i(x_i)}\right)$ , then  $\langle e_i^g | w^g \rangle \equiv \begin{cases} y_i, & \text{if } y_i \geq \frac{0.67}{m} |\{y_{i'} | y_{i'} \leq y_i\}| \\ 0 & \text{otherwise} \end{cases}$ . The **effect focus** is  $(|w^g\rangle, |w^s\rangle)$ .

Eqn. II.3.1.4.b) Refined effect focus based on correlations

In words, I replace the initial effect focus by weights that are based on absolute correlations times a factor that goes down to zero quadratically with the noise probability of these correlations. Final weights are defined relative to the maximum and all  $x$  that are  $\geq 50\%$  get full weight. Hence, the order of top genes in an effect is not influenced by different weights. Additionally, weights that are less than 0.67 times the quantile axis are set to zero exactly. Hence, any unspecific influence of (potentially very many) low weights is prevented.

The purpose of the effect focus remains the same after this refinement: Exclude dimensions that have nothing to do with the effect in order to compute as effect-specific as possible scores via weighted projections or correlations. Except for the initial focus, the effect focus  $(|w^g\rangle, |w^s\rangle)$  is always computed as defined here. This includes all focus updates following effect axes updates later during axes convergence in II.3.2.

For effect size estimation and qualification thresholds, mapping all  $x$ -values above 50% to full weight is not optimal; instead I would like to keep the dynamic range of weights for these tasks. To this end, I additionally define the extended effect focus  $(|v^g\rangle, |v^s\rangle)$  by increasing the upper threshold from 50% to 100% (i.e.  $y \equiv x$ ) and by decreasing the lower threshold from 0.67 to 0.4.

The refined effect focus  $(|w^g\rangle, |w^s\rangle)$  is utilized next to also refine twin axes<sup>(Eqn. II.3.1.3)</sup>, correlations<sup>(Eqn. II.3.1.4.a)</sup> and  $p$  values of correlations. Updated scores are focused more precisely on the effect and hence are more representative for it.

Due to the recursive nature of the refined effect focus formula<sup>(Eqn. II.3.1.4.b)</sup>, this focusing step could be iterated. It is iterated during axes convergence<sup>(II.3.2)</sup>. However, during screening for a new effect, a single focusing step suffices. Besides, multiple focusing steps for *every* screened candidate would have a massive negative impact on performance, as every focusing step requires weighted correlations with *all* measured genes and *all* measured samples to be recomputed.

To compare effect candidates and determine the best one, a scalar effect score is computed. During development, this score was effectively a functional degree of freedom whose formula had to be identified by trial and error in the context of test scenarios with synthetic data. This process was guided by two ideas: a) the higher correlations of genes and samples are to an effect, the more interesting is this effect and b) the more genes and samples are correlated to an effect, the more interesting it is.

To estimate effect sizes in each order dimension, i.e. to estimate the number of underlying genes or samples, the effect focus is summated:  $m_k \equiv \sum_{i=1}^m \langle e_i^g | v^g \rangle$  and  $n_k \equiv \sum_{j=1}^n \langle e_j^s | v^s \rangle$ . To determine a *single scalar that represents the effect's size*,  $m_k$  and  $n_k$  need to be combined. This should be done multiplicatively, because  $m_k n_k$  is proportional to the number of the measured values underlying the effect, i.e. the number of pixels in the signal matrix belonging to genes and samples with high correlation to the effect. To prevent the effect size factor from dominating the score, a square root is utilized below to balance its contribution relative to the correlation factor. To boost scores for effects that have broad and robust support in both order dimensions, I assign 90% geometric weight to the minimum of  $m_k$  and  $n_k$  and only the remaining 10% to their maximum. This causes broad effects in all order dimensions to be preferred over effect candidates that are broad in one but narrow in the other order dimension. The rationale behind this is that it is usually easier to interpret narrow effects *after* overlapping broad effects have already been dissected from the signal, rather than vice versa. Hence, the final effect size factor reads  $\sqrt{\min(m_k, n_k)^{0.9} \max(m_k, n_k)^{0.1}}$ .

The correlation information from  $|r^g\rangle$  and  $|r^s\rangle$  is also summarized by a representative scalar. This is realized by weighted averages of absolute correlations in the effect's focus. Abbreviating components  $v_i^g \equiv \langle e_i^g | v^g \rangle$ ,  $r_i^g \equiv \langle e_i^g | r^g \rangle$ ,  $v_j^s \equiv \langle e_j^s | v^s \rangle$  and  $r_j^s \equiv \langle e_j^s | r^s \rangle$ , these averages are:  $r_k^g \equiv (\sum_{i=1}^m v_i^g |r_i^g|) / m_k$  and  $r_k^s \equiv (\sum_{j=1}^n v_j^s |r_j^s|) / n_k$ . To determine a *single scalar representing the effect's average correlation*, another average using effect sizes as weights is computed:  $r_k \equiv (n_k r_k^g + m_k r_k^s) / (n_k + m_k)$ . Herein,  $r_k^g$  and  $r_k^s$  are weighted with the number of points supporting the correlations in  $|r^g\rangle$  respectively  $|r^s\rangle$ . This weighting is important when one order dimension has much higher resolution than the other. For instance, there are only three genes in the 3D example and their correlations  $r_i^g$  have been robustly determined by many samples, while sample correlations  $r_j^s$  are only computed over three points each and thus relatively unreliable.

Using effect scores that are directly proportional to  $r_k$  did not provide enough dynamic range, i.e. there was not enough difference between high and merely moderate correlations. This difference is required when multiplying with the effect size in order to prevent that a larger effect with only moderate correlations gets a higher score than a bit smaller effect with high correlations. The aim is to value specificity over size; this also helps dissecting overlapping effects. To solve this problem, I transform correlations via  $r_k / (1 - r_k^2)^{1/2}$ ; this emphasizes correlations near one. (This transformation was inspired by and adapted from the transformation of a correlation into a corresponding *t*-statistic when determining its significance<sup>(cf. II.5.2.1.)</sup>)

The final scalar effect score is the product of the effect's transformed average *correlation* and its summarized *size* as explained above.

$$\delta_k \equiv \frac{r_k}{(1 - r_k^2)^{1/2}} \sqrt{\min(m_k, n_k)^{0.9} \max(m_k, n_k)^{0.1}}$$

Eqn. II.3.1.6) Scalar effect score (average correlation times effect size)

$r_k$  Representative scalar correlation of the effect (see above).  
 $m_k, n_k$  Effect size in the gene respectively sample space (see above).

The score can be computed independent of the type of the initial candidate. Whether a gene or a sample is the better choice as initial representative for an effect depends on the effect's shape, i.e. how it extends to other

genes and other samples. In the 3D example, samples are always better representatives, as none of the three genes is specific to only one of the four effects. In high-dimensional signals<sup>(cf. II.6.1)</sup>, often genes are better initial representatives. The algorithm keeps treating roles of genes and samples symmetrically and both have equal chances of being selected as initial effect representative in every detection iteration.

### II.3.1.7 Search complexity, presorting candidates and a lookahead scheme

Finding the gene or sample with the globally maximal effect score has the following complexity: There are  $m + n$  possible initial candidates. Every gene candidate spawns a gene and a sample axis. Correlating the gene axis with all  $n$  samples has complexity  $O(nm)$ , as every single correlation is linear in the number of values per sample, i.e.  $O(m)$ . Correlating the gene candidate's sample axis with all  $m$  genes results in  $O(mn)$  as well. Likewise, a sample candidate spawns a gene axis and a sample axis that need to be correlated with all samples respectively with all genes to compute the effect score. Altogether, finding the globally maximal effect score has complexity  $O((m + n) \cdot (nm + mn)) = O((m + n) \cdot nm)$ . While this is much faster than the naïve approach that tries to enumerate the joint power set<sup>(cf. II.3.1)</sup>, this cubic complexity is still too slow for typical  $m$  and  $n$  in practice.

To facilitate a successful and fast lookahead for a local (and ideally the global) score maximum, a presorting of all genes and samples is needed that ideally sorts representatives by descending effect scores. This presorting should be based on local scores only that can be computed fast, i.e. scores that can be computed with just linear complexity from perspective of a single gene or sample. Hence, presorting cannot be perfect, as it only utilizes local information about each gene or sample rather than their interactions in form of correlations. However, this sorting does not need to be precise, as it *does not determine subsequent qualification* of genes or samples as effect candidates, but just the *order in which promising candidates are presented*.

A speed advantage is gained together with a lookahead scheme that breaks processing early, if no better score has been found for a certain amount of presented candidates (in the default 200). The better the presorting (i.e. ideally descending in effect scores), the faster this early break is reached. Hence, the factor  $(m + n)$  in the complexity gets replaced by a number  $l$  that in the theoretical worst case is still  $(m + n)$ , but in practice typically  $l \ll (m + n)$  for large  $m, n$ . The lookahead scheme itself is straight forward: When processing the presorted list, every candidate gene or sample has to qualify first<sup>(cf. II.3.1.8)</sup>. Then it competes with scores of other already qualified effect candidates within the lookahead interval. If no stronger candidate is found for the next 200 ranks in the presorting order, the candidate is accepted as initial representative.

The following presorting score has proven to provide a good speedup without hampering quality. It is based on the maximum of two *local* information sources about the initial gene respectively initial sample. The first local score is the candidate's uncentered standard deviation  $\sqrt{E(X^2)}$  (where  $X$  stands for the random variable sampled by all components of the initial candidate). Uncentered instead of usual standard deviations are utilized again for the same reason I chose uncentered correlations<sup>(cf. II.2.3.1)</sup>, i.e. since zero indicates baseline expression and a consistent offset from zero should lead to a higher effect score. The second local score is the candidate's maximal component in its absolute standardized signal, e.g.  $\max_j (|M_{k-1}^S(i, j)|)$  for gene candidate  $i$ . This maximal component is a purity score as explained<sup>(cf. II.3.1.2)</sup>, i.e. it is larger for effects that are concentrated in fewer dimensions rather than spread equally over many dimensions. Purer effects prospectively have higher average correlation, leading to higher effect scores. For each gene and each sample, both local scores can be computed in linear time, as they only need the signal for the respective gene or sample. Both are computed for all genes and all samples at the beginning of each detection iteration (complexity  $O(mn)$ ). Resulting scores are z-transformed (i.e. centered and divided by their usual standard deviation) to make both

sources of local information comparable. Finally, the maximum of these two  $z$  scores for each gene and each sample is employed to presort all genes and all samples by their descending scores. This results in a common list of candidates of length  $m + n$ . This list is then processed by the lookahead scheme as explained above.

### II.3.1.8

## Qualification of candidates

Several qualification scores and corresponding thresholds determine, whether a gene or sample is eligible as initial representative of an effect.

Effect scores only facilitate a relative ranking of initial candidate genes or samples, but they are not suitable for effect qualification. Qualification cutoffs need to be easy-to-configure based on easy-to-interpret properties of candidate effects.  $p$  values quantifying the significance of correlations or of signal strengths in the effect focus are natural choices for qualification. Both are approximated based on  $t$  statistics<sup>(cf. II.5)</sup>. (As no noise distribution has been estimated before the first effect detection<sup>(cf. II.5.1.1)</sup>, only correlation based significance is checked in iteration  $k = 1$ . Still, the effect's significance with respect to its signal strength can be assessed retrospectively at the end of iteration  $k = 1$ .)

Additionally, thresholds for the average correlation in the effect focus and for the effect size are defined for qualification. They can be used to exclude statistically significant, yet uninteresting effects. Default thresholds have been developed in context of synthetic data<sup>(cf. II.6)</sup> and with the aim to still detect even weak true positives, while terminating before the onset of many false positives. With real-world applications in mind, occasional false positives are considered less problematic compared to false negatives (i.e. non-detections), because cross-cohort validation of effects<sup>(cf. III.1.2)</sup> will filter these false positives out again. This additional validation step has to be performed anyway for real-world applications, as systematic errors in the signal like lab effects may be statistical true positives, but are biological false positives. (Hence, they cannot be excluded here by effect qualification as a matter of principle.)

The following table summarizes all parameters utilized for qualification of effect candidates.

| Qualification parameter   | Default threshold      | Comments   |
|---|------------------------|--|
| correlation significance $\alpha_r$                             | $\frac{10^{-10}}{m+n}$ | As $p$ values for correlations within an effect focus <sup>(II.5.2.1)</sup> decrease rapidly over effect size, a very strong significance threshold can be chosen. Because up to $m+n$ initial candidates are tested, the threshold is further decreased by applying the Bonferroni correction for multiple hypothesis testing. Corresponding $p$ values are defined in II.5.2.  |
| signal strength significance $\alpha_s$                         | $\frac{10^{-5}}{m+n}$  | The significance of the signal strength in the effect focus <sup>(II.5.1.2)</sup> is determined relative to the estimated noise level <sup>(II.5.1.1)</sup> . Again, the threshold is decreased by applying the Bonferroni correction of multiple hypothesis testing.  |
| minimal correlation in the effect focus $r_{\min}$              | 0.4                    | The average correlation in the effect focus $r_k$ <sup>(cf. II.3.1.6)</sup> can be utilized to exclude significant effects that are uninteresting, as their signal is too rough and inconsistent for interpretation purposes (even if it is still significant, i.e. if it cannot be explained by noise alone).   |
| minimal number of genes $m_{\min}$ participating in an effect   | $0.5 \cdot \log_2 n$   | The number of genes in an effect $m_k$ is defined in II.3.1.6. With the intention to define defaults that are as general as possible, I utilize a formula that is adaptive to the signal size. The more samples $n$ are measured, the more combinatorial possibilities exist to arrange them in a way such that a sample subset looks like a true effect for few genes. Hence, I require more genes in an effect, if there are more samples to choose from. The precise functional dependency has been derived from experience with the noise genes test scenario <sup>(cf. II.6.4.2)</sup> and the few samples test scenario <sup>(cf. II.6.4.3)</sup> .  |
| minimal number of samples $n_{\min}$ participating in an effect | $0.5 \cdot \log_2 m$   | The number of samples in an effect $n_k$ is defined in II.3.1.6. With the intention to define defaults that are as general as possible, I utilize a formula that is adaptive to the signal size. The more genes $m$ were measured, the more combinatorial possibilities exist to arrange a subset of them in a way that looks like a true effect for a few samples. Hence, I require more samples in an effect, if there are more genes to choose from. For a typical real-world $m = 50000$ this results in $n_{\min} \approx 8$ samples as minimum to qualify as "interesting" effect. The precise functional dependency has been derived from experience with the noise genes test scenario <sup>(cf. II.6.4.2)</sup> and the few samples test scenario <sup>(cf. II.6.4.3)</sup> . |

Table II.3.1.8) Qualification thresholds for effect candidates

Default thresholds have been defined based on experience with synthetic test scenarios<sup>(cf. II.6)</sup> and have been optimized to prevent most false positives, while keeping sensitivity even for small and weak effects in all but very noisy contexts<sup>(cf. II.6.4 for details)</sup>.

The gene or sample that has qualified relative to all these thresholds and that has the best effect score of all other candidates in the lookahead interval<sup>(cf. II.3.1.7)</sup> is finally selected as *initial representative*. Its gene and sample axes ( $|a^g\rangle, |a^s\rangle$ ) already point to the discovered effect that is regressed and dissected in this detection iteration. As a first step towards this dissection, effect axes are refined by convergence as described in the following section.

## II.3.2 Effect Axes Convergence

To optimize axes for the discovered effect and especially to become independent of potentially unreliable individual features of the initial representative, additional representative genes and/or samples are searched for and combined for effect axes estimation until convergence. Resulting converged axes are considered the final linear approximation of the effect's law of regulation. These axes usually put an emphasis on strong regulations by the effect, as they are often associated with higher correlations. Effect axes serve as basis for later bimonotonic regression<sup>(II.4)</sup>.

Their convergence and in particular the search for additional representative genes or samples must be guided. Samples from possibly overlapping foreign effects should not be selected to prevent uninterpretable mixtures of true effects. In case of suboptimal initial candidates that are located between two effects, convergence should guide away from the intermediate space and towards a single particular effect to dissect it and only it. I again utilize effect scores<sup>(II.3.1.6)</sup> and select those representatives that maximize it, i.e. I utilize *maximization of both correlation and effect size* as guide for convergence.

In the example<sup>(Figure II.3.2)</sup> the converged gene axis<sup>(yellow)</sup> is nearer to all blue points compared to the gene axis based only on the initial representative<sup>(Figure II.3.1)</sup>. Due to the nonlinear shape of the blue effect, no linear approximation can be a perfect match, but it can serve as starting point for regression of precise bimonotonic effect curves<sup>(II.4.1)</sup>. If an effect's law of regulation is actually linear like for the red effect, effect axes and effect curves are equal.

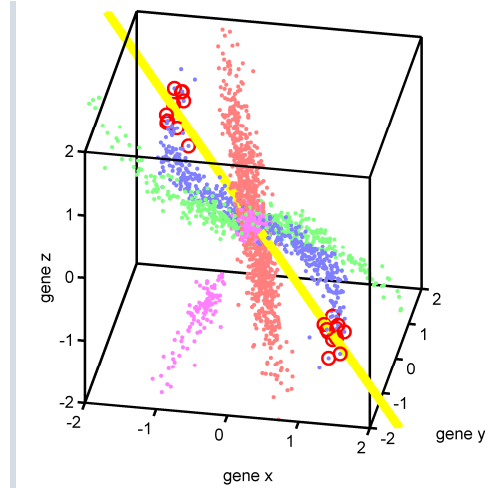


Figure II.3.2) 3D concept example, first detection iteration, generalizing representatives and converged gene axis for the blue effect

### II.3.2.1 Iterative selection of representatives

Let  $(|b_l^g\rangle, |b_l^s\rangle)$  denote current effect axes based on  $l$  representatives,  $(|w_l^g\rangle, |w_l^s\rangle)$  the current effect focus and  $(|r_l^g\rangle, |r_l^s\rangle)$  current correlations. For  $l = 1$ , these vectors equal the initial effect axes<sup>(Eqn. II.3.1.3)</sup> respectively the refined effect focus<sup>(Eqn. II.3.1.4.b)</sup> respectively the initial correlations<sup>(Eqn. II.3.1.4.a)</sup>.

At the beginning of each iteration, all genes and all samples are sorted descending by their absolute correlations  $|r_{i,l}^g|$  respectively  $|r_{j,l}^s|$  in a single joint list (again using component abbreviations  $r_{i,l}^g \equiv \langle e_i^g | r_l^g \rangle$  and  $r_{j,l}^s \equiv \langle e_j^s | r_l^s \rangle$ ). Points in the gene or sample space that most likely belong to the same effect are thus sorted to the top. Roles of genes and samples continue to be symmetric.

For a lookahead<sup>(similar to II.3.1.7)</sup> interval, top correlated points from this list (in the default 20) are candidate-added to the current set of effect representatives and are accumulated<sup>(detailed in II.3.2.2 unterhalb)</sup>; this results in

candidates for the next effect axes ( $|b_{i+1}^g\rangle, |b_{i+1}^s\rangle$ ), the next effect focus ( $|w_{i+1}^g\rangle, |w_{i+1}^s\rangle$ ) and the next correlations ( $|r_{i+1}^g\rangle, |r_{i+1}^s\rangle$ ).

Again, the effect score<sup>(Eqn. II.3.1.6)</sup> is computed for these updated axes. The candidate leading to the best effect score is selected as next representative. Often points with stronger folding are selected, as they are more likely to reach higher correlations to each other assuming a constant noise level and thus can lead to higher effect scores. However, the effect size factor in the score prevents selection of top-folded points, if they have no correlated partners.

Iterations continue until axes have converged and sufficiently many representatives have been selected<sup>(as defined in II.3.2.3 unterhalb)</sup>. Then the algorithm passes these converged axes on to effect regression and dissection<sup>(II.4)</sup>.

### II.3.2.2 Accumulating representatives and the update step

Let ( $|a_{i'}^g\rangle, |a_{i'}^s\rangle$ ) denote initial gene and sample axes<sup>(Eqn. II.3.1.3)</sup> for each so far selected representative gene or sample  $i' = 1 \dots l + 1$ , including the currently selected candidate  $l + 1$ . (The constant effect index  $k$  is suppressed for readability.) Then the updated accumulated axes are defined as their weighted arithmetic averages, using correlations to current accumulated axes as weights. This is more precise than assigning equal weights to all representatives, because effect axes may wander during convergence and, hence, previously selected representatives may become less correlated and thus less representative for the effect. Representative genes or samples may be either highly correlated or highly anti-correlated to the effect and to each other. To always accumulate them constructively, correlation signs are kept and utilized to align all representatives to each other in the sum<sup>(Eqn. II.3.2.2)</sup>.

The **accumulated sample axis**  $|b_{i+1}^s\rangle \in V^s$  for individual sample axes  $|a_{i'}^s\rangle$  of all so far selected representatives is defined as:

$$|b_{i+1}^s\rangle \equiv \frac{\sum_{i'=1}^{l+1} [a_{i'}^s | b_i^s]_{|w_i^s\rangle} |a_{i'}^s\rangle}{\sum_{i'=1}^{l+1} |[a_{i'}^s | b_i^s]_{|w_i^s\rangle}|}$$

The **accumulated gene axis**  $|b_{i+1}^g\rangle \in V^g$  for individual gene axes  $|a_{i'}^g\rangle$  of all so far selected representatives is defined as:

$$|b_{i+1}^g\rangle \equiv \frac{\sum_{i'=1}^{l+1} [a_{i'}^g | b_i^g]_{|w_i^g\rangle} |a_{i'}^g\rangle}{\sum_{i'=1}^{l+1} |[a_{i'}^g | b_i^g]_{|w_i^g\rangle}|}$$

Eqn. II.3.2.2) Accumulated effect axes over  $l$  effect representatives

Next, correlations ( $|r_{i+1}^g\rangle, |r_{i+1}^s\rangle$ ) for all genes and samples are computed for updated axes, analogous to initial correlations<sup>(Eqn. II.3.1.4.a)</sup>, but using the current effect focus ( $|w_i^g\rangle, |w_i^s\rangle$ ) as weights.

Finally, the effect focus is updated with the same formula that refined the initial effect focus<sup>(Eqn. II.3.1.4.b)</sup>, using current correlations ( $|r_{i+1}^g\rangle, |r_{i+1}^s\rangle$ ) and their significance.

Then the effect score is computed<sup>(Eqn. II.3.1.6)</sup> that also facilitates the identification of the best candidate for representative  $l$  from the lookahead interval<sup>(cf. II.3.2.1)</sup>.

### II.3.2.3 Checking for sufficient representatives and for convergence

Two criteria must be met before linear axes are considered sufficiently reliable and representative for an effect.

The convergence criterion checks that the change by addition of the last representative is below a threshold parameter  $\epsilon \equiv 10^{-4}$ . The change can be quantified by one minus the correlation to the preceding axis:  $\delta^g \equiv 1 - [b_i^g | b_{i+1}^g]_{|w_{i+1}^g\rangle}$  respectively  $\delta^s \equiv 1 - [b_i^s | b_{i+1}^s]_{|w_{i+1}^s\rangle}$ . (The correlations are always positive per construction of accumulated axes.) The precise condition is  $(\delta^g + \delta^s)/2 < \epsilon$ . In words, axes are considered converged, if

adding another representative gene or sample keeps them correlated with preceding axes  $> 0.999$  on average. As every additional representative can only change axes with a maximal weight of  $1/l$  (cf. Eqn. II.3.2.2), this always converges sooner or later in practice. (For the theoretical possibility of a convergence point  $l > m + n$  that never occurred for any test scenario (cf. II.6), all genes and samples are selected and a warning is issued.)

The second criterion demands a minimum amount of selected representatives. This is important, because if two nearly identical candidates exist in the raw signal by accident, the convergence criterion might already be reached for  $l = 2$ , but this is usually not yet representative nor robust. In the default, the minimum of fifteen and of 20% of the estimated effect size is required. The effect size is estimated as for the effect score (II.3.1.6). 20% are considered sufficiently representative for any effect. Fifteen has proven to be already sufficient in all test cases and there is no point to add more representatives and waste computation resources, if axes have converged. The relative cut at 20% is also important to demand less than fifteen representatives for very small effects that might not even have this many members (in this case, forcing more representatives would make effect axes less specific and less representative).

Once both criteria are met for a certain  $\hat{l}$ , effect axes  $(|b_i^g|, |b_i^s|)$  are considered the final linear approximation of the effect's law of regulation and thus are fixed hereafter. They are passed on to bimonotonic regression, where they serve as axes over that precise effect curves are monotonically regressed (II.4).



## II.4 Regression and Dissection

Once an effect has been pinpointed by its gene and sample axes as determined by the search strategy, the next step is to estimate and extract its and only its contributions to the signal in order to dissect it. This extraction must be precise, because if the effect's eigensignal is underestimated, hard-to-interpret shadows of the same effect would be detected later. If it is over-compensated, artifacts would be introduced, likewise leading to hard-to-interpret secondary effects. If it is compensated in a too fine-grained way, information about other overlapping effects would be lost, preventing their detection.

The bimonotonic effect model<sup>(cf. II.2.1.2)</sup> allows for a precise regression of the effect's eigensignal. Together with dissection strengths determined by correlations to the effect, the effect can be removed from the signal in a way that leaves signals from overlapping and even partially correlated foreign effects untouched. Additionally, effect curves based on the eigensignal can replace previous linear approximations with more precise monotonic approximations of the effect's law of gene regulation.

### II.4.1 Regression and Effect Curves

Based on discovered and representative effect axes  $(|b_i^g\rangle, |b_i^s\rangle)$ <sup>(cf. II.3.2.3)</sup>, the bimonotonic effect model<sup>(cf. II.2.1.2)</sup> is utilized to realize a corresponding regression of the effect's eigensignal. This eigensignal can be re-parameterized as effect curves<sup>(cf. II.2.2.3)</sup> in the gene or sample space. The resulting gene curve for the blue effect in the concept example demonstrates a much more precise approximation of the effect's nonlinear law of gene regulation<sup>(Figure II.4.1)</sup> compared to its previous linear approximation by effect axes<sup>(cf. Figure II.3.2)</sup>.

An overview of this important sub-algorithm is presented here, then individual steps are clarified in detail below. Eigensignals are regressed by one outer and one inner convergence loop. Every outer regression iteration  $i$  is structured as follows:

- II.4.1.1: Compute *effect strengths* for genes  $|u_{k,i}^g\rangle$  and for samples  $|u_{k,i}^s\rangle$
- II.4.1.2: Resort the current signal  $M_{k-1}$  to the effect's *empirical eigenorder*  $(I_{k,i}, J_{k,i})$  (as determined by effect strengths) in order to obtain a roughly bimonotonic presorting.
- II.4.1.3: Inner loop (index  $j$ ) for *bimonotonic regression of the empirical eigensignal*  $M_{k-1}(I_{k,i}, J_{k,i})$ . Results in  $\mathcal{M}_{i,\hat{j}}$  for convergence iteration  $\hat{j}$ .
- II.4.1.4: *Adaptive smoothing*  $\mathcal{S}$  by rescaling to the effect strength space and by 2D Fourier transforms.
- II.4.1.5: Convergence check. As soon as converged (index  $\hat{i}$ ), the result  $\mathcal{S}(\mathcal{M}_{\hat{i},\hat{j}})$  is passed on to compute the final eigensignal and to dissect the effect<sup>(II.4.2)</sup>.

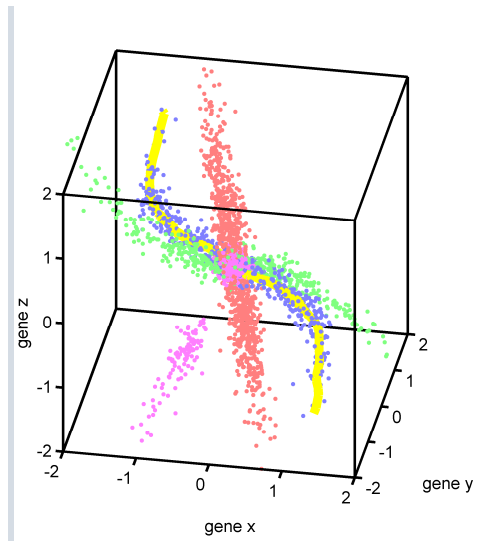


Figure II.4.1) 3D example, blue effect curve in gene space

Before the first bimonotonic regression ( $i = 0$ ), effect strengths are computed by projecting all genes on the effect's representative sample axis  $|b_i^s\rangle$  and all samples on the effect's representative gene axis  $|b_i^g\rangle$  that were determined by the search strategy. The current signal context for these projections and for genes  $|g_i\rangle$  and samples  $|s_j\rangle$  is still the signal matrix  $M_{k-1}$ . (For clarity, definitions below are also written out in terms of matrix based vector components.)

**Initial effect gene strengths**  $|u_{k,1}^g\rangle \in V^g$  are defined by weighted projections<sup>(II.2.3.2)</sup> in the sample focus  $|w_i^s\rangle$ :

$$\langle e_i^g | u_{k,1}^g \rangle \equiv \langle g_i | b_i^s \rangle_{|w_i^s\rangle}^0 = \langle M_{k-1}(i, J_0) | b_i^s \rangle_{|w_i^s\rangle}^0 / \|w_i^s\|$$

**Initial effect sample strengths**  $|u_{k,1}^s\rangle \in V^s$  are defined by weighted projections<sup>(II.2.3.2)</sup> in the gene focus  $|w_i^g\rangle$ :

$$\langle e_j^s | u_{k,1}^s \rangle \equiv \langle s_j | b_i^g \rangle_{|w_i^g\rangle}^0 = \langle M_{k-1}(I_0, j) | b_i^g \rangle_{|w_i^g\rangle}^0 / \|w_i^g\|$$

Eqn. II.4.1.1.a) Effect strengths by weighted projections on effect axes for  $i = 1$

Of note, these projections may effectively also be viewed as *classifications* of genes respectively samples with respect to an effect. Alternatively, correlations to the effect could be utilized for later classification purposes<sup>(cf. III.1.4)</sup>.

For regression iterations  $i > 1$ , effect curves<sup>(II.2.2.3)</sup> are employed instead of effect axes for higher precision. These curves are based on the smoothed and regressed eigensignal version  $\tilde{E}_{k,i-1} \equiv \mathcal{S}(\mathcal{M}_{i-1,\hat{\beta}})$  from the previous iteration  $i - 1$  (as defined and computed in the following sections; the corresponding matrix in reference order is again denoted without tilde, i.e.  $E_{k,i-1}(I_{k,i-1}^g, J_{k,i-1}^s) = \tilde{E}_{k,i-1}$ .)

**Refined effect gene strengths**  $|u_{k,i}^g\rangle \in V^g$  are defined by weighted projections<sup>(II.2.3.2)</sup> in the sample focus  $|w_i^s\rangle$ :

$$\langle e_i^g | u_{k,i}^g \rangle \equiv \langle g_i | c_{i-1}^s(i) \rangle_{|w_i^s\rangle}^0 = \langle M_{k-1}(i, J_0) | E_{k,i-1}(i, J_0) \rangle_{|w_i^s\rangle}^0 / \|w_i^s\|$$

**Refined effect sample strengths**  $|u_{k,i}^s\rangle \in V^s$  are defined by weighted projections<sup>(II.2.3.2)</sup> in the gene focus  $|w_i^g\rangle$ :

$$\langle e_j^s | u_{k,i}^s \rangle \equiv \langle s_j | c_{i-1}^g(j) \rangle_{|w_i^g\rangle}^0 = \langle M_{k-1}(I_0, j) | E_{k,i-1}(I_0, j) \rangle_{|w_i^g\rangle}^0 / \|w_i^g\|$$

Eqn. II.4.1.1.b) Effect strengths by weighted projections on effect curves for  $i > 1$

I utilize defined effect strengths<sup>(II.4.1.1)</sup> as empirical estimate for the effect eigenorder:

The **empirical gene eigenorder**  $I_{k,i}^g$  of the effect orders genes by their effect strengths  $\langle e_i^g | u_{k,i}^g \rangle$ , i.e.

$$\forall i \in [1, m - 1]: \langle e_{i+1}^g | u_{k,i}^g \rangle \geq \langle e_i^g | u_{k,i}^g \rangle.$$

The **empirical sample eigenorder**  $J_{k,i}^s$  of the effect orders samples by their effect strengths  $\langle e_j^s | u_{k,i}^s \rangle$ , i.e.

$$\forall j \in [1, n - 1]: \langle e_{j+1}^s | u_{k,i}^s \rangle \geq \langle e_j^s | u_{k,i}^s \rangle.$$

Eqn. II.4.1.2) Empirical effect eigenorder based on effect strengths

The correspondingly permuted signal matrix  $M_{k-1}(I_{k,i}^g, J_{k,i}^s)$  orders genes and samples by the discovered effect, resulting in a roughly bimonotonic signal already. This raw eigensignal provides the basis for precise regression below<sup>(II.4.1.3)</sup>. Notably, the final eigensignal cannot be defined with this empirical signal directly, because after subtracting it, the remaining signal would equal zero. Hence, all information about other possibly overlapping effects with different eigenorders would be lost.

Bimonotonic regression is realized based on 1D monotonic regressions, weighted averages and a convergence loop of the following structure:

- weighted monotonic 1D regression of each gene in the sample eigenorder  $J_{k,i}^S$  of the effect
- weighted monotonic 1D regression of each sample in the gene eigenorder  $I_{k,i}^G$  of the effect
- 2D Fourier smoothing of the signal (utilized for points with no or low weight in the effect focus)
- compute weighted averages
- check for convergence.

To realize 1D monotonic regressions for each gene and for each every sample, an established isotonic regression algorithm is utilized called Generalized Pool Adjacent Violator (GPAV)<sup>[84]</sup>. The algorithm has  $O(n^2)$  (respectively  $O(m^2)$ ) computational complexity in the worst case, but approaches linear complexity  $O(n)$  for already monotonically presorted data. Hence and in practice, it is much faster than the worst case, because genes and samples are already roughly presorted in the signal matrix  $M_{k-1}(I_{k,i}^G, J_{k,i}^S)$ . Importantly, this algorithm supports weights, which is utilized to put *regression emphasis on signals in the effect focus*.

Let  $j$  denote the iteration index of this inner convergence loop. Rather than using gene weights  $|w_i^g\rangle$  for the regression of each sample (or sample weights  $|w_i^s\rangle$  for the regression of each gene), every gene and every sample gets its own weights vector for regression as follows. Let  $W_j^{\tilde{G}} \in \mathbb{R}^{m \times n}$  represent sample weights for each gene and let  $W_j^{\tilde{S}} \in \mathbb{R}^{m \times n}$  denote gene weights for each sample. (The tilde indicates matrices in eigenorder rather than in reference order; see their following definition.) Both weight matrices are initialized at  $j = 0$  (i.e. before the first regression) with the symmetric outer product of the effect focus:

The **product effect focus**  $W \in \mathbb{R}^{m \times n}$  (of effect  $k$ ) is defined as components of the outer product  $|w_i^g\rangle \otimes |w_i^s\rangle$  of the final effect focus  $(|w_i^g\rangle, |w_i^s\rangle)$  (cf. II.3.2).

**Regression weights** for the initial inner iteration  $j = 0$  are initialized with this product effect focus in the current eigenorder  $(I_{k,i}^G, J_{k,i}^S)$ :  $W_0^{\tilde{D}} \equiv W(I_{k,i}^G, J_{k,i}^S)$ .

Eqn. II.4.1.3.a) The product effect focus and initial weights for 1D regressions

These product weights work like a logical AND-condition. Only (gene, sample) pixels associated with high gene weight  $\langle e_i^g | w_i^g \rangle$  and high sample weight  $\langle e_i^s | w_i^s \rangle$  also get high weights for regression and subsequent averaging (as detailed below).

Let  $\mathcal{M}_{i,j}$  denote the signal of outer iteration  $i$  and at the beginning of inner regression iteration  $j$ . Before the first regression ( $j = 0$ ),  $\mathcal{M}_{i,0}$  is initialized as the roughly bimonotonic signal  $M_{k-1}(I_{k,i}^G, J_{k,i}^S)$  in the empirical effect eigenorder. (For readability, the outer index  $i$  is suppressed for all local variables like weight matrices in this section.)

Now GPAV is applied to every gene row in  $\mathcal{M}_{i,j}$ , using corresponding rows in  $W_j^{\tilde{D}}$  as regression weights. This results in monotonically regressed genes that can be collected as matrix  $\tilde{G}_{j+1} \in \mathbb{R}^{m \times n}$ . Each gene row in  $\tilde{G}_{j+1}$  is a step function that consists of blocks of constant regressed gene expressions, while expressions of neighboring blocks are all either monotonically increasing or all monotonically decreasing. Each block corresponds to a sample interval in the sample eigenorder  $J_{k,i}^S$  of the effect. GPAV also updates sample weights (for each gene) by averaging input weights for each block; they can be collected as rows of a matrix  $W_{j+1}^{\tilde{G}}$ . Likewise, regressions of sample columns in  $\mathcal{M}_{i,j}$  are realized with GPAV, using corresponding columns of  $W_j^{\tilde{D}}$  as regression weights. This results resulting in a matrix of monotonic columns  $\tilde{S}_{j+1} \in \mathbb{R}^{m \times n}$  and updated gene weights for each sample  $W_{j+1}^{\tilde{S}} \in \mathbb{R}^{m \times n}$ .

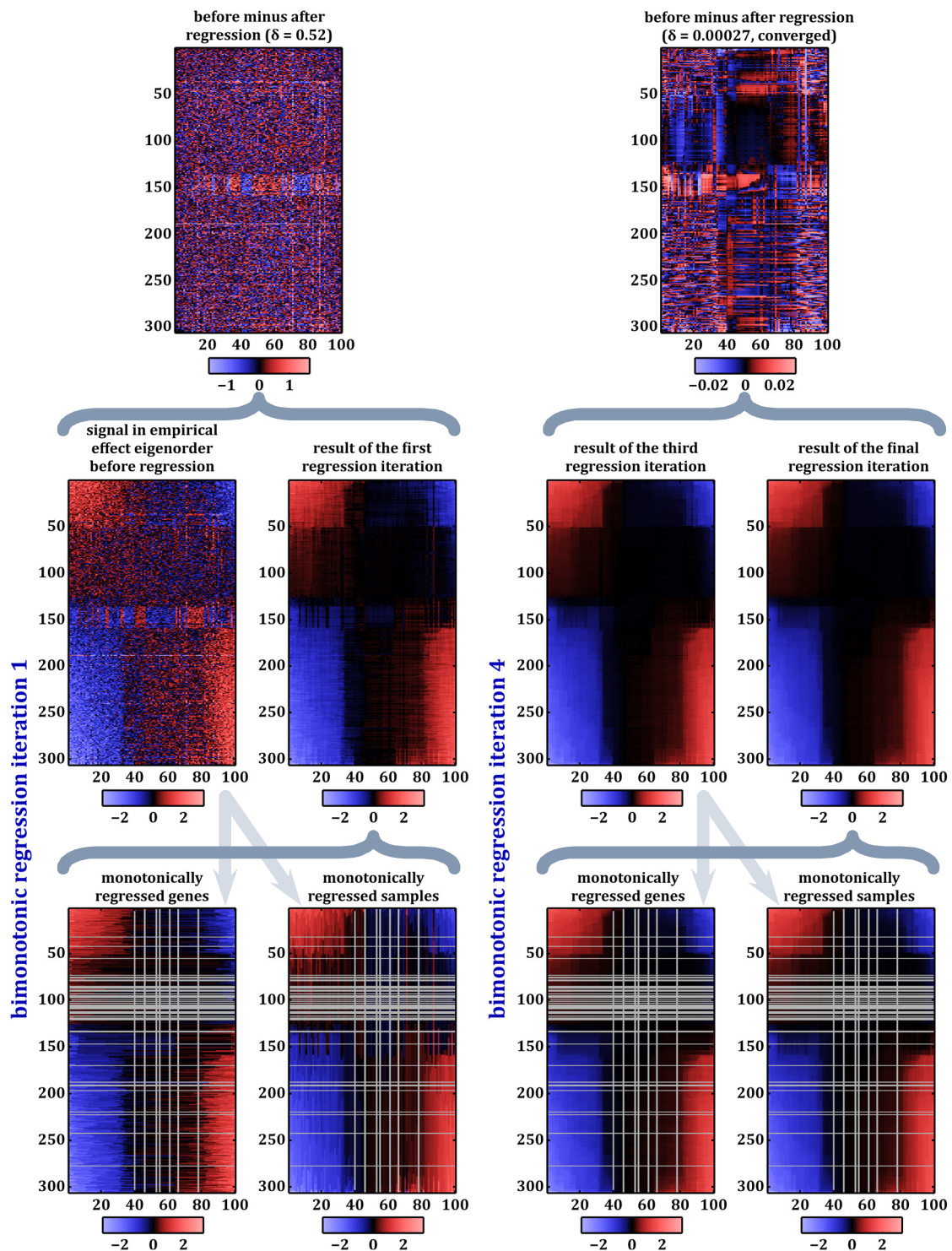


Figure II.4.1.3) Illustration of bimonotonic regression (inner convergence loop)

Starting from the presorted signal matrix  $M_{k-1}(I_{k,i}^g, J_{k,i}^s)$  in effect eigenorder, first genes and samples in the effect focus are monotonically regressed (bottom left). Results are combined via weighted averaging to obtain  $M_{i,1}$  (center row, second panel). This continues until the final iteration  $\hat{j} = 4$  that is depicted in the right half. The top row illustrates differences  $M_{i,j+1} - M_{i,j}$ ; color bars indicate the scaling. Provided  $\delta$  values are pixel standard deviations of depicted matrices in the effect focus. (Most of the 1000 simulated genes are not depicted, as they have zero weight in the effect focus. Hence, the effect eigensignal for them is determined solely by adaptive smoothing. Similarly, gray lines indicate genes and samples with zero weights because of their relatively low correlation to the effect's sample or gene axis compared to existing top correlated genes respectively samples.)

To interweave gene and sample regressions over iterations  $j$ , weights are mixed. To realize a soft convergence, additionally an average with a history of one iteration is computed. This average respects

imbalances in gene and sample counts (to prevent weights based on just a few samples to overwrite much more granular information from many genes, for example). The concrete definition of mixed weights reads:

$$W_{j+1}^{\bar{2D}} \equiv \frac{(nW_j^{\bar{2D}} + mW_{j+1}^S)}{2(n+m)} + \frac{(mW_j^{\bar{2D}} + nW_{j+1}^G)}{2(n+m)}$$

Eqn. II.4.1.3.b) Interweaved regression weights for genes and samples

Typically effects do not affect all genes and all samples in the signal. Often only a fraction of all measured genes are correlated to an effect. Hence, the effect focus  $(|w_i^g|, |w_i^s|)$  typically assigns zero weights to many genes and maybe also several samples. For zero weights, an adaptive smoothing of  $M_{k-1}(I_{k,i}^g, J_{k,i}^s)$  is employed instead of regressions to define the effect's eigensignal. In this way, usually most genes do not need to be regressed at all, which increases performance considerably. To this end, the same smoothing operator  $\mathcal{S}$  is applied to  $\mathcal{M}_{i,\hat{j}}$  as for final eigensignal polishing<sup>(cf. II.4.1.4)</sup>. This support for zero weights also has the useful side effect to support missing values in the input signal<sup>(cf. II.6.4.4)</sup>.

Finally and using the same interweaved weights as above, the next estimate for the bimonotonic eigensignal is formed as weighted average:

$$\mathcal{M}_{i,\hat{j}+1} = \frac{(nW_j^{\bar{2D}} + mW_{j+1}^S)}{2(n+m)} \cdot \tilde{G}_{j+1} + \frac{(mW_j^{\bar{2D}} + nW_{j+1}^G)}{2(n+m)} \cdot \tilde{S}_{j+1} + (1 - W_{j+1}^{\bar{2D}}) \cdot \mathcal{S}(\mathcal{M}_{i,\hat{j}})$$

Eqn. II.4.1.3.c) Bimonotonic regression step

The third summand gradually fills in the adaptively smoothed signal for all weights less than one. Hence, all pixels keep their norm, irrespective of how strongly correlated they are to the effect.

To illustrate this regression procedure, the first and last iterations of eigensignal regression for the discovered pattern #3 of a versatility test<sup>(cf. II.6.2)</sup> are depicted<sup>(Figure II.4.1.3)</sup>.

For convergence estimation, the pixel standard deviation of  $\mathcal{M}_{i,\hat{j}+1} - \mathcal{M}_{i,\hat{j}}$  can be utilized (weighted with the product effect focus  $W$ <sup>(Eqn. II.4.1.3.a)</sup>). Let  $\hat{j}$  denote the iteration as soon as this standard deviation is  $< \epsilon \hat{\sigma}(\mathcal{N}_{k-1})$ , where  $\epsilon \equiv 10^{-3}$  and  $\hat{\sigma}(\mathcal{N}_k)$  is the current estimated noise level<sup>(cf. II.5.1.1)</sup>. Then  $\mathcal{M}_{i,\hat{j}}$  is considered bimonotonically regressed to sufficient precision and passed on towards signal polishing<sup>(II.4.1.4)</sup>.

#### II.4.1.4 Adaptive smoothing by 2D Fourier transforms

Bimonotonic regression has already resulted in a signal matrix  $\mathcal{M}_{i,\hat{j}}$  that *fulfills the effect model*<sup>(II.2.1.2)</sup>. Adaptive smoothing alone could not have provided this, as monotonicity is a global property and smoothing is a local operation. Even for very large smoothing kernels, bimonotonicity could in general only be approximated and furthermore such blurry smoothing would prevent a precise dissection (in the limit of ever larger smoothing kernel sizes, the result would simply approach a constant). However, for relatively small and adaptive kernel sizes as explained below, smoothing has been found to increase quality over the step-function-like pure bimonotonic regression result  $\mathcal{M}_{i,\hat{j}}$ . As smoothing cannot change monotonicity, bimonotonicity is kept by this operation. Additionally and as explained in the last section, the same smoothing operation is also utilized for performance reasons when replacing the relatively slow regression by this smoothing for genes and samples with zero weight in the effect focus.

It is counterproductive to directly smooth  $\mathcal{M}_{i,\hat{j}}$  (or  $M_{k-1}(I_{k,i}^g, J_{k,i}^s)$  before regression), because neighboring genes or samples in the empirical eigenorder might have strongly different effect strengths. Smoothing such neighbors would result in an eigensignal that, if subtracted, *undercompensated* the stronger neighbor and *overcompensated* the weaker. Both would result in inaccurate approximations of the effect's law of regulation

and trigger hard-to-interpret secondary detections of signal remnants from this effect after its inevitable imprecise dissection. Hence, in case of rapid signal changes a very small (and for abrupt signal changes even zero-width) smoothing kernel is required to obtain a sharp eigensignal for the effect.

Furthermore, in eigenorder intervals with nearly constant effect strengths, it makes more sense to employ a large kernel that smooths many neighbors, because all of them contribute equally to the effect and hence should result in the same eigensignal for balanced dissection.

Unfortunately, the direct computation of a smoothing operation with an adaptive kernel size is computationally quadratic *in the number of pixels* and hence not feasible for practical purposes.

To realize this adaptive smoothing fast,  $\mathcal{M}_{i,j}$  is first *rescaled* using effect strengths  $\{u_{k,i}^g\}$  and  $\{u_{k,i}^s\}$  (cf. II.4.1.1). A constant smoothing kernel in this rescaled effect strengths space then corresponds to the requirement of an adaptive smoothing in the original eigenorder index space. For performance reasons, the resolution of this rescaled space is set to  $m^- \equiv 512$  rows and  $n^- \equiv 512$  columns respectively. This is sufficiently precise to represent a bimonotonic signal for all practical purposes. (Choosing resolutions as powers of two allows an

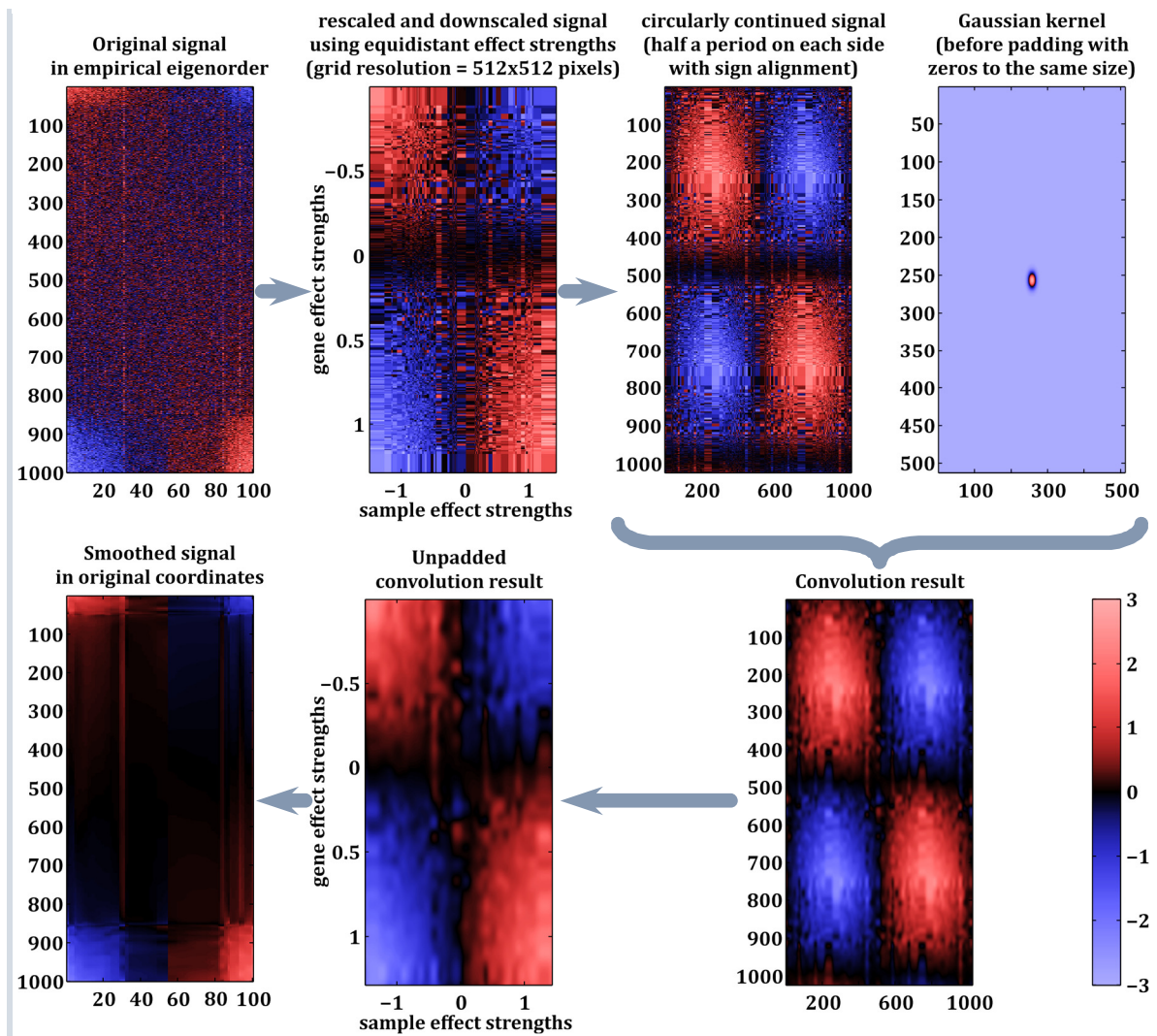


Figure II.4.1.4) Illustration of adaptive signal smoothing (realized by rescaling and 2D Fourier transformation)

Starting from the presorted signal matrix  $M_{k-1}(\{u_{k,i}^g\}, \{u_{k,i}^s\})$  in eigenorder, first the signal is rescaled and downsampled by aggregation and interpolation in the space with equidistant effect strengths (resolution: 512x512 pixels). To avoid border artefacts from 2D Fourier transforms the signal is circularly padded with sign alignment as depicted. With a Gaussian kernel and the convolution theorem described in the text, the smooth result on the lower right is obtained (i.e. the convolution). After unpadding and rescaling back to original coordinates by interpolation, the smoothed version of the start signal is obtained.

optimally efficient fast Fourier transformation below.) Coordinates of this rescaled and downscaled space correspond to *equidistant* effect strengths:

Basis vectors  $\{e_{i^-}^{g^-}\}$  of the **rescaled and downscaled gene space** correspond for  $i^- \in [1, m^-]$  to equidistant gene effect strengths  $u_{k,i}^g \left( I_{k,i}^g(1) \right) + \left( -u_{k,i}^g \left( I_{k,i}^g(1) \right) + u_{k,i}^g \left( I_{k,i}^g(m) \right) \right) \cdot \frac{i^-}{m^-}$ .

Basis vectors  $\{e_{j^-}^{s^-}\}$  of the **rescaled and downscaled sample space** correspond for  $j^- \in [1, n^-]$  to equidistant sample effect strengths  $u_{k,i}^s \left( J_{k,i}^s(1) \right) + \left( -u_{k,i}^s \left( J_{k,i}^s(1) \right) + u_{k,i}^s \left( J_{k,i}^s(n) \right) \right) \cdot \frac{j^-}{n^-}$ .

Eqn. II.4.1.4.a) Rescaled and downscaled gene and sample space for smoothing

The rescaled and downscaled signal  $\mathcal{M}_{i,\hat{j}}^- \in \mathbb{R}^{m^- \times n^-}$  can be computed from  $\mathcal{M}_{i,\hat{j}}$  by averaging in the corresponding effect strength grid cells (that are bordered on the right by defined effect strength cuts). In case of empty grid cells (i.e. no gene and no sample has an effect strength in the corresponding intervals), nearest-neighbor interpolation is employed.

This coordinate change does not only increase the resolution at steep changes of effect strengths (i.e. where the eigensignal also needs to change rapidly) but also reduces resolution for intervals with relatively flat effect strengths (i.e. where the eigensignal should also change little). Therefore,  $\mathcal{M}_{i,\hat{j}}^-$  can now be smoothed in this rescaled space using a *constant* window width, while still fulfilling the requirement of adaptive widths in the original index space.

Let  $G_{\sigma^{g^-}, \sigma^{s^-}} \in \mathbb{R}^{m^- \times n^-}$  denote the Gaussian kernel (centered at indices  $\left(\frac{m^-}{2}, \frac{n^-}{2}\right)$ ) with effect strength standard deviations corresponding to eight pixels:  $\sigma^{g^-} \equiv \left( -u_{k,i}^g \left( I_{k,i}^g(1) \right) + u_{k,i}^g \left( I_{k,i}^g(m) \right) \right) \cdot \frac{8}{512}$  and  $\sigma^{s^-} \equiv \left( -u_{k,i}^s \left( J_{k,i}^s(1) \right) + u_{k,i}^s \left( J_{k,i}^s(n) \right) \right) \cdot \frac{8}{512}$ . Eight pixels are sufficiently many to avoid numeric artefacts. And  $\frac{8}{512}$  is small enough to avoid problems with over- and under-compensation in the original space as explained above.

The smoothing result is the convolution  $\mathcal{M}_{i,\hat{j}}^- * G_{\sigma^{g^-}, \sigma^{s^-}}$ . Unfortunately its naïve computation has still a complexity that is quadratic in the number of points. Therefore, the convolution theorem<sup>[85 §15.3.1.3]</sup> is applied:

$$\mathcal{M}_{i,\hat{j}}^- * G_{\sigma^{g^-}, \sigma^{s^-}} = \mathcal{F}^{-1} \left( \mathcal{F}(\mathcal{M}_{i,\hat{j}}^-) \cdot \mathcal{F}(G_{\sigma^{g^-}, \sigma^{s^-}}) \right)$$

Eqn. II.4.1.4.b) Application of the convolution theorem reduces the smoothing task to 2D Fourier transforms and component-wise multiplication

This reduces the smoothing task to the component-wise multiplication of two Fourier-transformed matrices plus one inverse Fourier transform of the result. Fast Fourier transform implementations for  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  (provided by **fft2** and **ifft2** functions in MATLAB®) are employed that only have log-linear complexity.

To avoid numeric artifacts from border effects,  $\mathcal{M}_{i,\hat{j}}^-$  is circularly continued for half its size with sign alignment. This is implemented by padding of the signal matrix and corresponding unpadding after the inverse transformation<sup>(see Figure II.4.1.4)</sup>.

Finally, the unpadded convolution result is scaled back to original coordinates via 2D interpolation at effect strengths  $\langle e_i^g | u_{k,i}^g \rangle$  and  $\langle e_j^s | u_{k,i}^s \rangle$ . For abbreviation, let  $\mathcal{S}(\mathcal{M}_{i,\hat{j}}^-)$  summarize this smoothing operation.

Notably, while the result is already a good approximation of the eigensignal (and was even employed to estimate it in an earlier development stage of the algorithm), this approximation it is not necessarily bimonotonic and may contain hard to interpret parts, e.g. the visible red stripes<sup>(Figure II.4.1.4)</sup>. Especially for strong effect strengths (where eigensignals are also strong) precision is paramount to avoid introduction of artifacts that might be discovered as “pseudo effects” in later iterations. Additionally, this smoothing operation is not weighted, i.e. it cannot respect the effect focus. Only by enforcing bimonotonicity via weighted regression in the effect focus<sup>(II.4.1.3)</sup> and by using dissection strengths<sup>(II.4.2.1)</sup> in order to zero the eigensignal outside of the effect focus, the aspired precision and interpretability of resulting effects was obtained.

Convergence of the outer regression loop is finally checked by correlating all pixels of  $\mathcal{S}(\mathcal{M}_{i,\hat{j}})$  with the result from the previous iteration  $\mathcal{S}(\mathcal{M}_{i-1,\hat{j}})$ , again using the symmetric product effect focus<sup>(Eqn. II.4.1.3.a)</sup> as weights. The eigensignal is considered converged as soon as this correlation is  $>0.99$ . Let  $\hat{i}$  denote this iteration; then  $\mathcal{S}(\mathcal{M}_{i,\hat{j}})$  is passed on to the computation of the final eigensignal for effect dissection<sup>(II.4.2)</sup>.

Convergence is always reached, as the only change between outer iterations is the update of the empirical eigenorder by projections on regressed effect curves. In practice, often only the first iteration (from effect axes to the first effect curves) is associated with a relatively large change. After that, it usually takes only one additional outer iteration to reach the correlation threshold. If the effect is linear, it may even converge after the first outer iteration, as in this case effect curves equal effect axes.

After having bimonotonically regressed and smoothed the signal in the discovered effect eigenorder,  $\mathcal{S}(\mathcal{M}_{i,\hat{j}})$  is multiplied by dissection strengths based on the effect focus to obtain the final effect eigensignal.

The effect focus  $(|w_i^g|, |w_i^s|)$  obtained by the search strategy<sup>(cf. II.3.2)</sup> serves not only a purposes during regression but also during dissection. The first purpose was to define *weights for regression*<sup>(II.4.1.3)</sup> in order to estimate effect curves in the ideal case exclusively based on genes and samples *in the effect*, i.e. by those that are indeed significantly regulated by the discovered effect.

The second purpose during dissection concerns partially correlated effects, i.e. effects that share dimensions with partially co-ordered eigenorders. (For example, green samples in the 3D example<sup>(see e.g. Figure II.4.1)</sup> may show strong effect strengths when projected on the red effect axis, but simultaneously have high distance from the red effect.) In this case, even with a perfectly regressed effect curve for the current effect, effects cannot be dissected cleanly, as is demonstrated in this subchapter<sup>(Figure II.4.2.2.b)</sup>. Hence, the second purpose of the effect focus is to prevent explaining signal parts by for genes or samples that are significantly *out* of the current effect. Only then, these partially correlated effects can be detected and dissected in later iterations with effect curves that are specific to their *original* signal (rather than with hard to interpret effect curves that only fit remnants of their signal and effectively depend on all previously dissected effects).



### II.4.2.1 Dissection strengths, final eigensignal and remaining signal

Dissection strengths are essentially identical to the effect focus and guarantee that no signal is modified outside of the effect's focus. Like regression weights before (Eqn. II.4.1.3.a), dissection strengths are defined as the symmetric product effect focus in eigenorder.

**Dissection strengths**  $D \in \mathbb{R}^{m \times n}$  (for effect  $k$ ) are defined as square roots of the components of  $|w_i^g| \otimes |w_i^s|$ , i.e. of the outer product of the final effect focus  $(|w_i^g|, |w_i^s|)$  (cf. II.3.2).

$\tilde{D} \equiv D(I_{k,\hat{\lambda}}^g, J_{k,\hat{\lambda}}^s)$  denotes dissection strengths in the final eigenorder.

Eqn. II.4.2.1.a) Dissection strengths

Now the final eigensignal of effect  $k$  can be defined.

The **effect eigensignal** in the discovered eigenorder is defined as the component-wise product:

$$\tilde{E}_k \equiv \tilde{D} \cdot \mathcal{S}(\mathcal{M}_{\hat{\lambda}, \hat{\lambda}})$$

Eqn. II.4.2.1.b) Eigensignal of the discovered effect

Sorting back to reference orders  $I_0^g$  and  $J_0^s$  via  $E_k(I_{k,\hat{\lambda}}^g, J_{k,\hat{\lambda}}^s) \equiv \tilde{E}_k$  defines the final eigensignal  $E_k \in \mathbb{R}^{m \times n}$  that is compatible with signal matrices.

$$M_k \equiv M_{k-1} - E_k$$

Eqn. II.4.2.1.c) Remaining signal after dissection of the discovered effect  $k$

At this point, dissection is merely a matrix subtraction of the eigensignal. This results in the remaining signal  $M_k$ . It is the input for the next detection iteration  $k + 1$  that starts with the search strategy (cf. II.3), again.

### II.4.2.2 Effective clustering and conceptual limits of projection based methods

Implicitly, dissection *realizes an iterative soft clustering* of the signal. Even partially correlated effects can be properly dissected.

This is also demonstrated by the 3D example (Figure II.4.2.2.a). The blue effect has been successfully explained by its discovered and regressed eigensignal (see its gene curve in Figure II.4.1). All signal parts that remain from the blue effect are scattered around zero with distances that approximately correspond to the noise level.

In contrast, signals from all three other effects are still *untouched*. (The gene curve depicting the eigensignal for the green effect is already shown here as well. It will be dissected in the second iteration. All dissection steps for this 3D example are shown in II.6.1.)

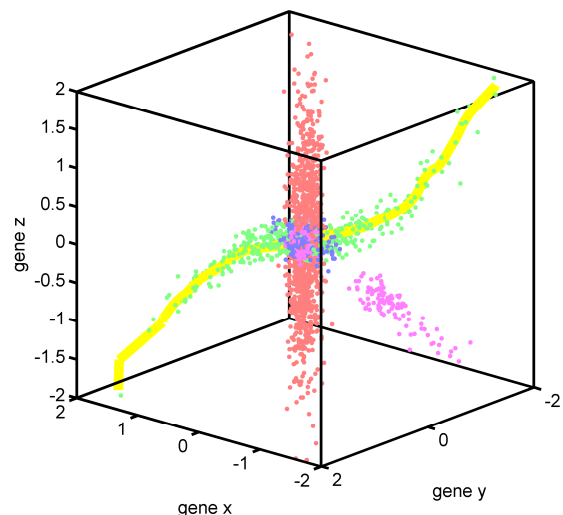


Figure II.4.2.2.a) 3D example, remaining signal  $M_1$  after dissection of the blue effect  $E_1$

Conceptually, this effective clustering by dissection goes beyond PCA and other methods that are equivalent to projections or to orthogonal coordinate transformations. For the 3D example, such methods could maximally provide three orthogonal gene axes, because after three projections only a point signal remains. However, there are four distinct effects in this 3D signal. Hence, these methods cannot dissect all four effects and, thus, original effects are hard to infer from resulting axes, e.g. from principal components<sup>(cf. II.6.1.2)</sup>.

To demonstrate these limits of projections and the importance of the effect focus in particular, the upper example uses only the regressed and smoothed  $\mathcal{S}(\mathcal{M}_{i,\hat{\beta}})$  for dissection<sup>(Figure II.4.2.2.b, upper panel)</sup>. This is compared with the dissection based on the final eigensignal  $\tilde{D} \cdot \mathcal{S}(\mathcal{M}_{i,\hat{\beta}})$  that cuts regression results with the effect focus<sup>(lower panel)</sup>.

Whenever partial correlations between effects are present in the signal, as in this example, then not only samples regulated by the current effect<sup>(blue)</sup> have a nonzero projection on the effect's gene axis<sup>(orange)</sup> but also samples from foreign effects. Hence,  $\mathcal{S}(\mathcal{M}_{i,\hat{\beta}})$  is not zero for samples regulated by foreign effects. Its subtraction moves points from these foreign effects towards a perpendicular plane, similar to a projection along the gene axis<sup>(see upper panel)</sup>.

This should be prevented in order to be able to detect and interpret *original* laws of gene regulation for the green, red and magenta effects, without any information loss or warping by other previously detected effects.

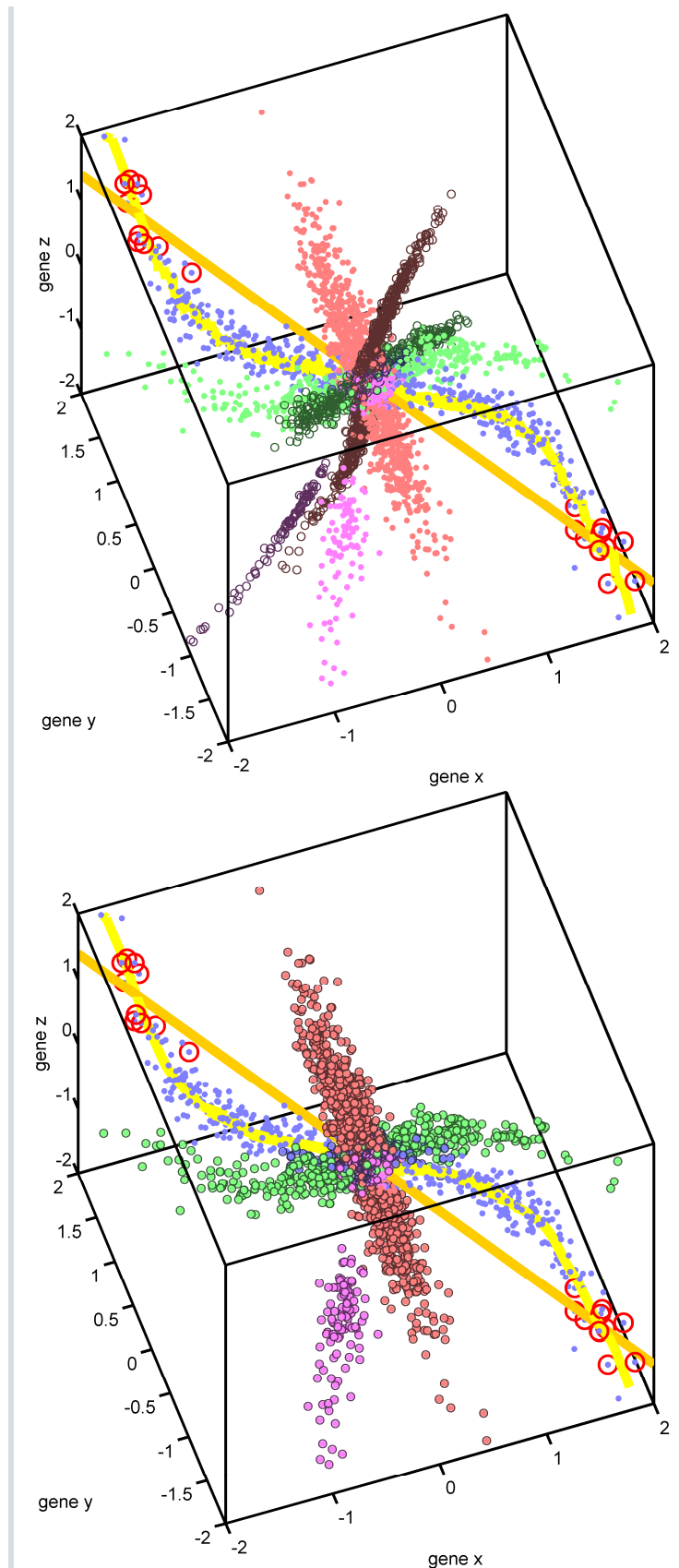


Figure II.4.2.2.b) Dissection with and without using the effect focus as dissection strengths

**Upper panel:** Naïve dissection based on the regressed effect curve. All points with nonzero projection on the effect axes are changes. **Lower panel:** Dissection in the effect focus based on the final eigensignal. Signals from partially correlated yet distinct effects are untouched, while the blue effect is removed.

In contrast, when using the final eigensignal  $E_k$  that makes use of the effect focus<sup>(cf. II.4.2.1)</sup>, signal parts from all other effects are untouched and only blue samples are moved by this dissection<sup>(Figure II.4.2.2.b, lower panel)</sup>, as intended.

### II.4.2.3 Visualization of high-dimensional eigensignals: the coordinate view

So far, most examples were visualized within their 3D gene space. This type of direct visualization is unfortunately not possible for high-dimensional signals. To still visualize the raw signal and the regressed effect curves for high-dimensional signals, coordinates of all gene and sample vectors and of the regressed effect curves can be depicted as heatmaps.

This is explained below for the second iteration of the 3D example that detects and dissects the green effect. The blue effect has already been dissected in the first iteration:

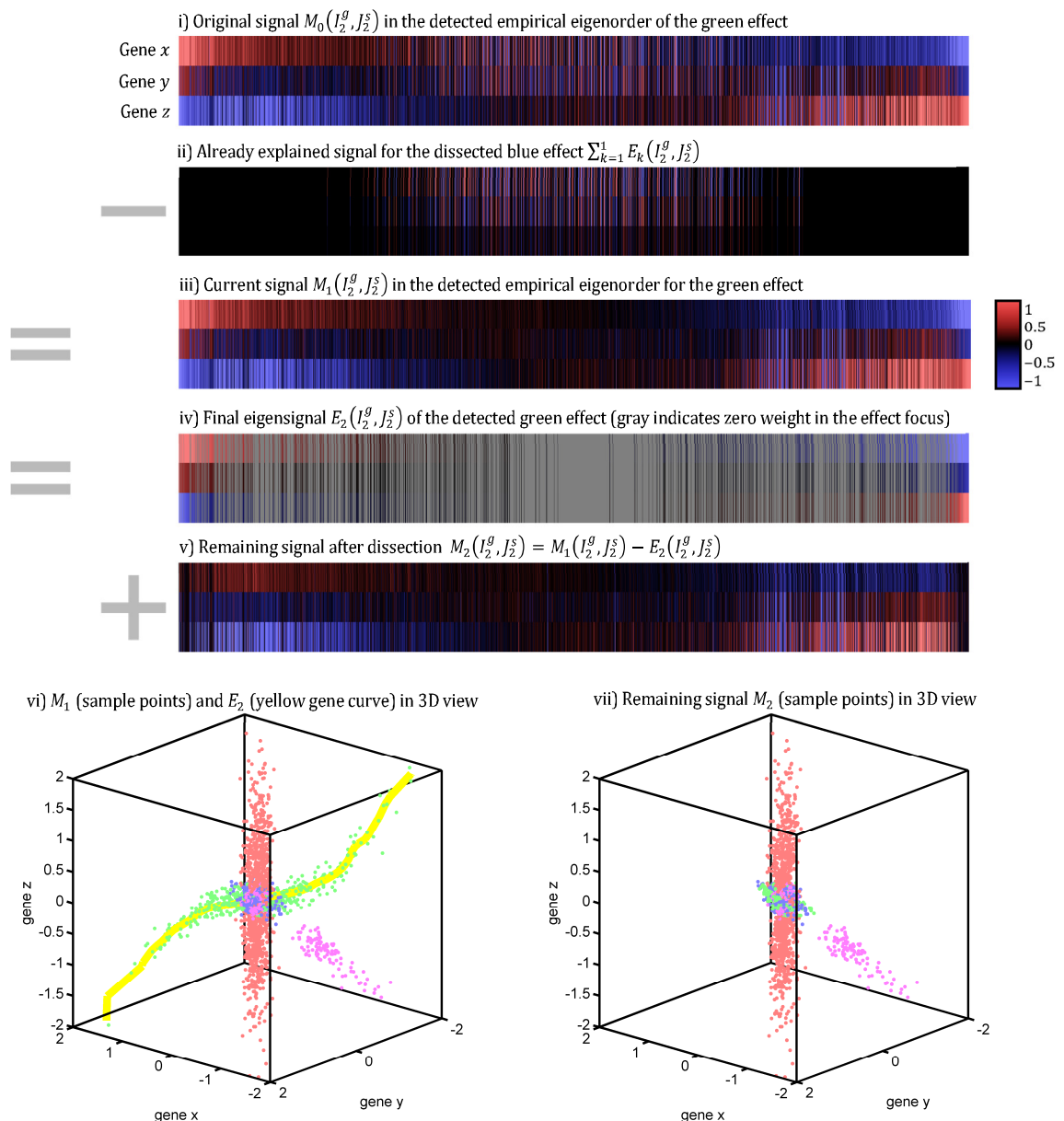


Figure II.4.2.3.a) Coordinate view for high-dimensional visualization; exemplary dissection of the green effect for the 3D example

This coordinate view is henceforth employed for visualization of all detections and dissections of high-dimensional effects. Higher dimensions simply correspond to more gene rows (or sample columns) in

heatmaps. A detailed explanation follows that clarifies the equivalence of this coordinate view to the gene and sample spaces:

- i) The original gene expressions  $M_0$  are depicted; they include signals from all four simulated effects. Columns correspond to samples and rows to genes.  $M_0$  is already presented in the empirical eigenorder  $(I_2^g, J_2^s)$  of the currently detected green effect.
- ii) The already explained parts of the signal are depicted in the empirical eigenorder for the detected green effect. Generally, this is the sum of all so far dissected effects. Here, it is just the final eigensignal  $E_1$  of the blue effect that was detected and dissected in the first iteration. As samples regulated by the blue effect have a signal near zero now, they are centered in the sample eigenorder  $J_2^s$  for the green effect.
- iii) The initial signal  $M_1$  for the current detection iteration is shown, again in the detected empirical eigenorder of the green effect. This ordered signal is already roughly bi-monotonic for top sample effect strengths (i.e. leftmost and rightmost columns). The green effect is partially correlated to the red effect with respect to the  $z$  direction. However, it is anti-correlated with respect to gene  $y$ . Only those samples that are co-regulated (i.e. that have the same color) for genes  $x$  and  $y$  belong to the green effect. Red samples may still have strong effect strengths (as quantified by their projections on the green effect axis), i.e. red and green samples are mixed by the empirical eigenorder.
- iv) This panel shows the eigensignal, i.e. the bimonotonically regressed signal of the green effect times dissection strengths. Points outside of the effect focus, as determined by relatively low correlations to the effect<sup>(cf. II.3.2.2 and Eqn. II.3.1.4.b)</sup>, are greyed out. Their signal is not modified by dissection. In particular, red samples are successfully filtered out, despite their partial correlation to the green gene axis.
- v) The final result of detection iteration  $k = 2$  is the remaining signal  $M_2$  after dissection of the detected green effect. Top samples of the green effect at the left and right end of the heatmap are now noticeably zeroed. The signal still contains the red and magenta effects; they are detected and dissected in subsequent iterations<sup>(cf. II.6.1)</sup>. (For instance, columns in the right half of this heatmap that show negative<sup>(blue)</sup> expression for all three genes correspond to samples regulated by the magenta effect.)
- vi) The initial signal  $M_1$  for the current iteration is shown in the 3D gene space. *Every sample point has coordinates depicted by one column in panel iii.* The eigensignal of the green effect is shown in form of the yellow gene curve of the effect. *Points along this curve correspond to columns in panel iv.*
- vii) Lastly, the final dissection result is shown in 3D view. *Every sample point corresponds to one column in panel v.*

Because every eigensignal is locally comparable to a signal average, dissection modifies the signal towards zero on average and ultimately only a zero signal can remain. Hence, termination is always guaranteed. Typically, already long before that, no effects with significant correlation and signal strengths remain in the signal. This is recognized in the search strategy when no gene or sample qualifies any longer<sup>(cf. II.3.1.8)</sup>. In this way, the method unsupervisedly determines the number of effects  $\hat{k}$  in the signal<sup>(cf. II.2.1.1)</sup>.

For the 3D concept example, the method correctly detects that after four iterations the remaining signal does no longer contain any relevant effects. Hence, the number of effects in the simulated 3D signal is determined to be  $\hat{k} = 4$  and the method terminates. More precisely, no candidate qualifies any longer with respect to the significance threshold for the signal strength<sup>(cf. II.3.1.8)</sup>. This is a *relative* statement and depends on the empirical noise estimation<sup>(cf. II.5.1)</sup>. This termination by remaining signal strength is visually clear in case of the 3D example when comparing the original simulated signal with the remaining signal after four dissection iterations:

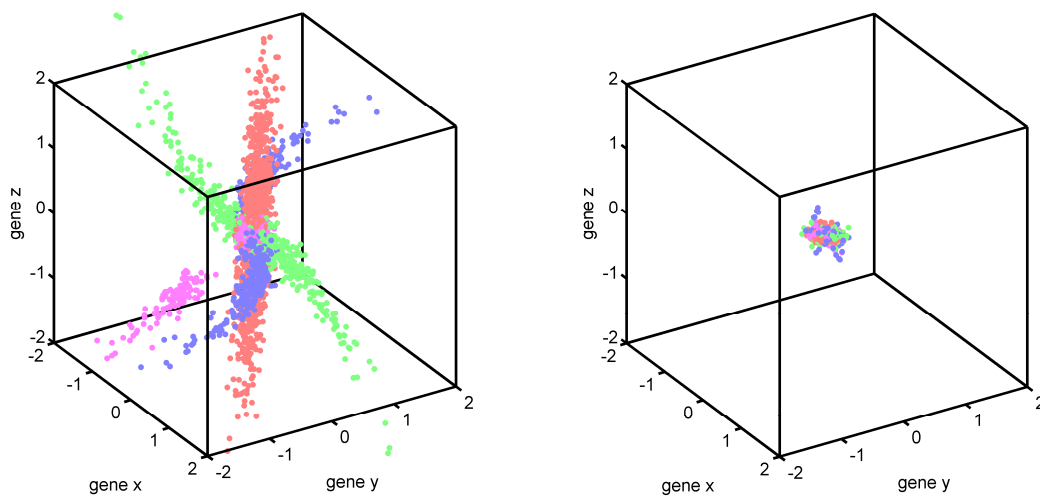


Figure II.4.2.3.a) 3D example, remaining signal after all four dissection iterations

*A vital part of a statistical solution for unsupervised detection of so far unknown effects is a statement of how reliable its findings are. This is typically quantified by  $p$  values, i.e. probabilities of observing detected effects (or even stronger effects) based on noise alone.*

*To obtain these probabilities, observables like correlations of genes with the effect's gene axis or correlations of samples with the effect's sample axis are tested for their significance. Likewise, the significance of observed signal strengths in the effect's focus is tested relative to the estimated noise level.*

*Based on these  $p$  values for observed signal strengths and correlations, effect candidates are qualified or disqualified by the search strategy<sup>(II.3.1.8)</sup>. As soon as no gene or sample qualifies any longer, the remaining signal is considered as noise and detection stops. Hence, noise estimation should be as precise as possible in order to minimize false negatives (due to overestimated noise) as well as false positives (due to underestimated noise).*

### II.5.1 Significance of Observed Signal Strengths

To calculate  $p$  values with respect to the signal strength of effects, first the true noise level of the overall signal needs to be estimated. This noise estimation is related to the field of blind denoising. It is a non-trivial sub problem, especially for signals containing many effects.

With the estimated noise distribution as reference, several statistical tests can be utilized to evaluate the observed signal in an effect's focus for significance. These tests differ in computation speed and how robust they are in practice against deviations from their theoretical assumptions by the actual signal.

#### II.5.1.1 Estimating the true noise level

A naïve approach might utilize global estimates like the empirical standard deviation of measured gene expressions  $\hat{\sigma}(M_0) \equiv (\sum_{i,j} (M_0(i,j) - \bar{M}_0)^2 / mn)^{1/2}$ . However, this can strongly *overestimate* the true noise level, because  $M_0$  still contains all true effects, including maybe strong and broad effects with a strong influence on such global estimators. In the 3D example, for instance,  $\hat{\sigma}(M_0) \approx 0.600$ , whereas the true simulated noise level is only 0.1. Hence, this reference would prevent the detection of weaker true effects (false negatives). On the other hand, using the standard deviation of the remaining signal  $\hat{\sigma}(M_{k-1})$  as reference at the beginning of each iteration  $k$  would ultimately strongly *underestimate* the true noise level in the initial signal, as dissecting effects propagates the remaining signal  $M_{k-1}$  ever nearer towards zero. Hence, using this estimate may lead to detection of many false positives.

Estimating the true noise level in a signal of unknown structure is a difficult problem that is also studied in the field of blind denoising. If the signal represents an image, one idea<sup>[86]</sup> is to first find patches of *weak texture* in this image and only utilize these patches as basis for noise estimation. However, this approach is based on the prescribed spatial order of image rows and columns. Another idea from blind denoising is to model projections of theoretical noise on all conceivable effect axes in a theoretical effect alphabet. These projections should be distributed normally and the supremum of these projections over the alphabet provides a useful cut

line between true effects and noise. This can then be utilized as stop criterion for denoising as demonstrated by the BIRD algorithm<sup>[87]</sup>. However, this approach needs a known and fixed alphabet of effects.

In order to adapt these two ideas for the unsupervised scenario with an initially empty effect alphabet and initially unknown and unfixed order of genes and samples, one needs to identify “weak textures” in this context. Let  $(|w_{k,i}^g|, |w_{k,i}^s|)$  denote the final effect focus for converged effect axes<sup>(cf. II.3.2)</sup> of a detected effect  $k$ . Gene expressions that are significantly correlated<sup>(cf. II.5.2)</sup> to this effect’s axes are not noise. They are “textured” in the image analogy and thus should be excluded for noise estimation. To deselect them, I demand that dissection strengths<sup>(cf. II.4.2.1)</sup>  $D_k$  for the current effect are near zero via  $D_k < 10^{-3}$  (the relation is taken component-wise, resulting in a Boolean mask  $\in \{\text{true}, \text{false}\}^{m \times n}$ ).

This condition alone is not yet sufficient to exclusively select noise, because there might exist many other effects with strong signals that are perpendicular (or partially correlated) to currently discovered effect axes. To deselect them as well, I utilize the standard assumption of a constant global noise level (rather than noise that varies for each effect). This allows testing for perpendicular effects based on standard  $z$  scores for the radial distances of genes and samples to the respective current effect axis. A cutoff condition of less than three noise standard deviations  $\hat{\sigma}_{\text{uc}}(\mathcal{N}_k)$ <sup>(defined below)</sup> has worked well in practice. From iteration  $k \geq 2$  onwards, the noise standard deviations can be estimated reliably based on the noise distribution  $\mathcal{N}_k$ . However, initially when  $k = 1$ , no noise distribution has been estimated yet. In this case, and again with the assumption of a global noise level, the standard deviation (weighted with dissection strengths  $D_1$ ) of pixels in  $M_0$  around the regressed eigensignal  $E_1$  can approximate the initial noise standard deviation.

Together, both conditions select only those gene expressions in  $M_{k-1}$  that are most probably “untextured” noise. These values are added to the initially empty estimated noise distribution  $\mathcal{N}_k$ . Every selected value for gene  $i$  and sample  $j$  in the noise distribution is assigned  $1 - D_k(i, j)$  as weight. The same is done for every detected effect, making the estimated noise distribution  $\mathcal{N}_k$  larger and more reliable with every dissection. For memory performance,  $\mathcal{N}_k$  and its weights matrix  $W_{\mathcal{N}_k}$  are implemented as matrices of the same size as  $M_k$ . Pixels never selected as noise samples get weight zero. Pixels selected as noise estimates in multiple detection iterations are averaged in  $\mathcal{N}_k$  and the maximum of their corresponding weights is retained in  $W_{\mathcal{N}_k}$ .

Now the true noise standard deviation of measured (gene, sample) pixel around zero can be estimated by  $\hat{\sigma}_{\text{uc}}(\mathcal{N}_k) \equiv (\sum_{i,j} W_{\mathcal{N}_k}(i, j) \cdot (\mathcal{N}_k(i, j) - 0)^2 / \sum_{i,j} W_{\mathcal{N}_k}(i, j))^{1/2}$ . I utilize the **un**centered standard deviation that is always computed relative to zero (and not relative to the empirical mean), as zero already represents the theoretically known mean for globally pre-centered data.

For the blue, green, red and magenta effects in the 3D concept example, this procedure yields estimates for the noise level of  $\hat{\sigma}_{\text{uc}}(\mathcal{N}_1) = 0.100$ ,  $\hat{\sigma}_{\text{uc}}(\mathcal{N}_2) = 0.108$ ,  $\hat{\sigma}_{\text{uc}}(\mathcal{N}_3) = 0.103$  and  $\hat{\sigma}_{\text{uc}}(\mathcal{N}_4) = 0.097$  respectively<sup>(cf. Figure II.5.1.2, page 79)</sup>, while the true simulated noise level is  $\sigma(\mathcal{N}_{\text{sim}}) = 0.1$ .

Besides providing a reference for significance estimation, the noise distribution can also be utilized to define signal-adaptive error thresholds. For example,  $\epsilon \equiv \hat{\sigma}(\mathcal{N}_k)/1000$  is used for bi-monotonic regression; there is no point in regressing the signal more precise than this. (For  $k = 1$ , error thresholds are estimated using  $\hat{\sigma}_{\text{uc}}(M_0)$ ; a probable overestimation is uncritical for this purpose.)

To answer whether an effect has significant signal strength, it is now possible to test the signal in the final effect focus against the obtained sampling  $\mathcal{N}_k$  for the true noise distribution. As the signal is signed, signal averages are not useful for this comparison. Instead the signal power (i.e. quadratic values) or signal amplitudes (i.e. absolute values) can be utilized. Several tests and statistics can be employed to this end, with different advantages and disadvantages.

■ *Permutation tests are non-deterministic and have too high computational cost*

A straight-forward approach would be a *permutation test*. Some distance statistic could be defined between signal amplitudes in the effect focus and noise amplitudes. Then the effect focus, i.e. gene weights and samples weights, could be randomly permuted among all available genes and samples. If the observed statistic for the effect focus is stronger than for sufficiently many permutations, the effect's signal strength may be called significant. This approach has two problems. First, it is non-deterministic because of random permutations. Secondly, it is far too slow in practice, because the significance with respect to signal strength has to be tested for every effect candidate during the search strategy<sup>(cf. II.3.1.8)</sup>.

■ *Kolmogorov-Smirnov tests are deterministic, but still too slow and not sensitive enough*

As a sample of  $\mathcal{N}_k$  was obtained (instead of just a scalar estimation of  $\sigma_{uc}(\mathcal{N}_k)$ ), one deterministic possibility to obtain a  $p$  value is to employ *Kolmogorov-Smirnov (KS) tests*. They can directly compare distributions of signal amplitudes. To this end, first the cumulative distribution function (CDF) for the current absolute signal matrix  $|M_{k-1}| \equiv (|M_{k-1}(i, j)|)_{i, j}$  is empirically estimated, using dissection strengths  $D_k$  as weights. Secondly, the weighted CDF for absolute signals of all noise samples  $|\mathcal{N}_k|$  is empirically estimated, using  $W_{\mathcal{N}_k}$  as weights. For a one-tailed test, whether signal amplitudes in the effect focus are higher than noise amplitudes, the KS statistic is simply the maximum of all point-wise differences of these two CDFs. Notably, the effect focus<sup>(Eqn. II.3.1.4.b)</sup> only utilizes correlations and their significance, but not signal strengths to define weights; otherwise weighting with  $D_k$  would be biased towards stronger expressions, which could lead to underestimated  $p$  values. Effectively, the KS test looks for larger counts (or larger weight mass) of higher expressions in the effect focus as can be expected by the sampled noise. Only the relative order of values is important for this test, rather than their absolute numeric values. Hence, this test is robust against outliers. To finally compute corresponding  $p$  values, the asymptotic upper tail of the Kolmogorov-Smirnov distribution<sup>(D statistic in table 1 of [88]; also see kstest2.m of the MATLAB® Statistics Toolbox for implementation details)</sup> can be integrated.

Though this KS test is robust and deterministic and utilizes all available information, it has a disadvantage: For strong yet small effects that differ only in the upper distribution tail when compared to  $\mathcal{N}_k$ , it might cause false negatives. This insensitivity with respect to tails is a known weakness of Kolmogorov-Smirnov tests; a solution using inverse signal variances as weights to focus on tails has been suggested<sup>[89]</sup>, but it is difficult to estimate the signal variance precisely. Additionally, for the search strategy weighted CDFs of the signal in every candidate effect focus would have to be estimated for these Kolmogorov-Smirnov tests. While this is practically possible, a faster alternative is presented next.

■ *Weighted t-tests are fast, deterministic and have other practical properties*

A relatively simple weighted  $t$ -test has been found to provide all needed properties for practice. It can be computed fast and deterministically. Additionally, it is robust and forgiving with respect to slight over- or underestimations of the true noise level. Most importantly, together with a significance threshold that can be chosen tightly as  $\alpha_{\text{signal}} \equiv 10^{-5}$ , these  $t$ -tests can correctly sort out false positives with respect to signal strength in most test cases, while still qualifying simulated weak true positives<sup>(cf. II.6)</sup>. This was confirmed over many simulations<sup>(e.g. II.6.2.5)</sup>.



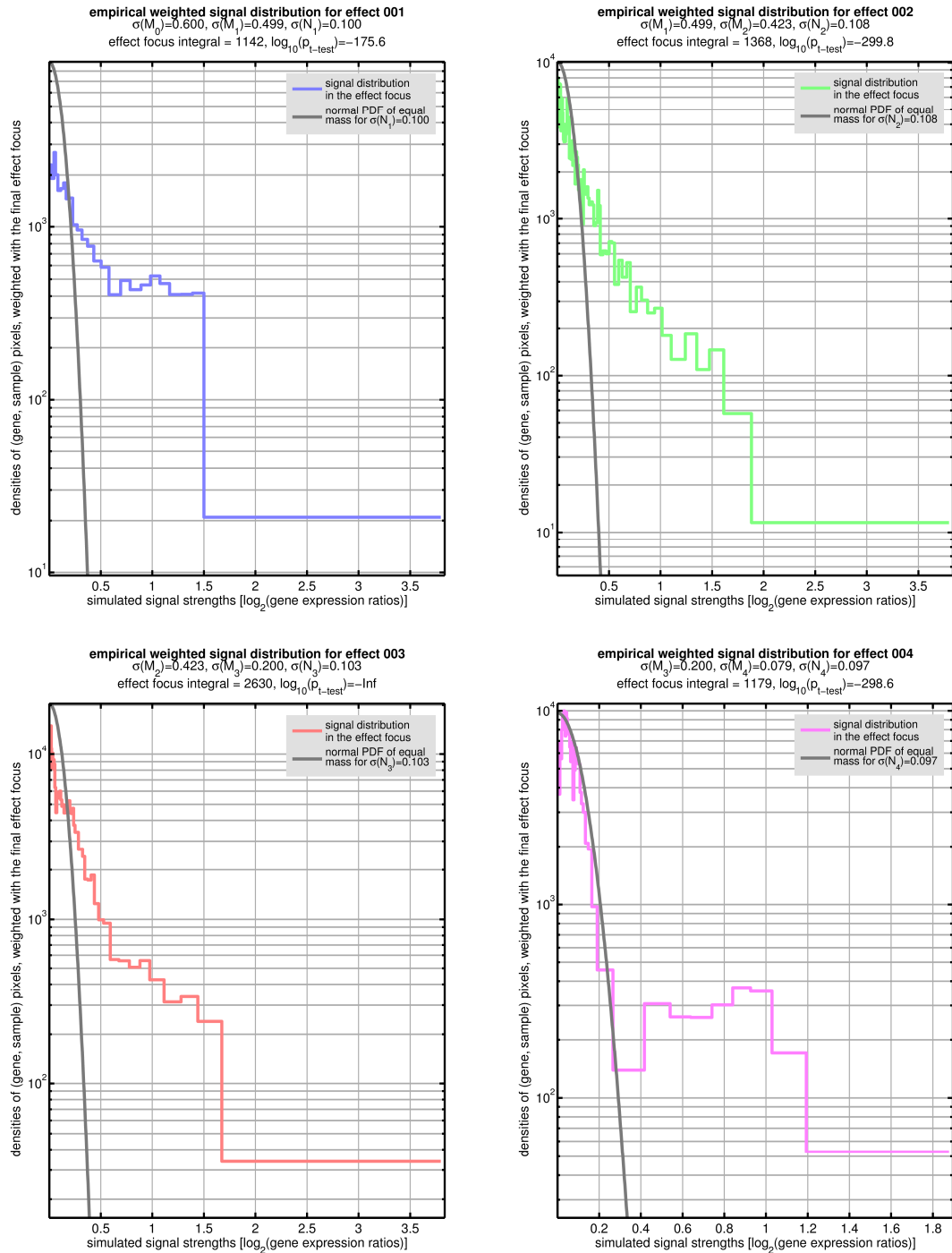


Figure II.5.1.2) Signal significance in the 3D example

Each panel shows the empirical density over signal strengths of (gene, sample) pixels in the focus of the discovered effect. The area under each colored curve corresponds to the sum of the product effect focus, i.e. the sum of the detected effect's dissection strengths  $\sum_i \sum_j D_k(i, j)$ . Theoretical half Gaussian distributions are depicted with the same mass and for the respective noise standard deviation  $\hat{\sigma}_{uc}(\mathcal{N}_k)$  ( $\sigma$  abbreviates  $\hat{\sigma}_{uc}$  in these plots). Standard deviations of  $M_k$  initially strongly overestimate the simulated noise level  $\sigma(\mathcal{N}_{sim}) = 0.1$  and in the end underestimate it ( $\hat{\sigma}_{uc}(M_4) = 0.079$ ). Approximate  $p$  values based on  $t$ -tests are provided in panel titles in  $\log_{10}$  scale. All are very small, as the signal of all four effects is clearly stronger than the simulated noise level. Additionally, all effects are carried by sufficiently many samples ( $-\text{Inf}$  corresponds to a numeric underflow to  $p = 0$ ).

In contrast, for instance,  $f$ -tests were relatively sensitive to the correct estimation of the true noise level. (These tests were also examined during development to compare signal powers between the effect's signal and the noise distribution.)  $f$ -tests caused both false positives and false negatives in several simulations. This corresponds to the known sensitivity of  $f$ -tests against non-normality<sup>[90]</sup>, while  $t$ -tests are known to be robust against moderate deviations from normality<sup>[91,92]</sup>.

The central limit theorem<sup>[85 §16.2.5.2]</sup> states that the arithmetic mean of many independent random variates from the same distribution (with finite mean and variance) is distributed approximately normal. This and the above-mentioned robustness of  $t$ -tests allow their application to only half-normally distributed values. More precisely, I utilize them to compare many half-normally distributed signal amplitudes in  $|M_{k-1}|$  in the effect focus against many half-normally distributed noise amplitudes in  $|N_k|$  (the absolute value is taken *component-wise*, again).

For computation,  $t$ -tests require only empirical means, empirical standard deviations and weight sums for both distributions to compare. In the unweighted case, weight sums simply equal the number of points. All these measures can be computed fast. For  $|N_k|$ , this even needs to be computed only once per detection iteration and can then be reused for all effect candidates visited by the search strategy. Let  $\hat{\mu}(|N_k|, W_{N_k})$ ,  $\hat{\sigma}(|N_k|, W_{N_k})$  and  $v_{N_k} \equiv \sum_{i,j} W_{N_k}$  denote the empirical weighted mean, the empirical weighted standard deviation and the weights sum for  $|N_k|$ . Identical to dissection strengths  $D_k$  (Eqn. II.4.2.1.a), let  $W_{\text{focus}}$  denote the product effect focus of some effect candidate. Then  $\hat{\mu}(|M_{k-1}|, W_{\text{focus}})$  denotes the weighted mean of signal amplitudes in this focus,  $\hat{\sigma}(|M_{k-1}|, W_{\text{focus}})$  is the weighted standard deviation of it and  $v_{W_{\text{focus}}} \equiv \sum_{i,j} W_{\text{focus}}$  is its weights sum. As all weights are  $\in[0,1]$ , each pixel can contribute at maximum 1 to this weights sum, i.e. no single independent measurement is allowed to count more than once. With the assumption of a constant global noise level, variances of both distributions can also be assumed equal. The  $t$  statistic for this comparison is now defined analogous to the unweighted two-sample  $t$ -test for independent samples with equal variance<sup>(cf. [79], eqn. 8.11)</sup> by replacing sample counts with weight sums:

$$\begin{aligned} v &\equiv v_{N_k} + v_{W_{\text{focus}}} - 2 \\ s &\equiv \sqrt{\left( (v_{N_k} - 1) \cdot \hat{\sigma}(|N_k|, W_{N_k})^2 + (v_{W_{\text{focus}}} - 1) \cdot \hat{\sigma}(|M_{k-1}|, W_{\text{focus}})^2 \right) / v} \\ t &\equiv \frac{\hat{\mu}(|M_{k-1}|, W_{\text{focus}}) - \hat{\mu}(|N_k|, W_{N_k})}{s \sqrt{\frac{1}{v_{N_k}} + \frac{1}{v_{W_{\text{focus}}}}} \end{aligned}$$

Eqn. II.5.1.2.a)  $t$  statistic for the difference between the mean signal amplitude in the focus of an effect and the mean noise signal amplitude

For signals stronger than noise, this  $t$  statistic for the effect is positive. To compute the  $p$  value for this statistic, the upper tail of the  $t$  distribution with  $v$  degrees of freedom is integrated (as implemented by the `tcdf.m` function of the MATLAB® Statistics Toolbox). This  $p$  value for an effect candidate or detected effect finally quantifies the significance of the observed signal strengths in the respective effect's focus relative to the estimated noise level.

Weighted correlations<sup>(cf. II.2.3.1)</sup> are utilized as base measure to detect interactions. For effect candidate qualification<sup>(cf. II.3.1.8)</sup> and to determine the effect focus<sup>(cf. II.3.2.2 and Eqn. II.3.1.4.b)</sup>, significance estimates for these correlations are essential.

#### ■ A *t*-statistic based approximation

In case of an unweighted Pearson correlation  $r$ , a corresponding  $t$  statistic can be derived from a two-dimensional normal distribution<sup>(see [93], section 12-8 for details)</sup>. It is computed as  $t = r\sqrt{\nu}/\sqrt{1-r^2}$ <sup>(see [93] eqn. 12-93 or [79] eqn. 11.20)</sup>, where  $\nu$  is the number of degrees of freedom that equals the number of correlated points minus two. For interpretation,  $p$  values should quantify the probability to observe *stronger correlations of the same sign* due to chance. The corresponding one-tailed  $p$  value can be computed by integrating the  $t$  distribution with  $\nu$  degrees of freedom in  $[t, \infty]$  (if  $t \geq 0$ ) respectively in  $[-\infty, t]$  (if  $t \leq 0$ ).

In the following, let  $|g\rangle$  denote an observed gene vector from the contextual signal matrix, let  $|a^s\rangle$  be a sample effect axis to which it is correlated, let  $|w^s\rangle$  denote sample weights of the contextual effect focus and finally let  $[g|a^s]_{|w^s\rangle}$  be the weighted correlation<sup>(cf. II.2.3.1)</sup> for this gene with the effect axis. A  $p$  value for this correlation needs to be computed. (Obtaining  $p$  values for correlations of samples with effect gene axes is analogous.)

By setting  $\nu = \sum_j \langle e_j^s | w^s \rangle - 2$  and  $r = [g|a^s]_{|w^s\rangle}$  in the above formula for the  $t$  statistic, an approximation of  $p$  values for weighted correlations can be obtained by assuming that the resulting statistic is distributed like a  $t$  statistic with  $\nu$  degrees of freedom. Whether this approximation is well-defined, is tested next.

#### ■ Comparison with null distribution based $p$ values

Similar to significance computation for signal strengths, random sampling techniques for the null distribution like permutation tests have the disadvantage to make the algorithm non-deterministic and relatively slow. Therefore, they cannot be utilized for practical purposes. However, they have the advantage to determine correct  $p$  values with ever higher precision over sample size of the null distribution, without needing to know its analytical form.

To test the applicability of the above  $t$  statistic based approximation, I sampled the null distribution of weighted correlations and then compared  $p$  values obtained by both methods.

One possibility to obtain a sampling of the null distribution is to randomly permute sample columns before computing gene correlations. Resulting weighted correlations represent correlations that could be explainable due to noise alone. A more direct approach is to simulate theoretically pure normal noise. (As correlations are scale invariant, the standard deviation of this noise signal can be chosen freely.) Then the same weighted correlations<sup>(cf. II.2.3.1)</sup> are computed with these simulated noise genes like before with the actually observed gene  $|g\rangle$ , i.e. using the same effect axis  $|a^s\rangle$  and the same weights  $|w^s\rangle$ . The resulting sampling of the null distribution is utilized next to approximate the  $p$  value directly, as in permutation tests. More precisely, the  $p$  value for  $[g|a^s]_{|w^s\rangle}$  is the ratio of stronger correlations in the null distribution (i.e. correlations that are nearer to +1 respectively to -1, depending on the correlation sign). As correlations are signed, the minimal achievable  $p$  value by this sampling method is  $2/(\text{sampling resolution})$ .

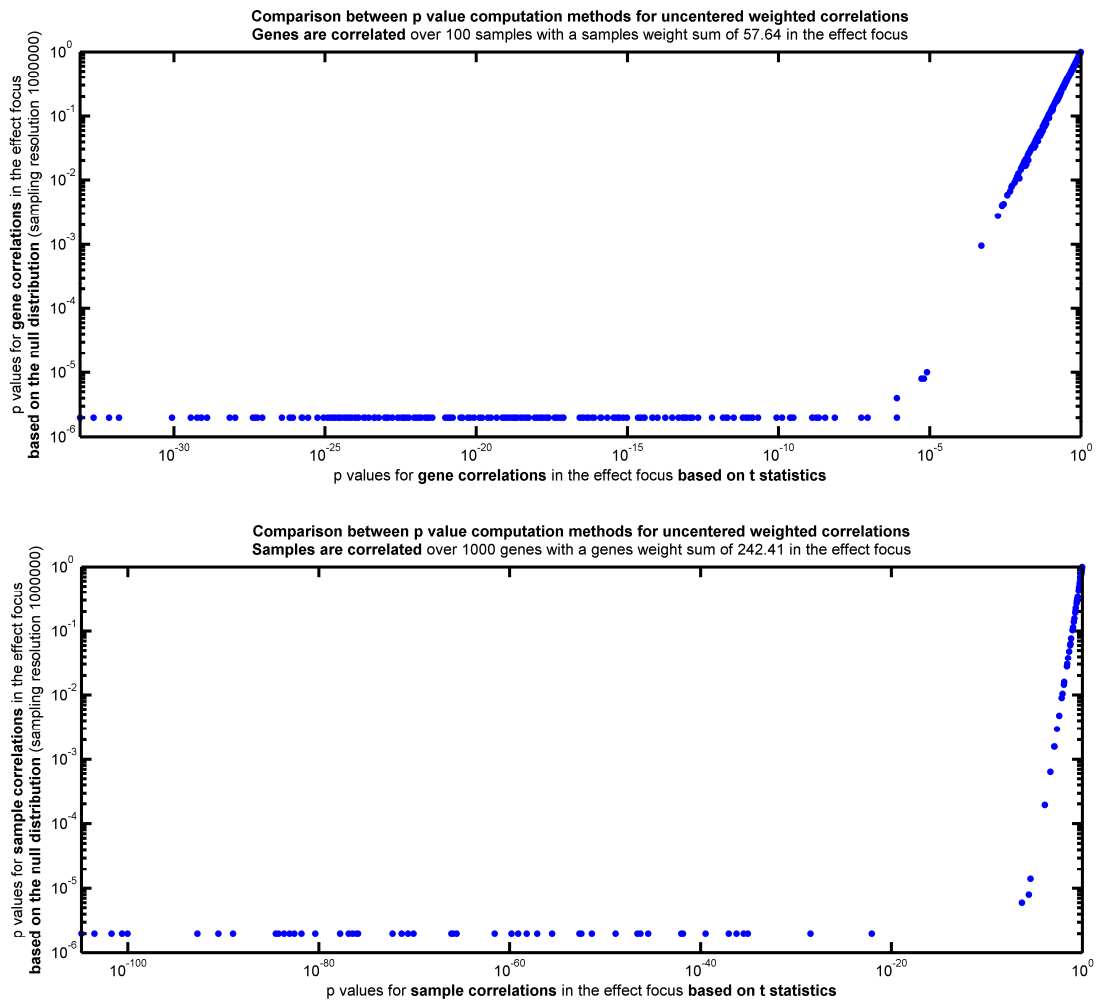


Figure II.5.2.1) Comparison of  $p$  value computation methods for weighted correlations

For this comparison, genes and samples simulated for a versatility test<sup>(II.6.2.1)</sup> are correlated with effect axes discovered for effect pattern #3. This pattern regulates 200/1000 simulated genes and 50/100 samples.

Non-deterministic  $p$  values based on a sampled null distribution (vertical axis, with sampling resolution 1.000.000) are compared to deterministic  $p$  values based on approximate  $t$  statistics as for weighted correlations. The former are truncated at  $2/\text{sampling resolution}$  (correlations are signed and  $p$  values are estimated separately for both tails; hence the factor two).

Above this resolution threshold, both methods agree nearly linearly, which provides confidence that the  $t$  statistic formula from unweighted correlations is still applicable to weighted correlations.

A detection iteration of effect pattern #3 from a versatility test<sup>(II.6.2.1)</sup> has been selected for this comparison (Figure II.5.2.1). For all 1000 genes<sup>(upper panel)</sup> and for all 100 samples<sup>(lower panel)</sup>,  $p$  values for their correlations to the respective effect axis in the effect focus have been computed by both methods. On the vertical axis,  $p$  values based on a null distribution with sampling resolution 1.000.000 are depicted. On the horizontal axis,  $p$  values are obtained via the approximate  $t$  statistic<sup>(see above)</sup>. While  $p$  values based on the null distribution are truncated at  $2/\text{sampling resolution}$ , a *nearly linear agreement* above that resolution threshold provides confidence that the  $t$  statistic based approximation formula from unweighted correlations is still valid for weighted correlations. (No closed form of the precise analytical PDF for correlations using arbitrary weights seems to be known yet. A stochastic derivation of  $p$  values for weighted correlations could help to solidify this significance measure. Maybe it could be derived similar to the unweighted case<sup>(see [93], pages 188-199)</sup> under the null hypothesis of uncorrelated signals by introducing weights as another random variable that is constrained by a constant weights sum.)

### II.5.2.2 Significance of all gene and sample correlations for an effect

---

In the last section,  $t$  statistics and  $p$  values have been obtained for weighted correlations of each single gene  $r_i^g \equiv [g_i | a^s]_{|w^s}$  to an effect's sample axis  $|a^s\rangle$  in its sample focus  $|w^s\rangle$ . Likewise, correlations  $r_j^s \equiv [s_j | a^g]_{|w^g}$  and their significance for each single sample are computed. For effect qualification<sup>(cf. II.3.1.8)</sup> these results need to be summarized in form of a scalar  $p$  value for all correlations.

As confirmed by the sampled null distribution<sup>(cf. II.5.2.1)</sup>, all corresponding  $t_i^g$  gene statistics approximately follow a  $t$  distribution with  $\sum_j \langle e_j^s | w^s \rangle - 2$  degrees of freedom and all  $t_j^s$  sample statistics follow a  $t$  distribution with  $\sum_i \langle e_i^g | w^g \rangle - 2$  degrees of freedom.

If all observed  $\{|t_i^g| | i = 1 \dots m\}$  for genes or all observed  $\{|t_j^s| | j = 1 \dots n\}$  for samples are significantly higher than absolute values expected by the respective  $t$  distribution, then the effect is carried by significant correlations. This can be tested by two Kolmogorov-Smirnov tests in the gene respectively sample effect focus<sup>(cf. II.5.1.2)</sup>, resulting in  $p_{r,genes}$  for all gene correlations (that is for their  $|t_i^g|$  statistics) and in  $p_{r,samples}$  for all sample correlations.

Finally,  $p_r \equiv \min(p_{r,genes}, p_{r,samples})$  provides a scalar  $p$  value for all correlations with the current effect axes  $(|a^g\rangle, |a^s\rangle)$  and is tested for significance during effect qualification<sup>(cf. II.3.1.8)</sup>.

## II.6 Method Validation and Comparison

To validate that the method is capable of dissecting signals into meaningful effects, several test scenarios have been designed.

The main scenario tests the versatility of the method by simulating 13 different effects for randomly selected genes and patient. These effects comprise seven distinct patterns that imitate different real world gene expression effects of biological nature or lab effects<sup>(II.6.2)</sup>. The second scenario tests the superposition limit of still dissectible overlapping effects for three different patterns<sup>(II.6.3)</sup>.

Finally, several detection limits are tested<sup>(II.6.4)</sup>: the minimal signal strength relative to the noise level that is required for detection, the maximally tolerable ratio of missing values, the exclusion of false positives in the few samples limit and the exclusion of false positives in the many noise genes limit.

### II.6.1 3D Concept Example

Before starting with high-dimensional effects, an overview of all results for the concept example is provided here. Principal components for the same signal are also provided for a comparison and to illustrate why they are hard to interpret in terms of original simulated effects.

#### II.6.1.1 Overview of all dissection steps

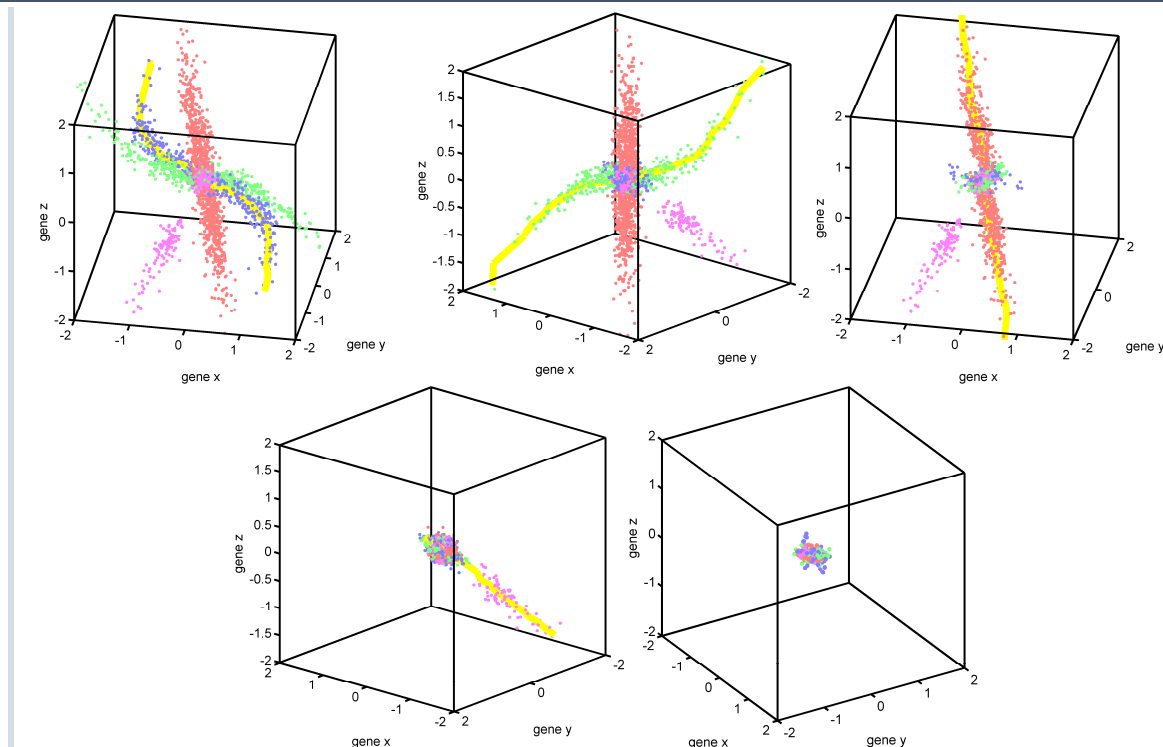


Figure II.6.1.1) 3D concept example, all detected gene curves, all dissection steps and the remaining signal

Four simulated pathways corresponding to four distinct sample groups that are governed by different regulation laws for the same three genes have been simulated (cf. II.1.1). First the blue effect is detected, regressed and dissected. This is followed by the green, the red and the magenta effect. The remaining signal has a strength below the estimated noise level and hence signal dissection terminates after four iterations. (Details for all steps are explained in II.3, II.4 and II.5.)

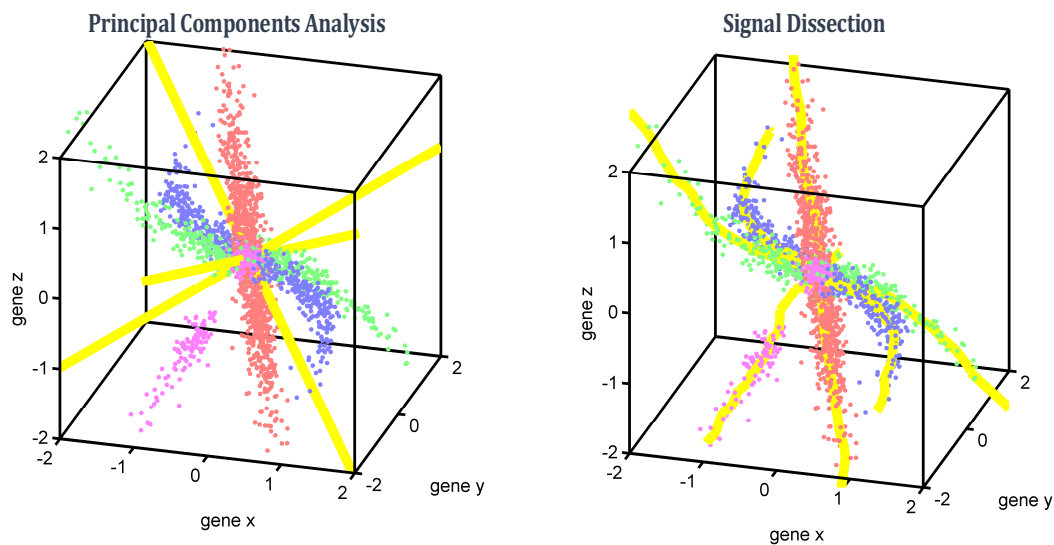


Figure II.6.1.2) 3D concept example, comparison of final detected gene curves and principal components

Clearly, depicted principal components are not strongly correlated to any of the four simulated effect axes. Hence, no principal component represents a simulated law of gene co-regulation. Instead, all four effects have non-zero projections on all three principal components. Hence, every principal component (i.e. the new coordinate defined by it) represents a mixture (i.e. a linear combination) of four distinct effects. This is the reason, why these principal components are hard to interpret and not helpful when the goal is to discover new and distinct effects in an unknown signal.

In contrast, signal dissection detects and regresses all four simulated (monotonically nonlinear) effects, i.e. empirically derives the original laws of gene co-regulation. Every effect is dissected iteratively<sup>(cf. Figure II.6.1.1)</sup> and its associated effect focus defines the samples in it, thereby effectively realizing a clustering of the signal, on top of explaining laws of gene regulation.

The interpretability of a simulated effect based on detection results can be quantified by the best correlation of the simulated effect axis to any of the detected effect axes (respectively principal components). This can be used to illustrate the interpretability of method results for high-dimensional multi-effect signals, as is defined in II.6.2.3.

As it is used often for unsupervised effect discovery in biosciences<sup>(cf. I.2.2.1)</sup>, results of hierarchical clustering for the 3D concept example are illustrated next. Hierarchical clustering can be configured with various distance metrics and linkage methods. First, results for the default Euclidean distance and for average linkage are demonstrated. In this case<sup>(cf. Figure II.6.1.3.a)</sup>, most of the samples are grouped by a zero-centered large cluster, while tips of simulated effects are separated by several smaller clusters.

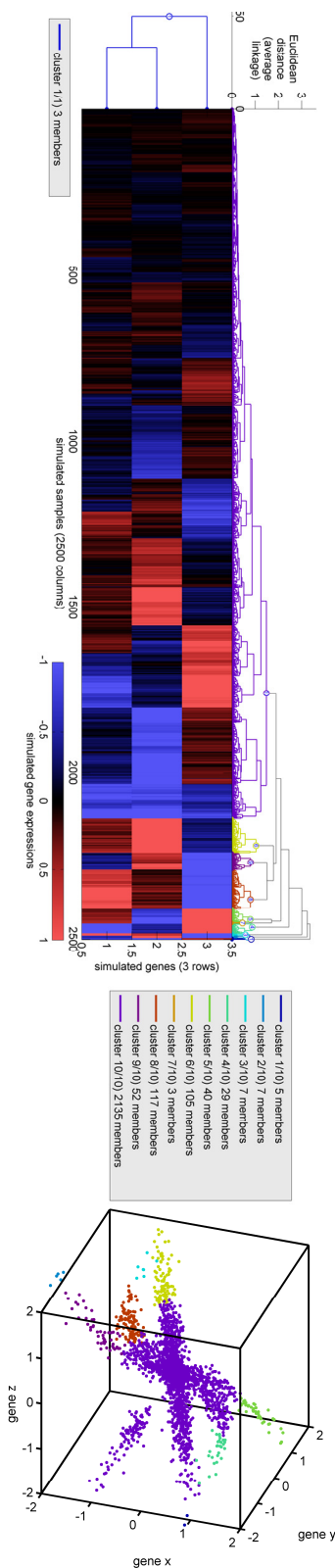


Figure II.6.1.3.a) 3D concept example, results from hierarchical clustering (Euclidean distance metric, average linkage)

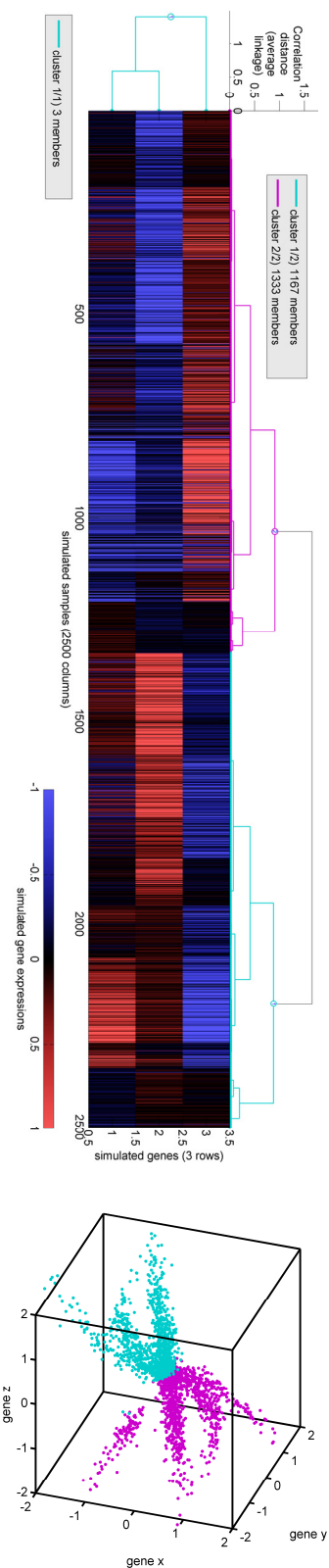


Figure II.6.1.3.b) 3D concept example, results from hierarchical clustering (Correlation distance metric, average linkage)

One conceptual problem of hierarchical clustering for this signal is its distance-based model for interactions: Points from different simulated effects are near to each other at the zero crossing, but points from opposite tips of the identical effect are geometrically far from each other.

The correlation based distance (instead of the Euclidean distance) is a step towards solving this problem and results in a more balanced dendrogram, but it still cannot make out true simulated effects (cf. Figure II.6.1.3.b).

Other linkage methods and distance metrics have been tested (data not shown); they lead to similar uninterpretable results for this signal.

Another comparison with hierarchical clustering for high-dimensional data follows in II.6.2.6; all other comparisons are performed with PCA, as it seems to be the more powerful competitor compared to hierarchical clustering.



The purpose of the versatility test is to challenge the method a wide variety of different effect forms that imitate different real world biological effects or lab effects. These effects may overlap each other. After defining simulated effect patterns, an exemplary run demonstrates how the method works for these high-dimensional signals. For method comparison, detection statistics are computed for signal dissection and for PCA. In addition, a brief comparison with hierarchical clustering is provided.

## II.6.2.1

## Scenario definition and 7 distinct effect patterns

A signal with  $m = 1000$  genes and  $n = 100$  samples is simulated by superposition of seven effects and normal noise with standard deviation  $\sigma = 0.5$ .

The seven simulated effects differ in size, form and signal strength. For every effect, first genes and patients are randomly permuted. For predefined effect sizes (percentages), a bi-monotonic signal has been generated at the top of these permutation orders, as depicted (cf. Figure II.6.2.1) (for other genes and samples, the simulated effect has zero eigensignal).

To test the dissection of outshining broad and strong effects, pattern #1 mimics different experimental setups (e.g. the labeling protocol effect in GSE10846<sup>[5]</sup>). Due to its size, it is relatively easy to detect, but important to dissect without information loss. Pattern #2 simulates a strong binary cluster for only 10 measured genes; this is typical for, e.g., gender specific gene expressions. To test the detection of subclass-only effects, pattern #3 simulates a broad biological effect that only exists in half of the samples. Pattern #4 simulates a medium sized one-sided effect that is also typical for real-world biological effects. Pattern #5 simulates a small gradual signal of medium strength. It is mainly a one-sided effect, but includes one anti-correlated gene to test detection specificity. To test detection sensitivity with respect to noise, pattern #6 adds a weak effect that has a maximum absolute signal at the noise level and that gradually declines to zero. Finally, pattern #7 tests the sensitivity with respect to narrow effects by only affecting 5% of all simulated samples.

The superposition of all seven effects in a common gene and patient reference order plus normal noise with standard deviation  $\sigma = 0.5$  serves as the input signal  $M_0$  for detection methods. The detection task is to recover simulated effect axes and depicted eigensignals for the respective random subsets of genes and samples.

Besides this test scenario with 7 effects, and in order to increase detection and dissection difficulty, another larger versatility scenario is defined with 13 effects. To this end, the biologically most typical patterns #2, #3 and #4 are simulated three times each (for different random permutations of all genes and all samples).

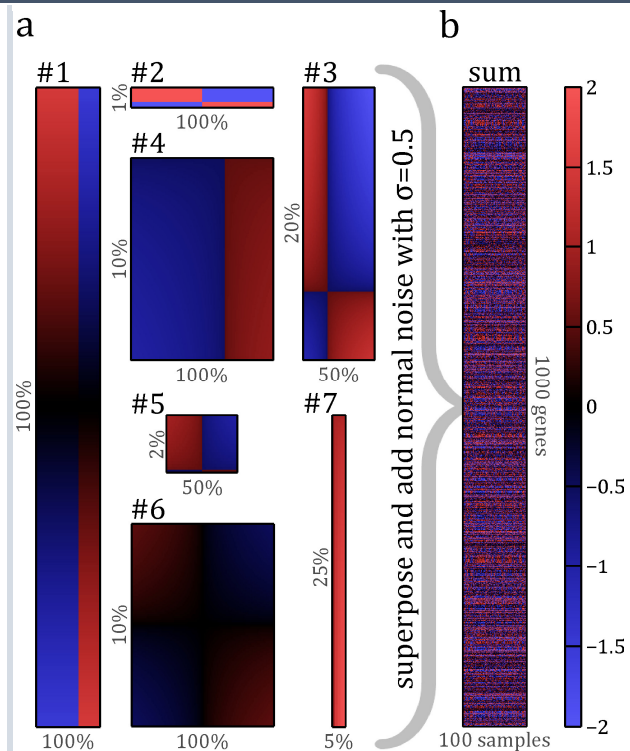


Figure II.6.2.1) Versatility test, 7 simulated effect patterns and the superposed signal

(a) Seven effects of depicted size and signal shape are simulated for randomly selected genes and samples. They are detailed in the text. Smaller effects are zoomed.  
(b) Superposed effects are depicted in a common reference order for all genes and samples, including simulated normal noise of standard deviation  $\sigma = 0.5$ .

To illustrate how signal dissection works for high-dimensional signals, several detection and dissection iterations for the versatility test scenario with 7 effects are depicted and explained in this section.

First, the large superposed lab effect #1 has to be dissected. Because of its dominant size (resulting in a dominant effect score; cf. II.3.1.6) it is always discovered first:

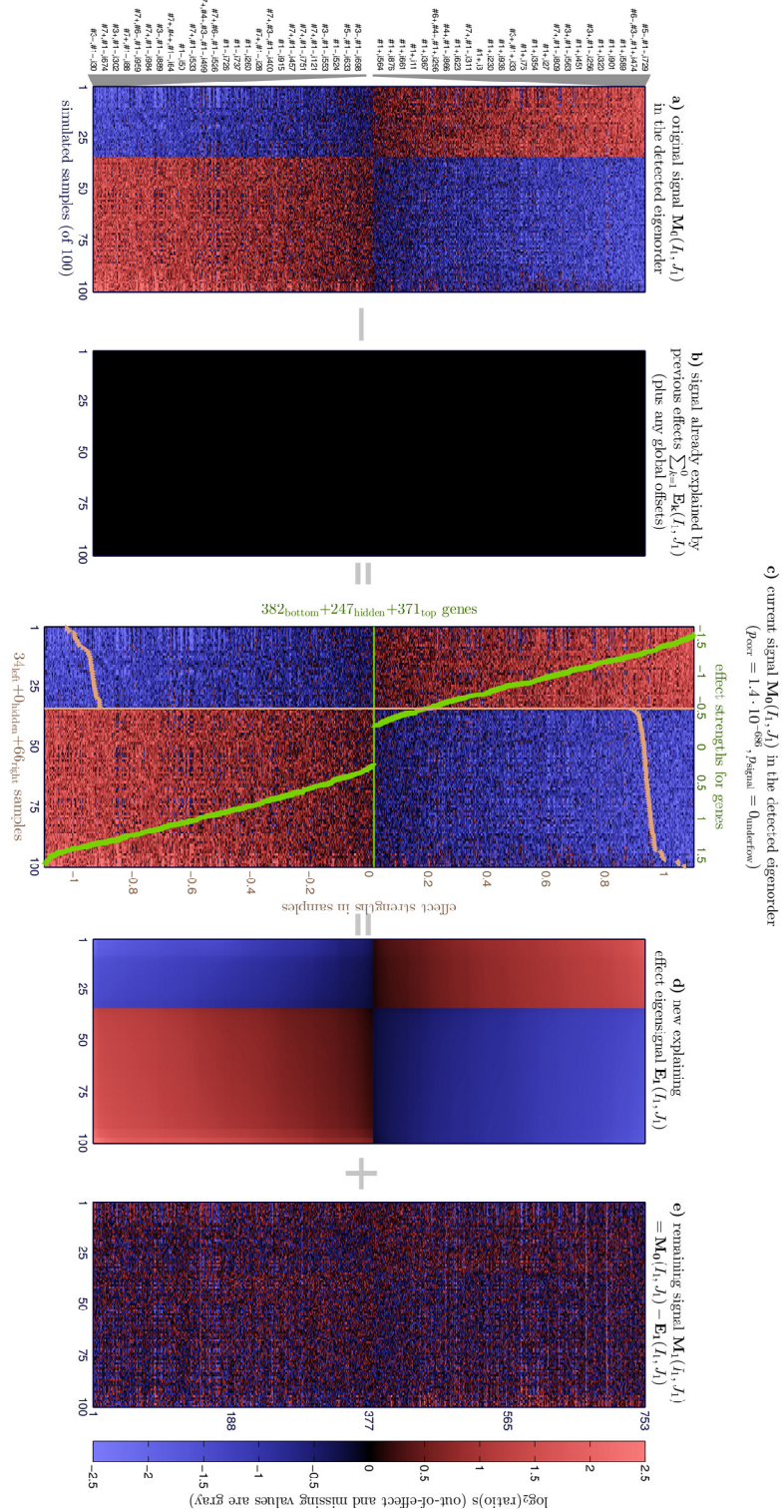


Figure II.6.2.2.a) Versatility test with 7 effects, detection and dissection of the large overlapping lab effect (pattern #1, iteration  $k = 1$ )

Labels of top genes are depicted; each label lists all patterns that influence a particular gene. Plus and minus signs indicate whether they are co-regulated or anti-regulated by an effect. Clearly, genes that are correlated to the detected effect (“#1+”) are sorted to the top, while bottom ranks are exclusively occupied by genes that are anti-correlated to the effect (“#1-“). Final gene strengths  $|u_{1,\hat{i}}^g|$  (center panel, green curve; cf. II.4.1.1) for this effect correctly reflect its linear gradual signal and sample effect strengths  $|u_{1,\hat{i}}^s|$  (center panel, light orange curve) reflect the constant and strong signal in two different sample groups of size 2:1. Dissection of this effect results in a remaining signal (panel e) that is indistinguishable between those sample groups for strong effect strengths. Towards zero signal (and hence towards low correlation) a slight shadow remains due to decreasing dissection strengths (cf. II.4.2.1). This is intended, because in general this shadow could also be just ordered noise. And with dissection of partially correlated effects in mind, dissection should not extend into the uncorrelated regime (cf. II.4.2.2).

Next, the remaining signal is screen for other simulated effects. Pattern #3 is detected at rank #2 (cf. Figure II.6.2.2.b, page 90), because it has the next-highest effect score. (Genes and samples with very weak effect strengths are hidden from the plot to zoom on the effect.) The already explained and dissected signal from pattern #1 is presented in the eigenorder of the currently detected effect (panel b). Due to the first dissection iteration, the empirical signal of pattern #3 (center panel) is much clearer than it was in the original signal (panel a). By virtue of low or zero dissection strengths (gray shading), its eigensignal (panel d) does not extend to samples that are not affected by this effect. Hence, in a real-world gene expression signal comprised of genetically heterogeneous samples, effects that are specific to only a subset of samples (e.g. to a yet unknown disease subtype) can also be detected and dissected. The signal of samples belonging to other subtypes does neither disturb this subtype-specific detection nor is it disturbed by subsequent dissection of this detected effect. This effectively allows a flexible detection of partly overlapping or partly hierarchical effects that may be shared by several but not necessarily by all samples.

To demonstrate the specificity of the method, the small pattern #5 is depicted next (cf. Figure II.6.2.2.c, page 91). It simulates only one anti-regulated gene (of 1000 measured genes). This particular gene occupies the top rank of the detected empirical eigenorder. As the same gene was also regulated by effects #1 and #3 in this simulation (see its gene label), its correlation to other genes regulated by effect #5 is hardly visible in the original signal (panel a), but clearly revealed in the current signal (center panel). Detecting this single anti-regulated gene of pattern #5 is more difficult when many additional noise genes are simulated (cf. II.6.4.2).

The pattern with the weakest simulated signal is detected last (cf. Figure II.6.2.2.d, page 92). Due to overlapping foreign effects, it is barely visible in the original signal (panel a), but still robustly detected (center panel).

Detection and dissection iterations for patterns #2, #4 and #7 are similar. Definition plots and tables of all seven effects are available at [Method Validation\versatility7.single \(nG=1000, nP=100\)\sim 001\B=detected orders](#).

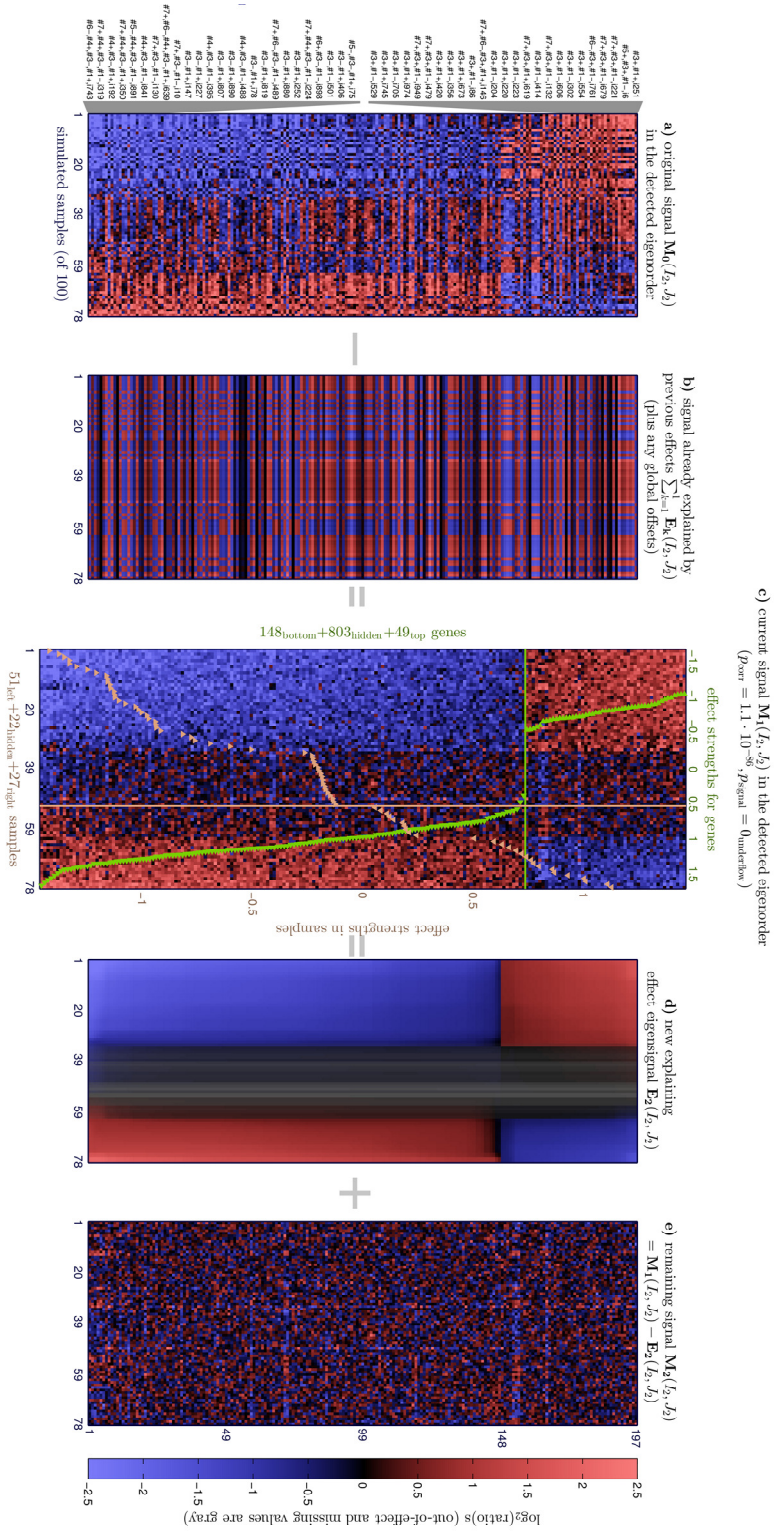


Figure II.6.2.2.b) Versatility test with 7 effects, detection and dissection of pattern #3 (in iteration  $k = 2$ )

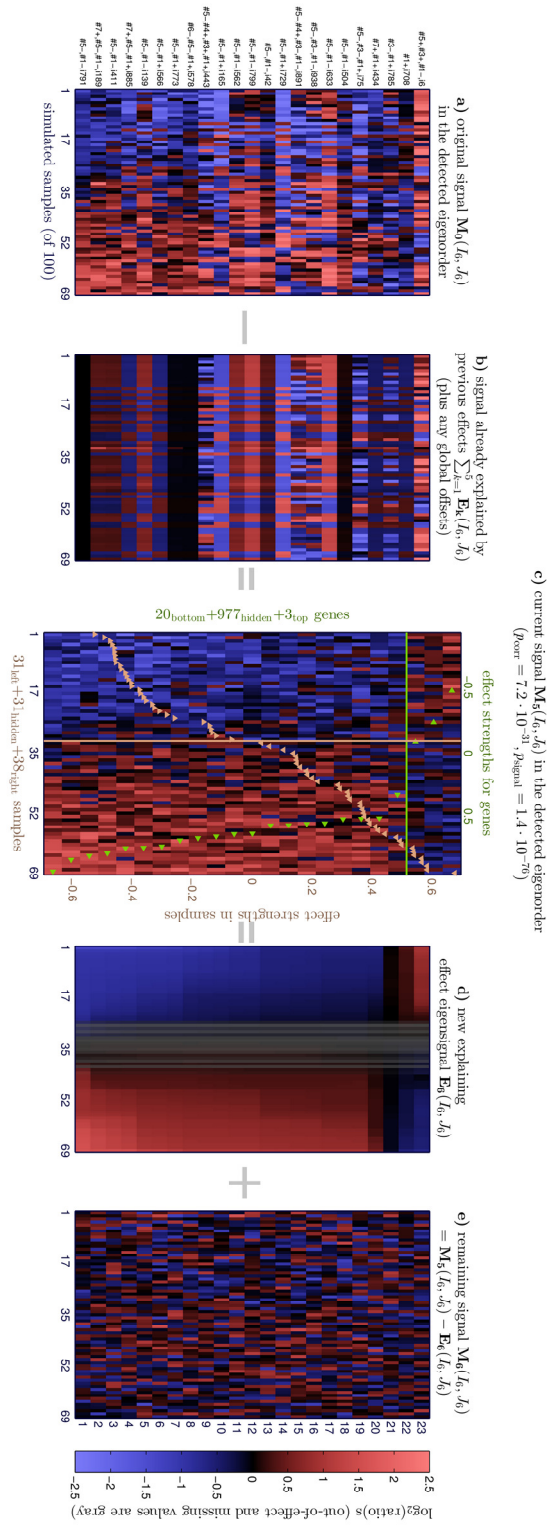


Figure II.6.2.2.c) Versatility test with 7 effects, detection and dissection of pattern #5 (in iteration  $k = 6$ )

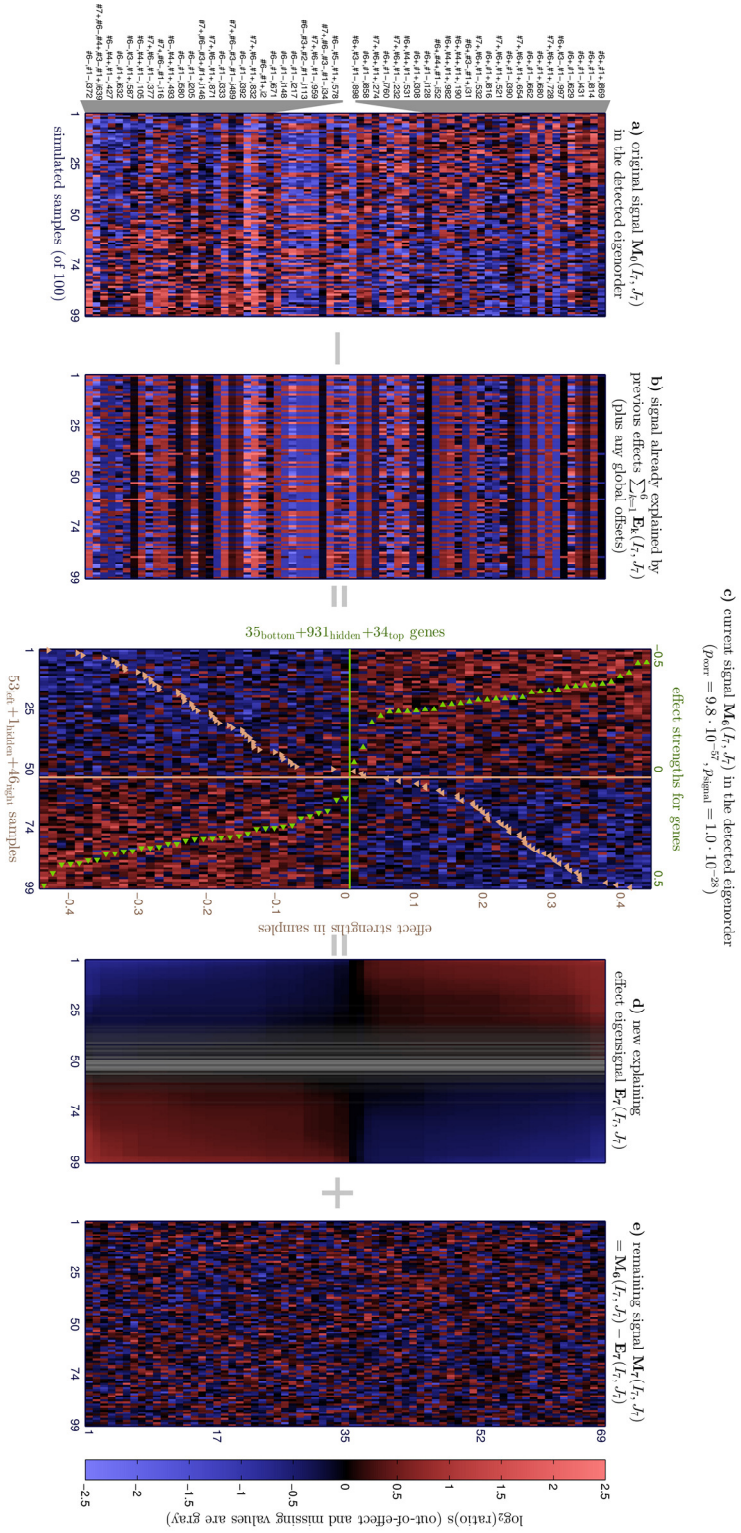


Figure II.6.2.2.d) Versatility test with 7 effects, detection and dissection of pattern #6 (in iteration  $k = 7$ )

### II.6.2.3 Comparison of detected and true simulated effects and results

To quantify detection quality and to facilitate a comparison with simulated effects, I utilize the same uncentered weighted correlations<sup>(cf. II.2.3.1)</sup> that were already used for effect detection.

With a fair comparison to PCA results in mind, I correlate effect axes, i.e. the linear approximations of detected laws of gene regulation. They are directly comparable to principal components. Monotonically regressed effect curves may be more precise approximations of effects, but PCA has no analogue for a direct comparison. Likewise, PCA has no concept of an effect focus. Hence, the detected effect focus should not be used to define correlation weights as there is no analogue when testing PCA results. Instead, identical correlation weights should be used for both, i.e. for correlations of effect axes detected by signal dissection to simulated axes as well as for correlations of principal components to simulated axes. These correlations are defined below based on simulated and detected effect axes. As PCAs always return as many principal components as input dimensions (i.e. they do not determine the number of effects in the signal), only those principal components that show the best correlations with simulated effects are used for comparison. In case of signal dissection, always all detected effects are compared and displayed (even if they have only a weak correlation to all simulated effects). (This is a slight bonus for PCA in the presentation of comparison results.)

Correlation weights should include all genes respectively samples that are strongly involved in the effect with full weight. They should exclude those with weak or zero regulation, because otherwise noise might make the correlation unspecific with respect to the tested effect. As the simulated gene axis  $|a_{\kappa}^{g,sim}\rangle$  of pattern  $\kappa$  is known, it can be utilized to define full weights for all genes with  $\geq 50\%$  simulated signal strength  $\langle e_i^g | a_{\kappa}^{g,sim}\rangle$ . Below this threshold, weights decrease linearly. (Formally, let  $a_{\kappa,i}^{g,sim} \equiv \langle e_i^g | a_{\kappa}^{g,sim}\rangle / \max_i |\langle e_i^g | a_{\kappa}^{g,sim}\rangle|$  denote relative signal strengths. Then  $|w_{\kappa}^{g,sim}\rangle \equiv \sum_{i=1}^m \min(100\%, a_{\kappa,i}^{g,sim} / 50\%) |e_i^g\rangle$ .) Let  $|a_k^g\rangle$  denote the actual gene axis of the  $k^{\text{th}}$  detected effect (respectively principal component). The correlation should not only test the sensitivity of the redetection within the simulated focus  $|w_{\kappa}^{g,sim}\rangle$ . Correlations should also become weaker, if the detected effect axis  $|a_k^g\rangle$  is strong *outside* of the simulated effect (i.e. if there are false positive genes). To this end, let  $|w_{\kappa}^{g,actual}\rangle$  denote analogously defined gene weights based on the *detected* gene axis  $|a_k^g\rangle$ . Balanced weights  $|w_{\kappa,k}^g\rangle$  defined by the maxima  $\langle e_i^g | w_{\kappa,k}^g\rangle \equiv \max(\langle e_i^g | w_{\kappa}^{g,sim}\rangle, \langle e_i^g | w_{\kappa}^{g,actual}\rangle)$  can finally be utilized to facilitate a both sensitive and specific correlation. (For example, if a gene has only 25% weight in  $|w_{\kappa}^{g,sim}\rangle$  because of a relatively weak simulated signal strength, but has a strong false positive signal in a detected effect, it will still be weighted with 100% in the comparison.)

With these balanced weights, finally weighted correlations<sup>(cf. II.2.3.1)</sup> to quantify the rediscovery of simulated gene axes are computed. Analogously, correlations between simulated and detected sample axes can be obtained.

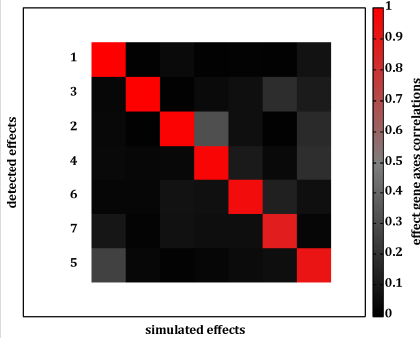


Figure II.6.2.3.a) Versatility test with 7 effects, correlation of detected and simulated gene axes

$$[a_{\kappa}^g | a_{\kappa}^{g,sim}]_{|w_{\kappa,k}^g\rangle}, [a_{\kappa}^s | a_{\kappa}^{s,sim}]_{|w_{\kappa,k}^s\rangle}$$

Eqn. II.6.2.3) Correlations of detected and simulated effect axes

|  |  |
|--|--|
| $ a_{\kappa}^{g,sim}\rangle,  a_{\kappa}^{s,sim}\rangle$ | Simulated effect axes for pattern $\kappa$ |
| $ a_k^g\rangle,  a_k^s\rangle$                           | Detected effect axes for effect $k$        |
| $ w_{\kappa,k}^g\rangle,  w_{\kappa,k}^s\rangle$         | Balanced weights (see text above).         |

For the dissected simulation of the versatility test scenario presented in the last section, correlations between all simulated and all detected gene axes are depicted<sup>(cf. Figure II.6.2.3.a)</sup>. These correlations demonstrate that each simulated effect has been redetected by exactly one detected effect with a correlation near one<sup>(red diagonal)</sup>. No detected effect axis tries to explain parts of other effects<sup>(black off-diagonal pixels)</sup>. (Results are depicted in best match order; the actual detection order of effects is shown by the respective values of  $k$  on the left of the matrix.)

Analogously and likewise, correlations  $[a_k^s | a_k^{s, sim}]_{w_{k,k}^s}$  compare simulated and detected *sample axes* (cf. Figure II.6.2.3.b). They are redetected with similar quality in this simulation.

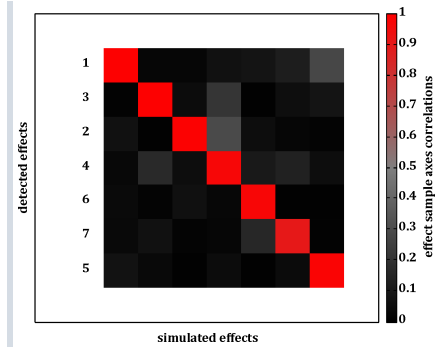


Figure II.6.2.3.b) Versatility test with 7 effects, correlation of detected and simulated sample axes

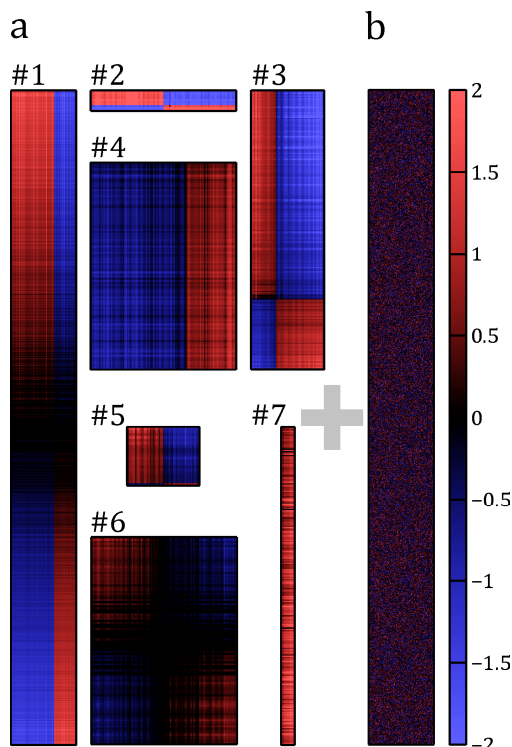


Figure II.6.2.3.c) Detected effect eigensignals in their true simulated eigenorder for the versatility test with 7 effects

Eigensignals are bimonotonic per construction in their respective discovered empirical effect eigenorders. To visualize the compatibility of these empirical eigenorders with the actually simulated ones, regressed eigensignals are displayed here in the respective true effect eigenorders. (a) Empirically estimated eigenorders clearly reproduce all simulated large-scale signal differences of effects. However, local gradual signals are not perfectly reproduced because of noise. (b) The remaining signal does not contain any true positive effects any more, as all 7 simulated effects have been detected and dissected.

While correlations of effect axes are useful for comparison and validation, they summarize (re)detection quality by only one scalar per effect. To visualize the achievable detection quality with respect to the full signal caused by effects, all detected effect eigensignals are presented here (Figure II.6.2.3.c). They are not depicted in their respective detected eigenorder (where they are perfectly bimonotonic per construction), but in their true simulated eigenorder. Indeed, simulated gradual signals are retrieved with high similarity (cf. Figure II.6.2.1). Only for pattern #6 that has a true signal at or below the noise level, the reconstructed signal is relatively rough. This was expected, because this effect is nearly invisible in the original signal (cf. Figure II.6.2.2.d). Still, it was robustly detected. (The corresponding detection limit with respect to minimal signal strength relative to the noise level is analyzed in II.6.4.1.)

Altogether, these results for the presented detection run (cf. II.6.2.2) already are a first validation of signal dissection against the versatility scenario.

#### II.6.2.4 Comparison with PCA (versatility scenario with 7 effects)

A more comprehensive validation for the small versatility scenario with 7 effects is presented in this section. Additionally, a first systematic comparison with PCA results is provided.

To cover the randomness of effect simulation (cf. II.6.2.1), each scenario has been simulated 49 times (hence, results can be presented compactly as 7x7 block matrices). Each simulated signal has been dissected. Detected gene and sample axes are compared to simulated ones as explained (cf. II.6.2.3).



Identical simulated signals have been passed to principal components analyses. Resulting gene axes<sup>(aka eigengenes in [68]; cf. I.2.2.2)</sup> and sample axes<sup>(aka eigenarrays in [68])</sup> are likewise correlated to simulated effect axes<sup>(cf. II.6.2.3)</sup>.

For the 7-effect versatility test, results are summarized by the following block matrices. Each block corresponds to one simulation and shows correlations as before<sup>(cf. II.6.2.3)</sup>. Blocks at the same position correspond to the same input signal; hence a direct visual comparison to results from PCA is possible.

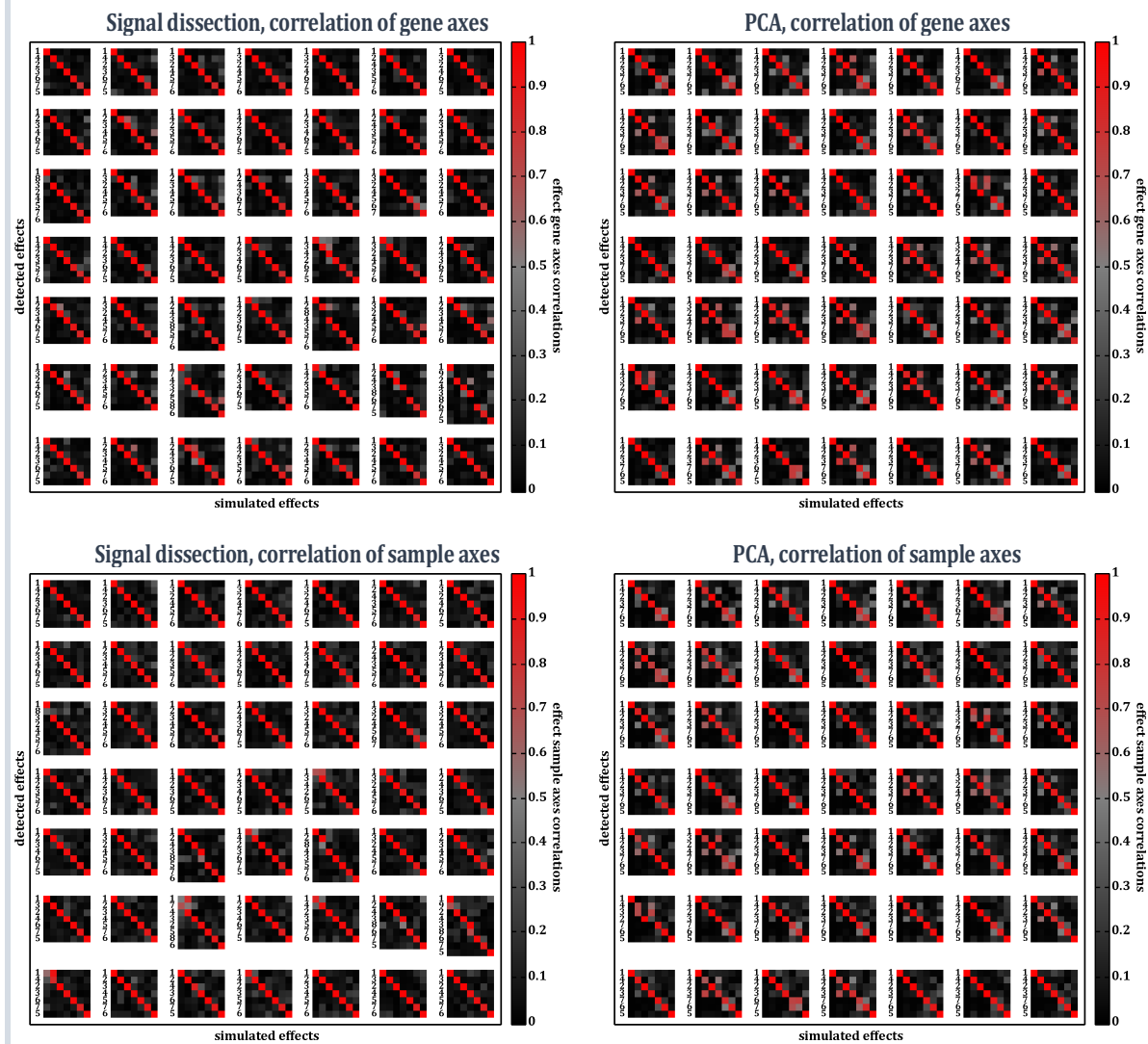


Figure II.6.2.4.a) Versatility test with 7 effects, 49 runs, correlations of effect axes and comparison of signal dissection with PCA

While signal dissection results are precise and consistent, also PCA detects most effects reliably in this test scenario.

Sometimes signal dissection finds more than 7 effects (e.g. row 6, column 7), i.e. the remaining signal is considered as noise later than the optimum. However, minimizing false negatives is a higher priority than minimizing false positives, especially with cross-cohort validation for real-world data in mind<sup>(cf. III.1.2)</sup>. In case of PCA, always the top correlated principal components are displayed (from all  $m$  respectively  $n$  returned principal components), because PCA cannot determine the number of effects in the respective signal.

Looking at details, principal components much more frequently mix two or more distinct simulated effects compared to signal dissection (as depicted by high correlations that are not on the diagonal). This problem gets more severe for more complex signals, as tested next.

### II.6.2.5 Comparison with PCA (versatility scenario with 13 effects)

To increase the detection and dissection difficulty, the same simulations and comparisons as presented in the last section are repeated for the versatility scenario with 13 effects<sup>(cf. II.6.2.1)</sup>.

While the quality of signal dissection stays approximately constant, PCA results are considerably less reliable for this more complex signal. In all simulations, several principal components represent mixtures (i.e. linear combinations) of more than one simulated effect. This is analogous to 3D examples before<sup>(see Figure II.6.1.2 and cf. Figure II.4.2.2.b)</sup>. Supposedly, the higher rate of overlapping effects causes directions in gene space to show highest variance *where effects overlap*. Hence, principal components follow these directions instead of effect axes along which simulated laws of gene regulation really extend. This is the case for both gene axes<sup>(cf. Figure II.6.2.5.a, page 97)</sup> and for sample axes<sup>(cf. Figure II.6.2.5.b, page 98)</sup>.

To summarize and quantify these visual comparison results, two questions are tested:

- 1) Does signal dissection show significantly *higher* correlation to the respective true simulated effect?
- 2) Does signal dissection show significantly *lower* correlation to other effects?

The first question effectively asks for a form of detection sensitivity, while the second asks for detection specificity. These questions can be answered with paired *t*-tests for each simulated effect. Correlations to simulated axes<sup>(cf. Eqn. II.6.2.3)</sup> for signal dissection results and PCA results are compared (and paired by simulation runs, i.e. by identical input signals). For the first question, correlations of the best-matching effect axis (respectively principal component) with the respective simulated effect<sup>(as depicted by diagonal pixels in Figure II.6.2.5.a and Figure II.6.2.5.b)</sup> are compared between signal dissection and PCA over all simulation runs. For the second question, average correlations to all other simulated effects<sup>(as depicted by all non-diagonal pixels in the same column)</sup> are compared. Results for each simulated effect are listed to the right<sup>(cf. Table II.6.2.5)</sup>. (For each question, *p* values for both methods add to one.)

Due to its signal strength and size, pattern #1 defines a clear direction of maximal variance that coincides with its simulated effect axis. Hence, PCA is excellent at detecting this overlapping lab effect. For example, in all 49/49 simulations a principal component exits with a correlation > 0.95 to the gene axis of pattern #1, while this is only the case in 37/49 simulations for signal dissection. However, both methods detect this strong effect with gene axes correlations > 0.9 in all 49/49 simulations.

Additionally, PCA is better at detection of the weak signal pattern #6 (e.g. gene axes correlations are > 0.8 in 38/49 runs for PCA, but only in 26/49 runs for signal dissection). Due to its weak signal relative to the noise level, neither method can

| Gene axes correlations |              |                          |         |                          |         |
|------------------------|--------------|--------------------------|---------|--------------------------|---------|
| sim. effect            | sim. pattern | sensitivity (question 1) |         | specificity (question 2) |         |
|                        |              | Sig. diss.               | PCA     | Sig. diss.               | PCA     |
| 1                      | 1            | 0.9963                   | 0.0037  | 1.0000                   | 7.2E-24 |
| 2                      | 2            | 1.6E-08                  | 1.0000  | 2.8E-24                  | 1.0000  |
| 3                      | 3            | 1.2E-12                  | 1.0000  | 0.0029                   | 0.9971  |
| 4                      | 4            | 3.8E-17                  | 1.0000  | 9.5E-12                  | 1.0000  |
| 5                      | 5            | 0.3803                   | 0.6197  | 1.8E-17                  | 1.0000  |
| 6                      | 6            | 0.9999                   | 7.6E-05 | 7.6E-13                  | 1.0000  |
| 7                      | 7            | 2.3E-09                  | 1.0000  | 7.1E-10                  | 1.0000  |
| 8                      | 2            | 2.4E-10                  | 1.0000  | 3.2E-26                  | 1.0000  |
| 9                      | 3            | 3.8E-12                  | 1.0000  | 1.5E-05                  | 1.0000  |
| 10                     | 4            | 1.5E-13                  | 1.0000  | 1.6E-09                  | 1.0000  |
| 11                     | 2            | 1.5E-08                  | 1.0000  | 6.8E-22                  | 1.0000  |
| 12                     | 3            | 4.7E-12                  | 1.0000  | 0.0013                   | 0.9987  |
| 13                     | 4            | 9.0E-17                  | 1.0000  | 7.6E-24                  | 1.0000  |

| Sample axes correlations |              |                          |         |                          |         |
|--------------------------|--------------|--------------------------|---------|--------------------------|---------|
| sim. effect              | sim. pattern | sensitivity (question 1) |         | specificity (question 2) |         |
|                          |              | Sig. diss.               | PCA     | Sig. diss.               | PCA     |
| 1                        | 1            | 1.0000                   | 1.9E-05 | 1.0000                   | 1.1E-17 |
| 2                        | 2            | 3.6E-22                  | 1.0000  | 1.0E-21                  | 1.0000  |
| 3                        | 3            | 6.7E-18                  | 1.0000  | 5.4E-05                  | 0.9999  |
| 4                        | 4            | 1.7E-26                  | 1.0000  | 2.8E-23                  | 1.0000  |
| 5                        | 5            | 2.5E-15                  | 1.0000  | 5.2E-16                  | 1.0000  |
| 6                        | 6            | 0.3351                   | 0.6649  | 1.8E-10                  | 1.0000  |
| 7                        | 7            | 0.7967                   | 0.2033  | 4.6E-17                  | 1.0000  |
| 8                        | 2            | 1.3E-22                  | 1.0000  | 3.4E-19                  | 1.0000  |
| 9                        | 3            | 3.6E-15                  | 1.0000  | 8.5E-07                  | 1.0000  |
| 10                       | 4            | 3.1E-19                  | 1.0000  | 1.3E-20                  | 1.0000  |
| 11                       | 2            | 9.2E-23                  | 1.0000  | 5.2E-19                  | 1.0000  |
| 12                       | 3            | 2.5E-19                  | 1.0000  | 2.6E-11                  | 1.0000  |
| 13                       | 4            | 1.4E-27                  | 1.0000  | 2.5E-18                  | 1.0000  |

Table II.6.2.5) Versatility test scenario with 13 effects, 49 runs, comparison between signal dissection and PCA results

Based on depicted correlations between simulated and detected gene axes<sup>(cf. Figure II.6.2.5.a)</sup> respectively sample axes<sup>(cf. Figure II.6.2.5.b)</sup>, signal dissection and PCA are compared. For detection sensitivity with respect to each simulated effect, correlations of the best-matching detected effect axis respectively of the best-matching principal component are compared over all 49 simulation runs (question 1). For detection specificity, average correlations of the same detected effect axis respectively principal component to all other simulated effects are compared (question 2). Both comparisons are realized by one-tailed paired *t*-tests (paired by simulation runs, i.e. by identical input signal).

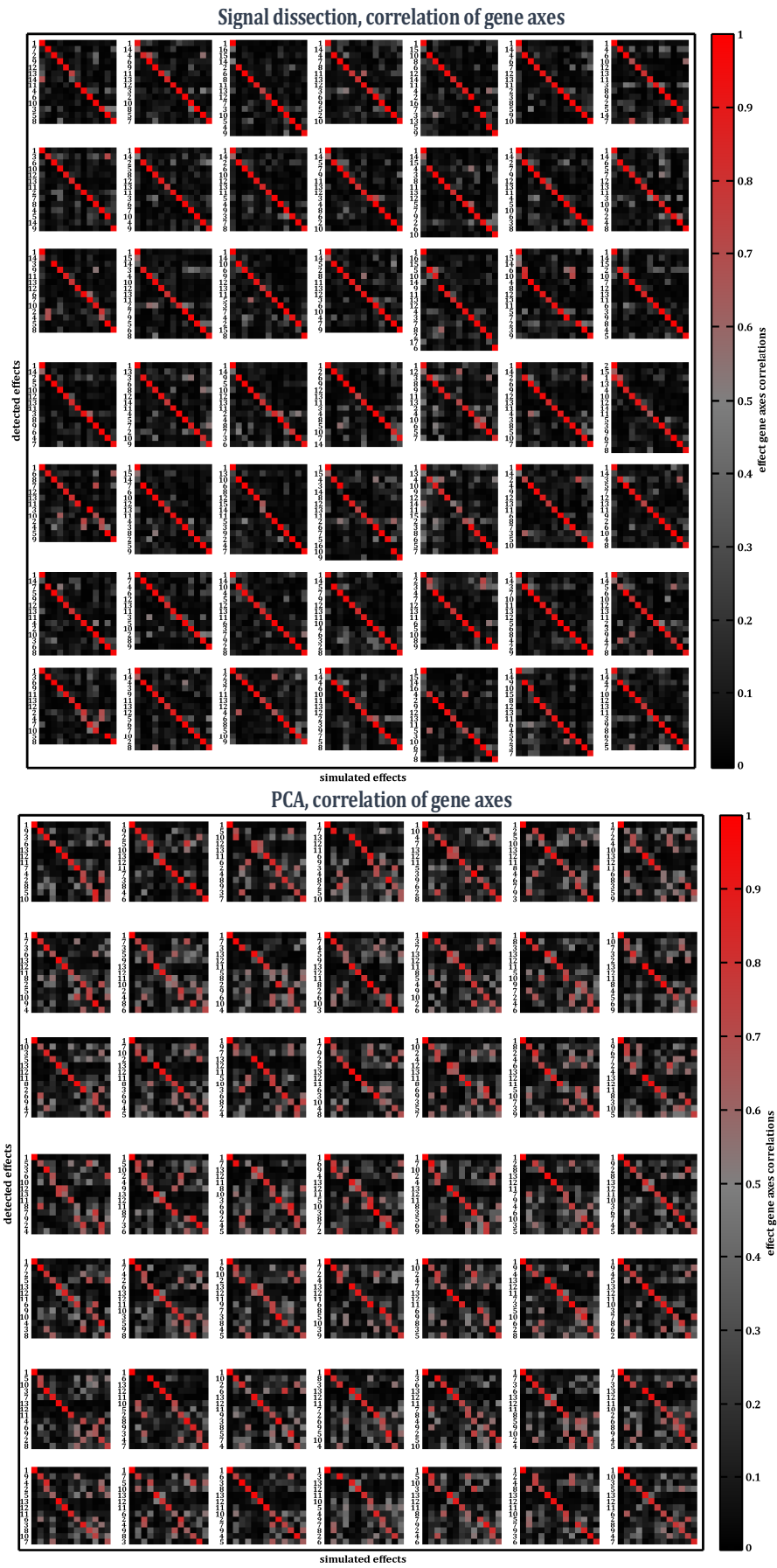


Figure II.6.2.5.a) Versatility test with 13 effects, 49 runs, correlations of gene axes and comparison of signal dissection with PCA

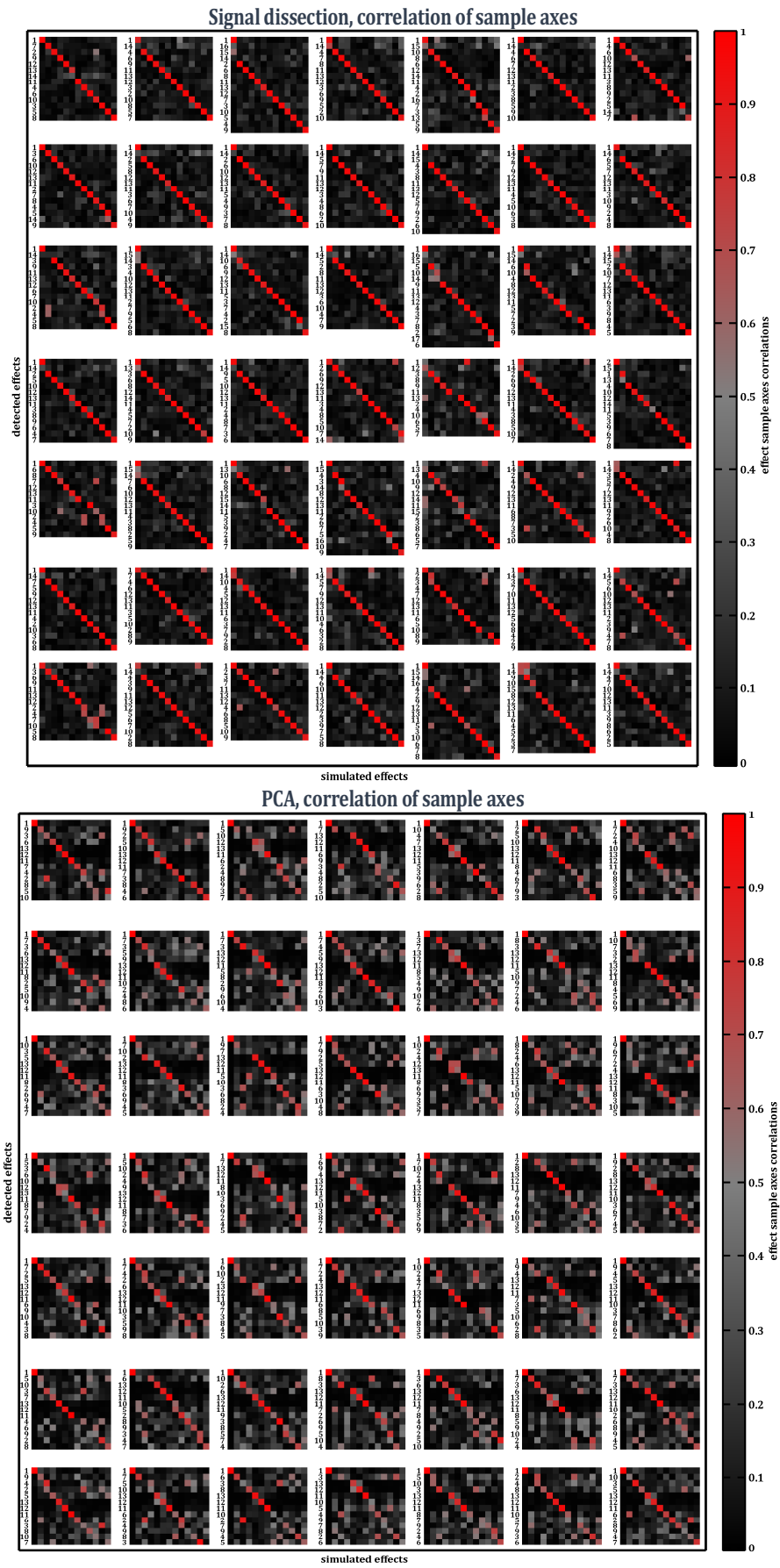


Figure II.6.2.5.b) Versatility test with 13 effects, 49 runs, correlations of sample axes and comparison of signal dissection with PCA

detect the original simulated gene axes with high precision (0/49 for correlations  $> 0.95$  for both methods).

With respect to the narrow pattern #7, the correct 5% of samples influenced by this effect are detected by both methods with similar reliability (sample axes correlations  $> 0.95$  in 43/49 runs for signal dissection and in 42/49 runs for PCA).

*All other effects are detected significantly better by signal dissection.* For example, all three simulated instances of the strong yet small signal pattern #2 are detected by PCA with correlations  $> 0.8$  only in 3/49, 1/49 respectively 3/49 runs. Signal dissection detects the same effects in 45/49, 45/49 respectively 44/49 runs. For a correlation cutoff  $> 0.95$  still 30/49, 34/49 respectively 34/49 detections remain, while PCA detects zero effects of pattern #2 with this correlation.

*Additionally, PCA results for most effects are significantly less specific,* as principal components are also correlated with several other effects instead of exclusively with their best-matching simulated effect (cf. *p* values on the right side in Table II.6.2.5). I.e. effects cannot be cleanly dissected in coordinates of these principal components (visible by off-diagonal correlations in Figure II.6.2.5.a and Figure II.6.2.5.b).

In brief, these results thoroughly validate signal dissection against the 13-effect versatility test. PCA returns for identical input signals comparably unreliable and hard to interpret results.

To compare with hierarchical clustering, it has been applied to the versatility simulation with 7 effects that has been presented and dissected in II.6.2.2. Based on 3D results for hierarchical clustering<sup>(cf. II.6.1.3)</sup> and its typical configuration for gene expression signals<sup>(cf. I.2.2.1)</sup>, the correlation distance is chosen (with average linking).

As expected, the distance metric is dominated by the overlapping strong lab effect (pattern #1) and resulting dendrograms organize genes and samples accordingly<sup>(cf. Figure II.6.2.6)</sup>. Hence, pattern #1 can be considered detected. As discussed for method design goals<sup>(cf. II.1.2)</sup>, the overlapping effect dominates the clustering and cannot be removed by hierarchical clustering as a matter of method concept. Hence, all other effects are split by this independent lab effect. Additionally, they split each other into sub clusters, if they overlap. This is visible in form of red or blue “blocks”. Genes belonging to the same simulated effect are clustered together in these blocks, but only in a fragmented form.

In practice, it may be possible to manually focus on sub clusters in the gene dendrogram and then apply hierarchical clustering recursively in order to obtain a sample clustering for each gene sub cluster<sup>(cf. II.1.2.2)</sup>. But this leads to hard-to-compare results, as the manual or visual determination of gene sub clusters is often difficult to reproduce.

As hierarchical clustering has no concept of gene or sample axes, it cannot be compared directly and systematically with signal dissection as is possible for PCA. Therefore and because PCA is conceptually more competitive, all subsequent test scenarios are compared with PCA only.

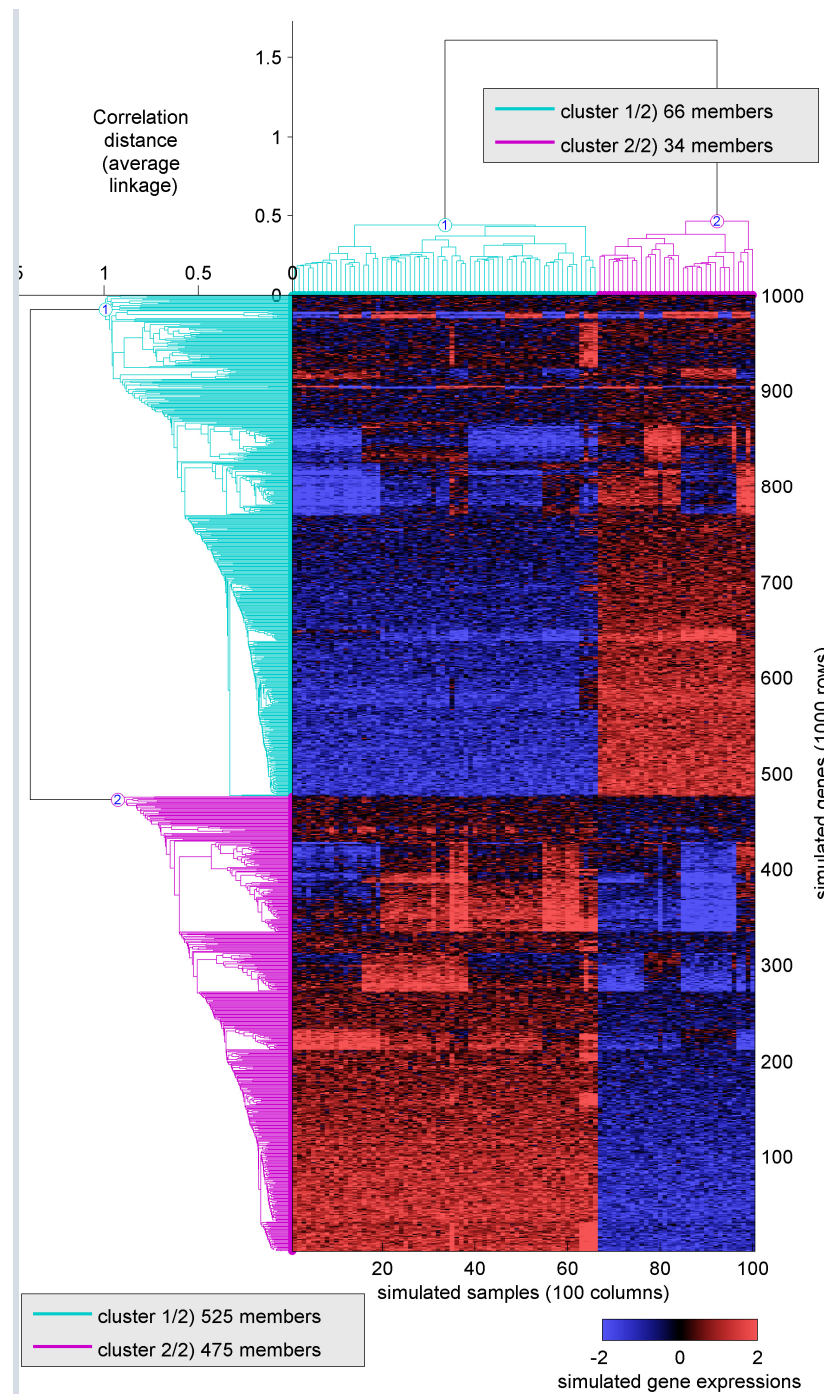


Figure II.6.2.6) Versatility test with 7 effects, results from hierarchical clustering (corr. distance, average linking)

Besides a versatile detection of effects, dissection should work even if the same genes and samples are affected by multiple superposed effects.

Using the same size of the signal matrix as in the versatility scenario (1000 genes, 100 samples), this is tested here by simulating *one to 20 instances* of effects of the same form (either pattern #3, pattern #4 or pattern #6 from the versatility scenario; cf. II.6.2.1).

For higher superposition depths, all genes or samples are regulated by multiple simulated effects eventually. This stresses dissection. These scenarios also stress the estimation of the true noise level<sup>(cf. II.5.1.1)</sup>, because initially the signal standard deviation is much higher than the actual noise level.

Results from signal dissection and from PCA are again correlated to all simulated gene and sample axes and are depicted as before<sup>(e.g. Figure II.6.2.5.a)</sup>.

*Interestingly, PCA fails completely for these superposition scenarios, even for relatively few superposed effects.* This can possibly be explained by PCA's model for interactions again, namely that it searches and computes directions of maximal variance. Maximal variance is found where signals of effects overlap constructively. Hence, these directions of maximal overlapping are returned by PCA (rather than simulated effect axes). Additionally, no variance difference due to different effect sizes exists any longer compared to the versatility scenario. Given that in real-world gene expression signals effects of similar size are common, PCA's inability to deal with these scenarios is this is a major disadvantage with biological interpretability of discovered gene expression effects in mind.

### II.6.3.1 Results and comparison with PCA for one to 20 times pattern #3

Results from signal dissection and PCA for simulated signals comprised of one to 20 times pattern #3 are presented<sup>(cf. Figure II.6.2.1)</sup>. The same noise level as for the versatility scenario is simulated ( $\sigma = 0.5$ ).

While signal dissection detects precise and highly correlated effect axes for all simulated effects in all simulations, PCA fails to provide interpretable results in any simulation except for the single-effect signal<sup>(cf. Figure II.6.3.1.a and Figure II.6.3.1.b)</sup>.

For higher superposition depths, signal dissection often detects more effects than simulated. These are either false positives (if they are correlated to no simulated effect) or duplicates. The reason for the former is that the true noise level is hard to estimate for busy signals with many overlapping true positive effects. Besides, it is typically more useful to accept few false positives rather than accepting false negatives, because false positives can be sorted out in real-world scenarios via cross-cohort validation later<sup>(cf. III.1.2)</sup>. PCA does not provide any noise estimate and always returns as many principal components as input dimensions (including many false positive axes). Only top correlated principal components are selected and depicted.

The high overlapping rate causes correlations to degrade, resulting lower dissection strengths<sup>(cf. II.4.2.1)</sup>. Hence, only effect genes with relatively low overlapping grade are dissected first. The remaining parts of the same effect are detected and dissected in later iterations, after overlapping and disturbing signals from other effects have been dissected, i.e. after the signal has been "cleaned". This is visible as duplicate detections. These duplicates could be easily identified and merged via post-processing, as their detected effect axes are highly correlated. Hence, duplicate detections are preferable over overoptimistic dissection strengths that might result in larger detected effects that represent mixtures of overlapping simulated effects.

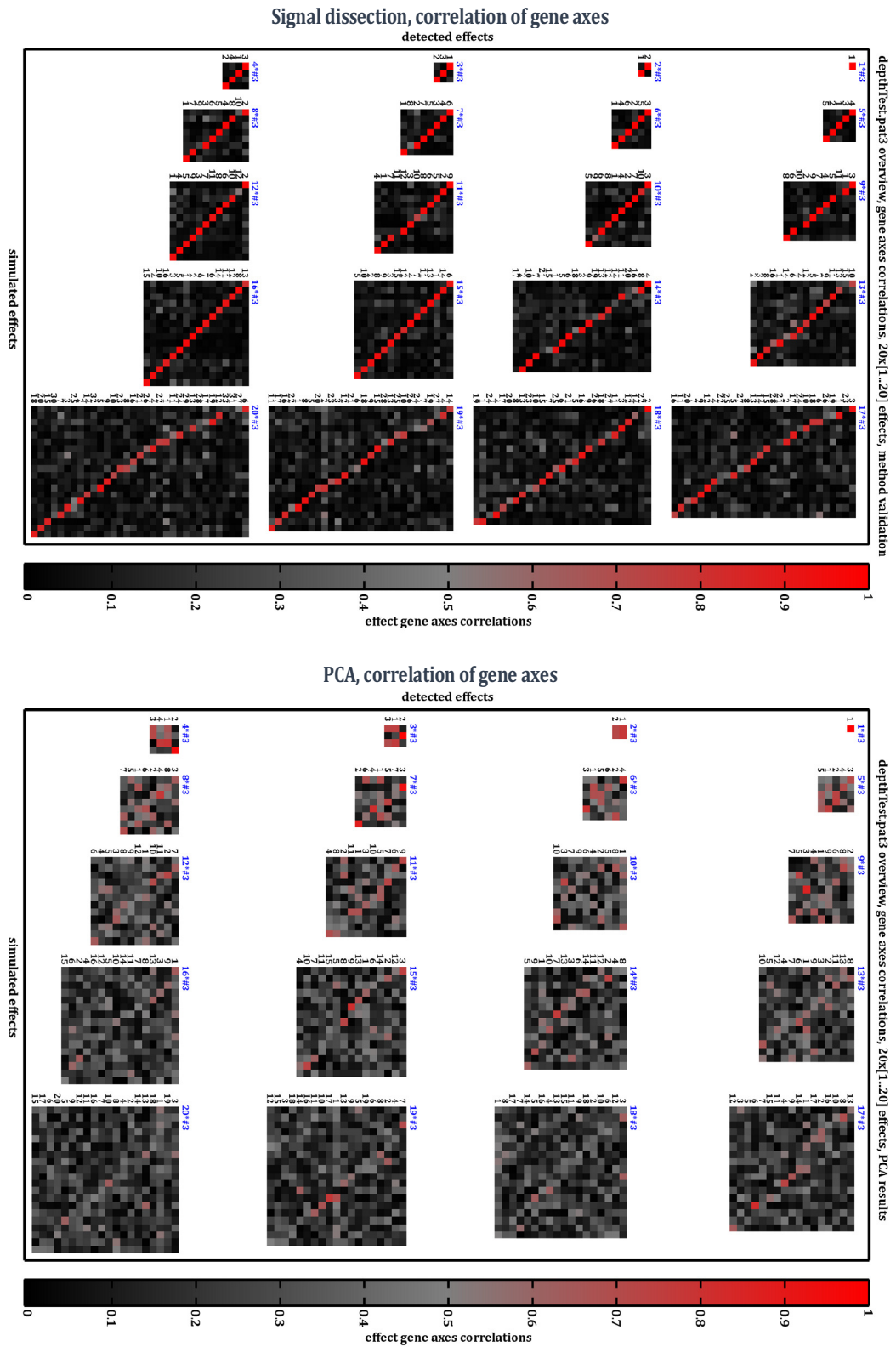


Figure II.6.3.1.a) Superposition tests based on pattern #3, 1 to 20 instances, correlations of gene axes and comparison of signal dissection with PCA



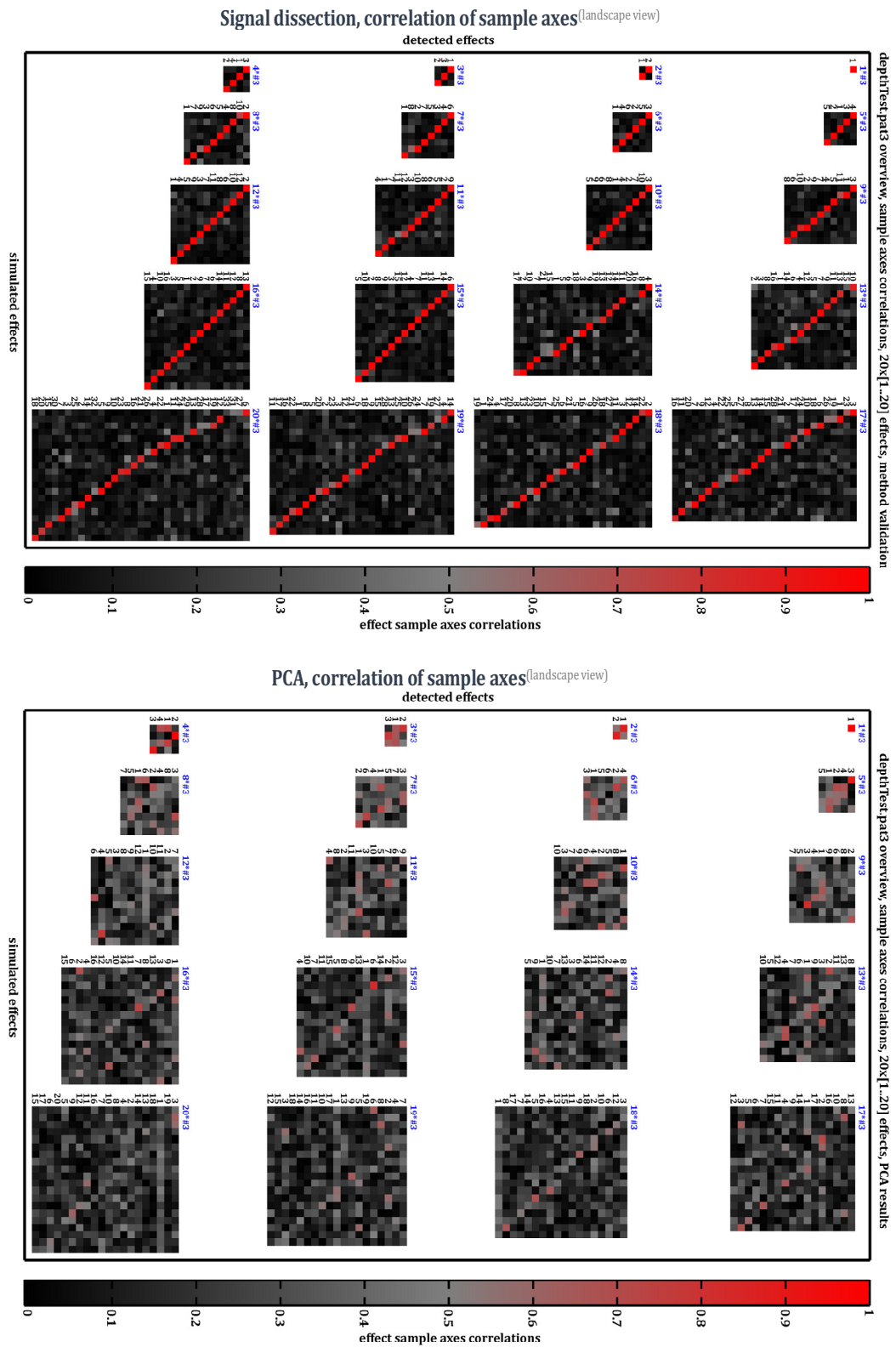


Figure II.6.3.1.b) Superposition tests based on pattern #3, 1 to 20 instances, correlations of sample axes and comparison of signal dissection with PCA

To demonstrate the grade of overlapping up to which effects are detected here, the following figure shows the detection of the 19<sup>th</sup> instance of pattern #3 (detected in iteration  $k = 20$ ). Gene labels at the top and bottom list all effects that regulate these genes simultaneously in the original signal. Subtracting already detected effects (panel b) reveals the cleaned effect (as depicted in its discovered eigenorder in the center panel).

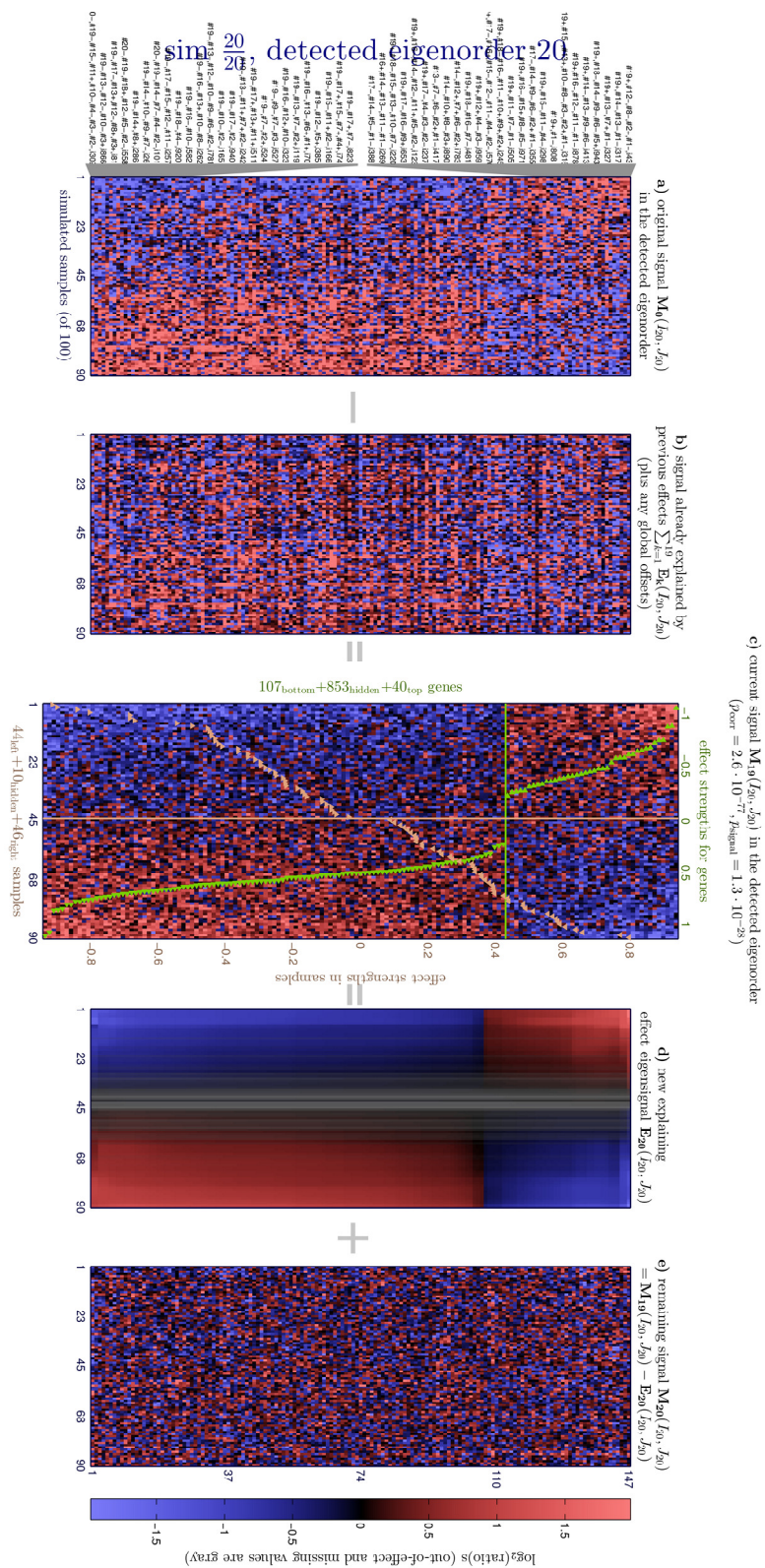


Figure 11.6.3.1.c) Superposition test based on pattern #3, dissection of the signal with 20 superposed effect instances, detection iteration 20

### II.6.3.2 Results and comparison with PCA for one to 20 times pattern #4

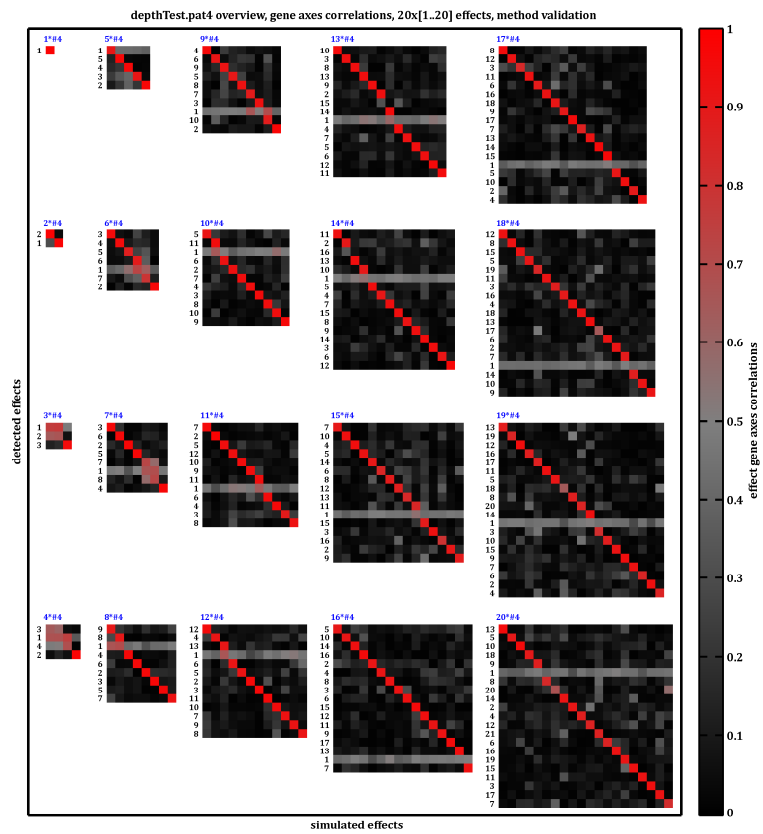
---

In the last section, the two-sided effect pattern #3 (having both correlated and anti-correlated genes) has been simulated and dissected. To demonstrate that this still works for one-sided effects (that regulate all genes in the same direction and have no anti-correlated genes), the superposition test has been repeated with effect pattern #4 (cf. Figure II.6.2.1).

Correlations of detected and simulated gene axes (cf. Figure II.6.3.2.a) again show that signal dissection reliably discovers simulated effects in most simulations, while PCA cannot compete. Exceptions are signals with four or less effects as explained below. (Correlations for sample axes are similar. Corresponding plots are available at [Method Validation\depthTest.pat4 \(nG=1000, nP=100\)](#), together with all other effect dissection plots for each simulation of this test scenario.)

Superposing many effects of this one-sided pattern *accumulates a global nonzero offset*. I purposely keep it in the signal for additional difficulty and to test how the method can handle it. From five effect instances onwards, the *first detected effect represents the combined global offset* (gray rows). After dissecting it, all simulated effects are detected with high correlation (red pixels). Hence, signal dissection realizes a signal normalization on the fly by dissecting an offset effect. A depicted example (cf. Figure II.6.3.2.b) shows that top genes of this offset effect are the most-overlapped genes, as expected. For only three or four instances of #4 in the simulated signal however, there is an ambiguity. Does the emerging constructive global offset already have higher (uncentered) correlation or still any single simulated effect? For this ambiguity, the method is as unsuitable as PCA for three or four instances of the effect. Hence typical normalization steps (e.g. subtracting the median expression per gene) should not be omitted in general.

## Signal dissection, correlation of gene axes



## PCA, correlation of gene axes

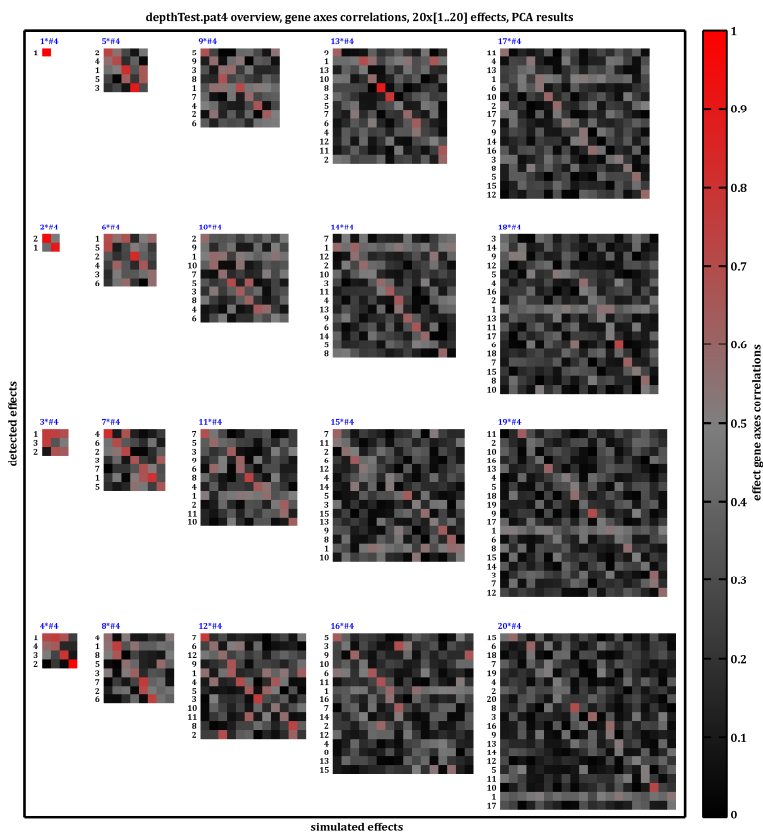


Figure II.6.3.2.a) Superposition tests based on pattern #4, 1 to 20 instances, correlations of gene axes and comparison of signal dissection with PCA

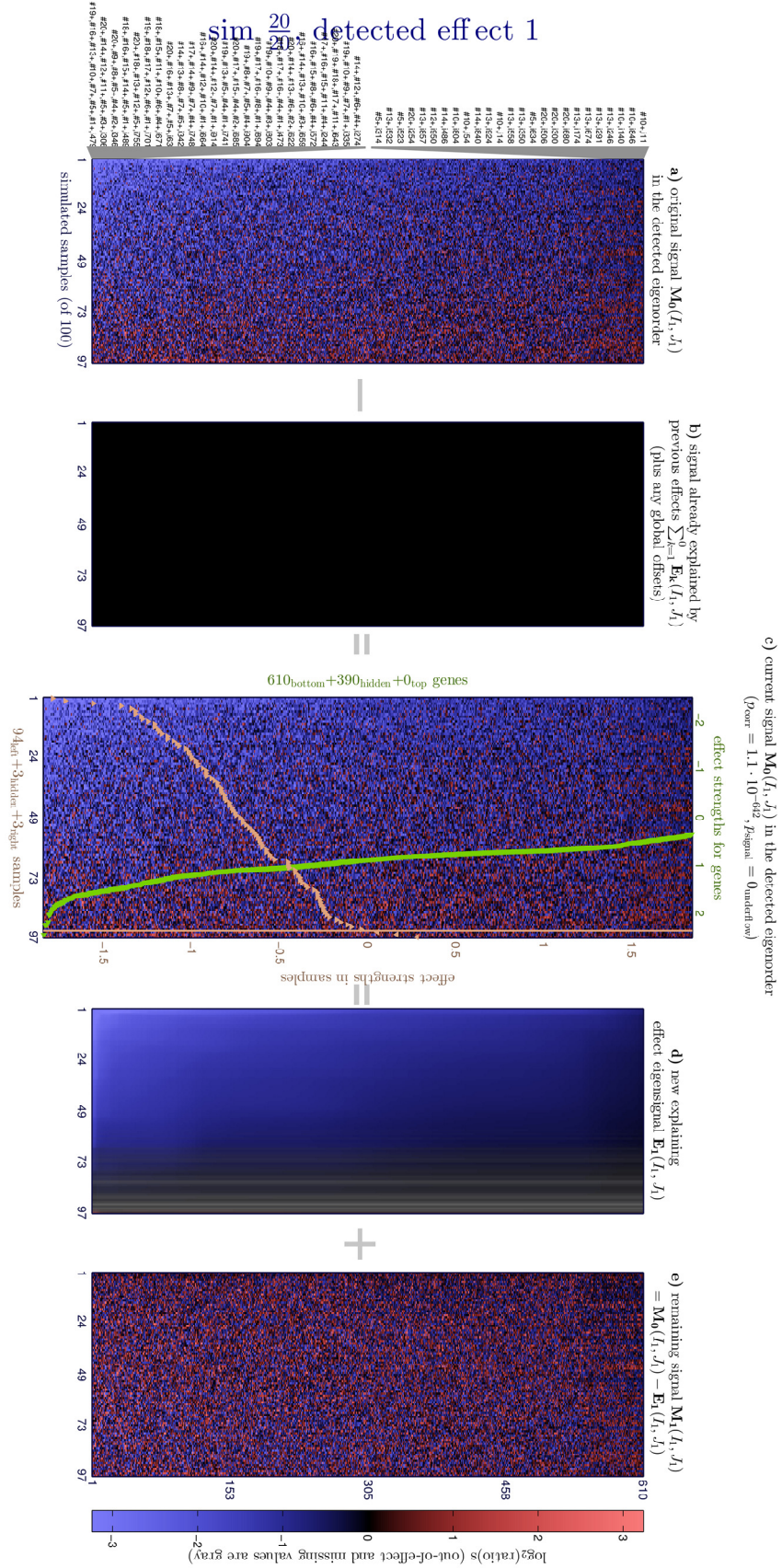


Figure II.6.3.2.b) Superposition tests based on pattern #4, 20 instances, detection and dissection of the accumulated offset effect

### II.6.3.3 Results and comparison with PCA for one to 20 times pattern #6

---

To approach the limits of signal dissection, the superposition scenario has been repeated with pattern #6, i.e. with effects that have a signal at or below the simulated noise level everywhere. This signal is gradual and linearly decreases from top regulated genes respectively samples to zero<sup>(cf. Figure II.6.2.1)</sup>.

Here, signal dissection can no longer reliably dissect all simulated effects. Like for the other superposition tests before, PCA cannot compete.

(Results are displayed for gene axes correlations only; sample axes correlations are comparable and can be found at `Method Validation\depthTest.pat6` ( $nG=1000$ ,  $nP=100$ ). All other effect dissection plots for each simulation of this test scenario are provided in subfolders.)

Detection limits are investigated more systematically in the next section.

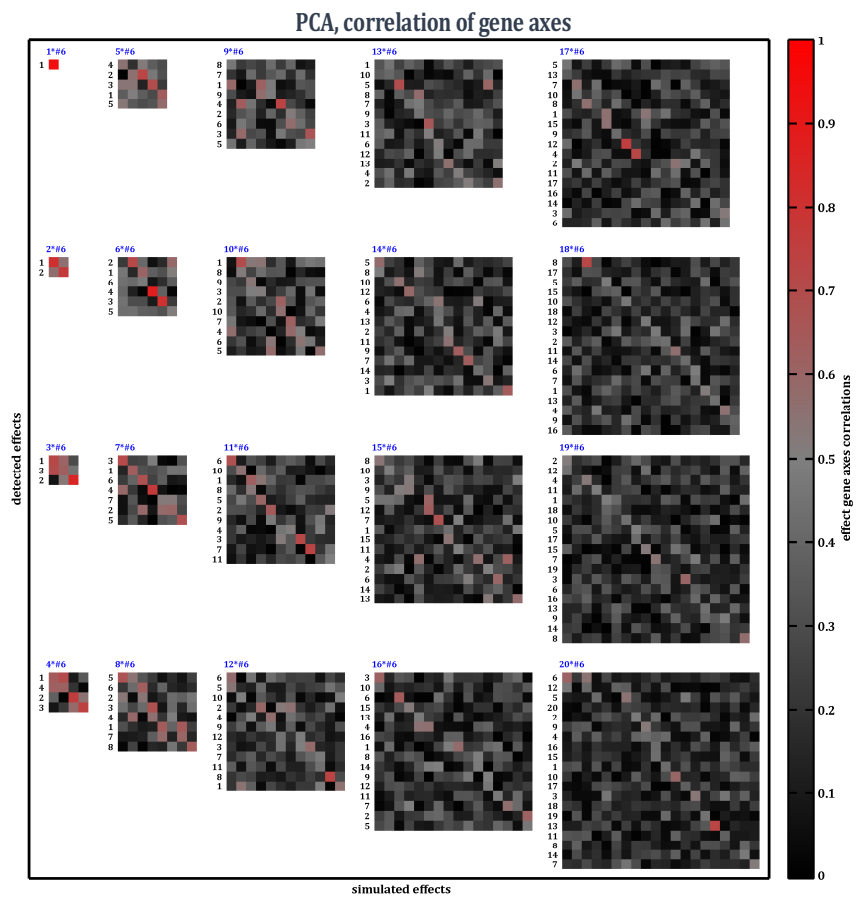
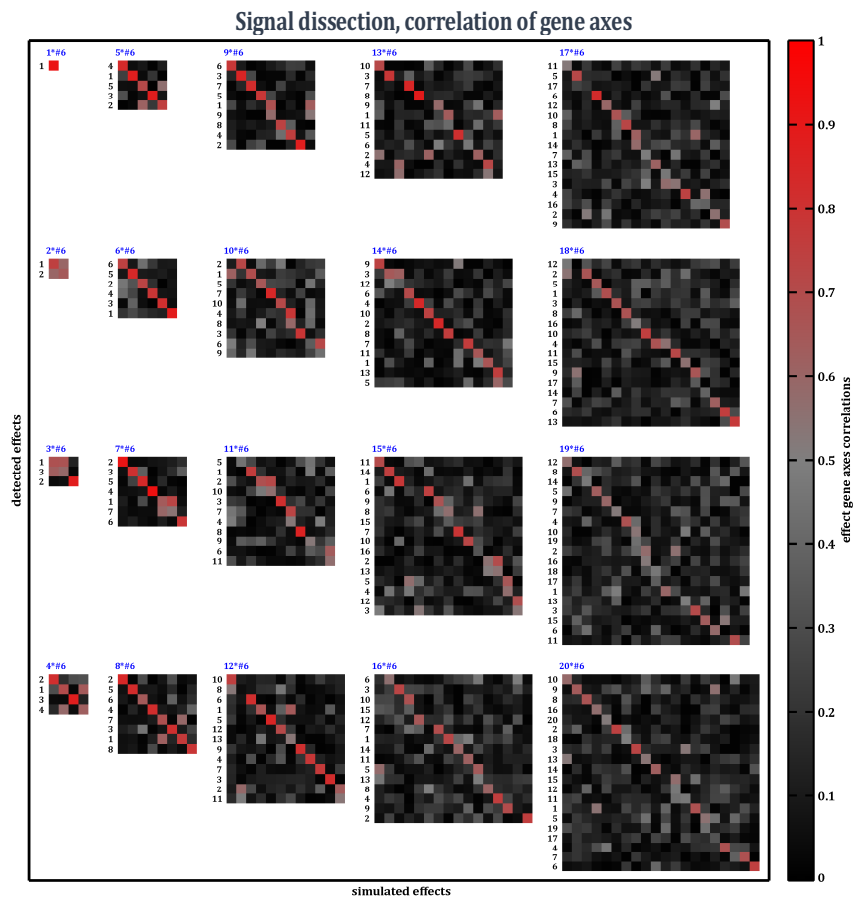


Figure II.6.3.3] Superposition tests based on pattern #6, 1 to 20 instances, correlations of gene axes and comparison of signal dissection with PCA

To systematically test detection limits of signal dissection, several specialized test scenarios have been devised. Again, PCA results for each scenario are provided for comparison.

To quantify the detection limit with respect to signal strength, signals are simulated that contain a single effect of decreasing signal strength relative to the constant simulated noise level  $\sigma$ . Only a single effect is simulated per signal matrix (still of size 1000 genes and 100 samples).

First, effects of the same shape as pattern #6 are simulated. In the versatility test, this pattern has the weakest signal of all effects. More precisely, its absolute signal equals  $1\sigma$  for its top regulated genes and samples and then decreases linearly to zero (cf. Figure II.6.2.1). Below, this shape is simulated 100 times with decreasing top signal strength, as indicated in units of the simulated noise level  $\sigma$  (horizontal axis).

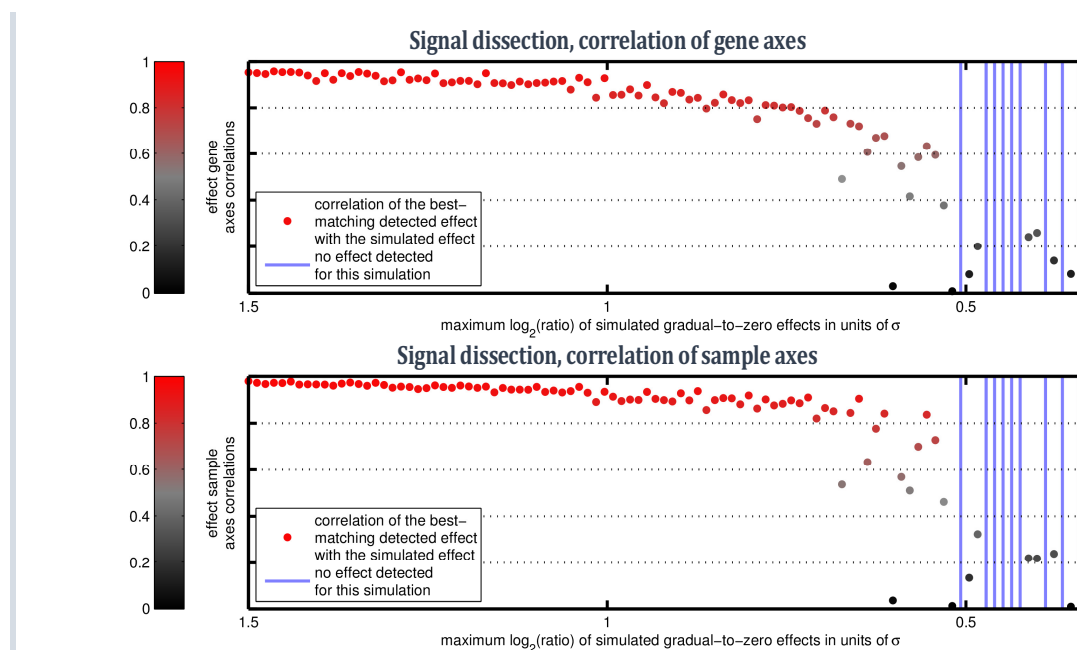


Figure II.6.4.1.a) Detection limit with respect to signal strength, effect pattern #6, signal dissection results

Signal dissection detects the effect down to approximately  $0.5\sigma$ . Below that, more often than not signal dissection terminates without any detected effect.

It is possible to extend this limit by increasing the significance threshold for effect signal strengths (cf. Table II.3.1.8). However, this would not only result in detections of the simulated effect but also of many noise effects (i.e. false positives).

Already before  $0.5\sigma$ , correlations of detected axes to true simulated axes break down. Hence, the detection limit for practice is approximately  $0.75\sigma$ , because in practice not just a detection is required but also a high correlation to true effect axes for interpretability.

Hence, the chosen significance threshold has been configured as intended, *as it enables detection of true positives as long as effect axes may be determined with sufficient correlation, but stops short before accepting many false positives.*



PCA has no concept of noise estimation and does not determine the number of effects in the signal. Instead, it always returns as many principal components as input dimensions. In the following comparison, always the principal component *with maximal variance* is selected and its correlation with the respective simulated effect axis is depicted.

Because simulated signals only contain one true positive effect in this test scenario, axes of maximal variance and simulated effect axes would coincide, if there was no noise. Hence, these PCA results can be seen as the empirical optimum for this scenario. Indeed, correlations are higher than for signal dissection, although differences become negligible for higher signal strengths.

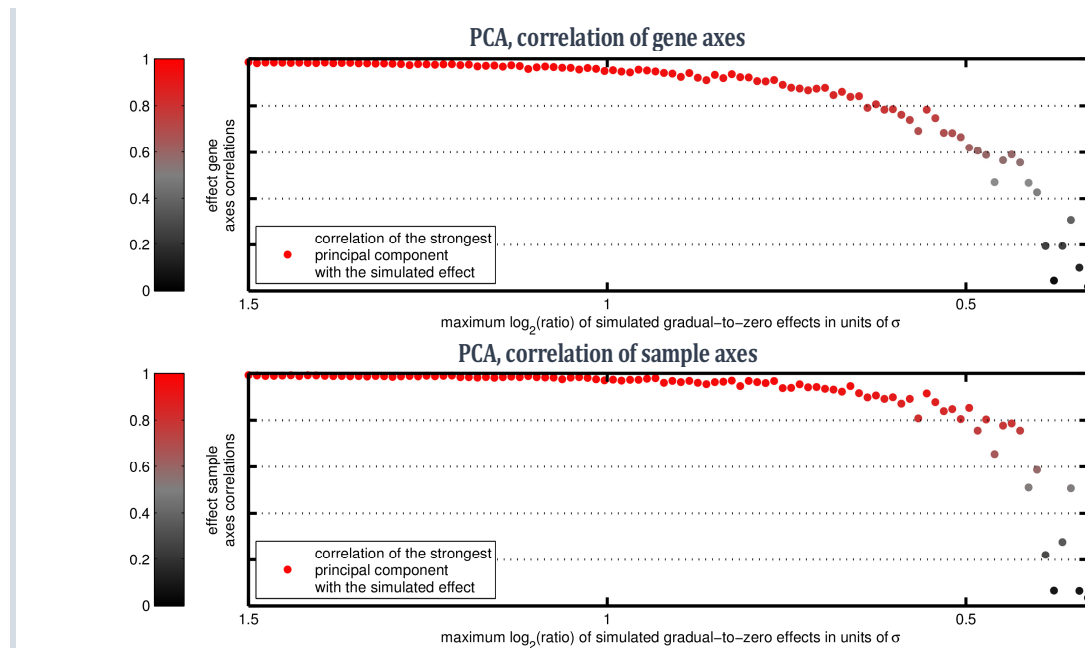
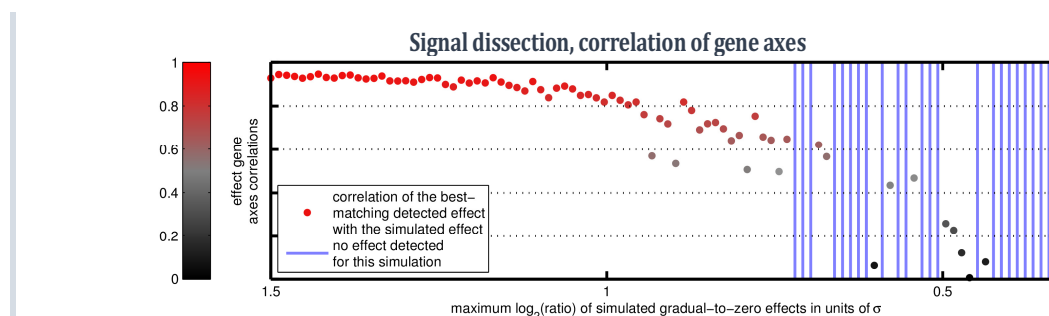


Figure II.6.4.1.b) Detection limit with respect to signal strength, effect pattern #6, PCA results

The same test has been repeated for effects shaped like pattern #3 in the versatility test, i.e. a two-sided signal that starts at a nonzero offset (cf. Figure II.6.2.1). As this effect only affects 50% of simulated samples, correlations to its effect axes decrease faster with the simulated signal strength as for the pattern tested above. Hence, also the onset of missing detections is reached earlier at approximately  $0.75\sigma$ . As before, PCA shows higher correlations and the correlation breakdown is prolonged to weaker signal strengths. Again, correlation differences become negligible for stronger signals.



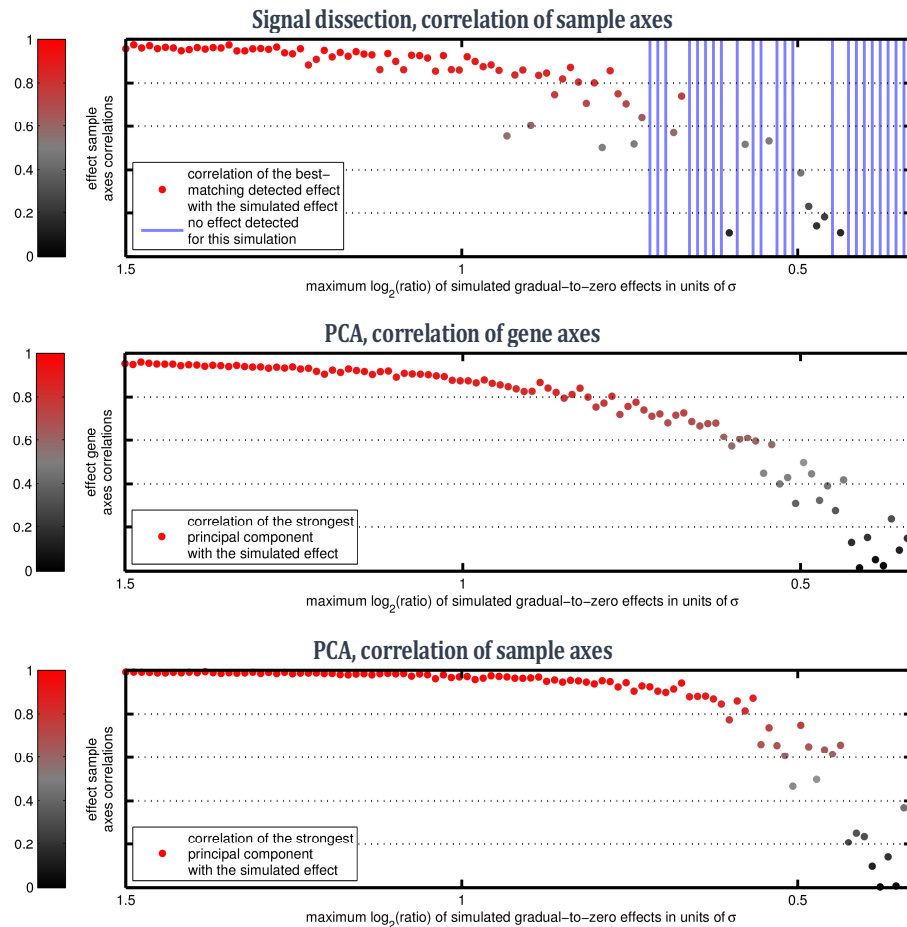


Figure II.6.4.1.c) Detection limit with respect to signal strength, effect pattern #3, signal dissection and PCA results

#### II.6.4.2

### Many noise genes

Besides finding true positives and preventing false positives in the limit of weak signals, the same goals need to be targeted in the limit of many noise genes.

Discovering small true effects that exist only in few of all measured genes is demanding. For example, a small effect with a moderate signal strength for e.g. 10 genes may be robustly detectable without inviting any false positives in a measured signal matrix of 1000 genes. If the same effect was embedded in a signal matrix of 20000 measured genes, it may still be detectable using constant significance thresholds, but at the cost of also detecting many false positives due to multiple hypothesis testing and due to more possibilities for 10 genes to be correlated by chance. To correct for that, significance thresholds are defined adaptively to the signal size using Bonferroni corrections<sup>(cf. II.3.1.8)</sup>.

To investigate this limit and to test these adaptive significance thresholds, the versatility scenario with 7 effects (that is simulated for 1000 genes and 100 samples as before) is now embedded in a larger signal matrix. More precisely, noise genes are appended, i.e. genes that are affected by the same global noise level, but that are not regulated by any of the simulated effects. Starting with the unchanged versatility scenario, noise genes are added in steps of 2000 up to a signal size with 55000 simulated genes.

Results are depicted for gene axes correlations<sup>(cf. Figure II.6.4.2.a)</sup>. Sample axes correlation are comparable and are available in `Method Validation\versatility7.overNoiseGenes (nP=100)`. All detected effects by signal

dissection are depicted, including false positives. As PCA has no concept of noise estimation, its seven top correlated principal components for the respective simulation are selected and depicted for comparison.

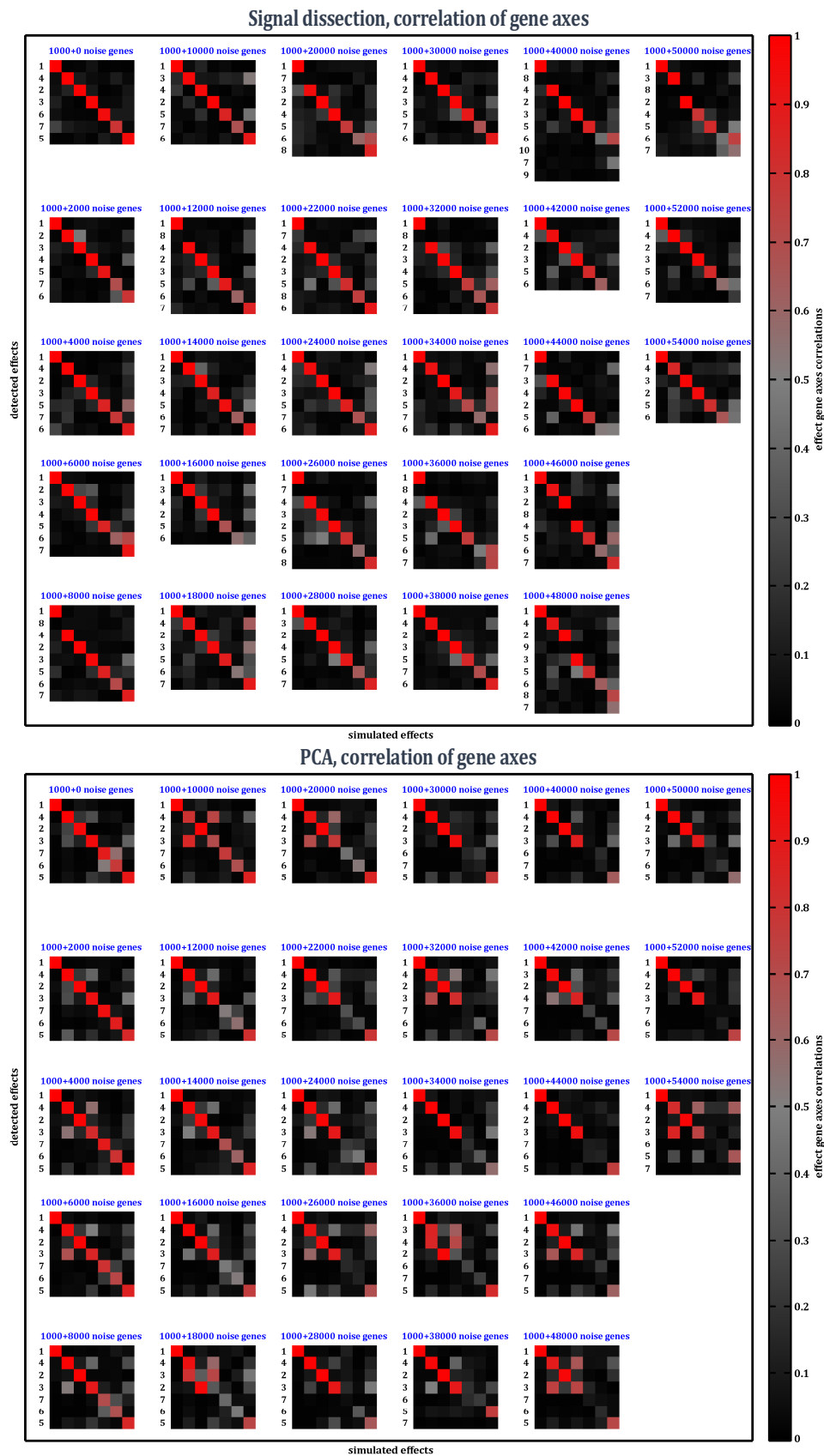


Figure II.6.4.2.a) Versatility test with 7 effects, 1000...5000 genes, correlations of gene axes and comparison of signal dissection with PCA

### ■ Comparison with PCA for each simulated pattern

For zero added noise genes, signal dissection and PCA have already been systematically compared<sup>(cf. II.6.2.4 and II.6.2.5)</sup>. Results are similar for relatively few added noise genes, i.e. both signal dissection and PCA discover all seven simulated effects with high correlation in most simulations. Differences in the context of many noise genes are discussed below for each simulated pattern<sup>(cf. Figure II.6.2.1)</sup>.

**Pattern #1:** As expected and due to its size, pattern #1 is detected reliably in all simulations with high correlation by both methods.

**Pattern #2:** The strong but small pattern #2 is reliably detected by signal dissection. Likewise, PCA always returns a principal component with high correlation to this effect's gene axis. But often this principal component is not specific. This can be seen by its simultaneous correlation to either pattern #3 or pattern #4<sup>(two red pixels in column 2)</sup>.

**Pattern #3:** Like pattern #2, this effect is reliably and specifically detected by signal dissection, while top-correlated principle components often mix it with pattern #2 or pattern #4. Hence, underlying laws of gene regulation are hard to interpret, again.

**Pattern #4:** (Like patterns #2 und #3.)

**Pattern #5:** For the narrow pattern #5 the difference between the two methods is striking. While signal dissection discovers this effect reliably even for highest amounts of noise genes, PCA returns unreliable correlations from 12000 noise genes onwards and does not detect this effect at all for 26000 noise genes onwards. This effect is only comprised of 19 correlated and 1 anti-correlated genes. Probably this effect is too small to detect it without adaptive effect focusing in a sea of noise genes.

**Pattern #6:** While pattern #5 was hard to detect in context of many noise genes due to its small size, pattern #6 is hard to detect due to its weak signal strength. For 10000 noise genes and lower, both methods still detect it reliably, albeit with weakening correlation. While signal dissection detects it in approximately every second simulation even for highest numbers of simulated noise genes, PCA fails to discover it from approximately 22000 noise genes onwards.

**Pattern #7:** From 16000 added noise genes onwards, signal dissection does not discover the narrow-shaped effect pattern #7 reliably. Either it is not detected at all or with low gene axis correlation to the simulated effect and with unspecific correlations to other effects. Likewise, PCA cannot always discover this effect and often discovers it with only moderate correlation<sup>(light red or gray in column 7)</sup>, but tends to be more specific for this pattern (i.e. corresponding principle components have less correlation to other effects).

### ■ Dilution of top genes by noise genes

Lower correlations between gene axes correspond to higher dilution of top genes by noise genes. This dilution can be observed especially for the weak-signal pattern #6 and for the narrow pattern #7. Increasing dilution by false positives generally causes increasing difficulty for effect interpretation. Unfortunately, these false positives top genes cannot be circumvented, as they have *signal properties equal to true positives*.

Especially for pattern #7 that only regulates 5/100 simulated samples, it is relatively easy to simulate noise genes that are correlated by chance. These false positives with high correlation to true positives naturally become more frequent if more noise genes are simulated. An exemplary discovery of pattern #7<sup>(cf. Figure II.6.4.2.b)</sup> illustrates this problem. While top ranks at the bottom are occupied by true positives<sup>(non-black signal in panel b)</sup>, all



### Specific discovery of small effects despite many present noise genes

The most striking difference between signal dissection and PCA in the current scenario has been observed for the small pattern #5, as explained above. PCA is unable to detect it in the presence of many noise genes, probably because PCA lacks a concept for adaptive effect focusing. As *small biological effects measured by large whole genome microarrays are typical, this is a major drawback of PCA.*

An exemplary discovery (cf. Figure II.6.4.2.c) of pattern #5 demonstrates the accuracy of signal dissection and effect focusing in context of many noise genes: Only 19 correlated and one anti-correlated true positive genes exist in this simulated signal of 55000 genes. This signal size is typical when measuring gene expressions on probeset level (cf. III.1.1.1). While again several false positives (black in panel b) are correlated to the empirically derived effect axes, all 20 true positives are discovered (see gene labels). Furthermore, they occupy top ranks in this effect without interruption by any false positives. Especially the fact that the solitary anti-correlated true positive gene has higher effect strength than all false positives was unexpected, as pattern #5 has only moderate signal strength and only affects 50% of all simulated samples. Additionally, this gene was overlapped by patterns #1 and #6 in the depicted simulation. Hence, its obvious signal correlation in the center panel is barely visible in the original signal (panel a).

### A false positive example

An *exemplary false positive discovery* (cf. Figure II.6.4.2.d) shows significant signal strength as well as significant correlations. This is expected, as it passed effect qualification (cf. II.3.1.8). In fact, its original signal in the empirically discovered eigenorder looks like a solid true positive effect. Typically however, such false positives are relatively small, either with respect to the number of involved genes or with respect to the number of involved samples.

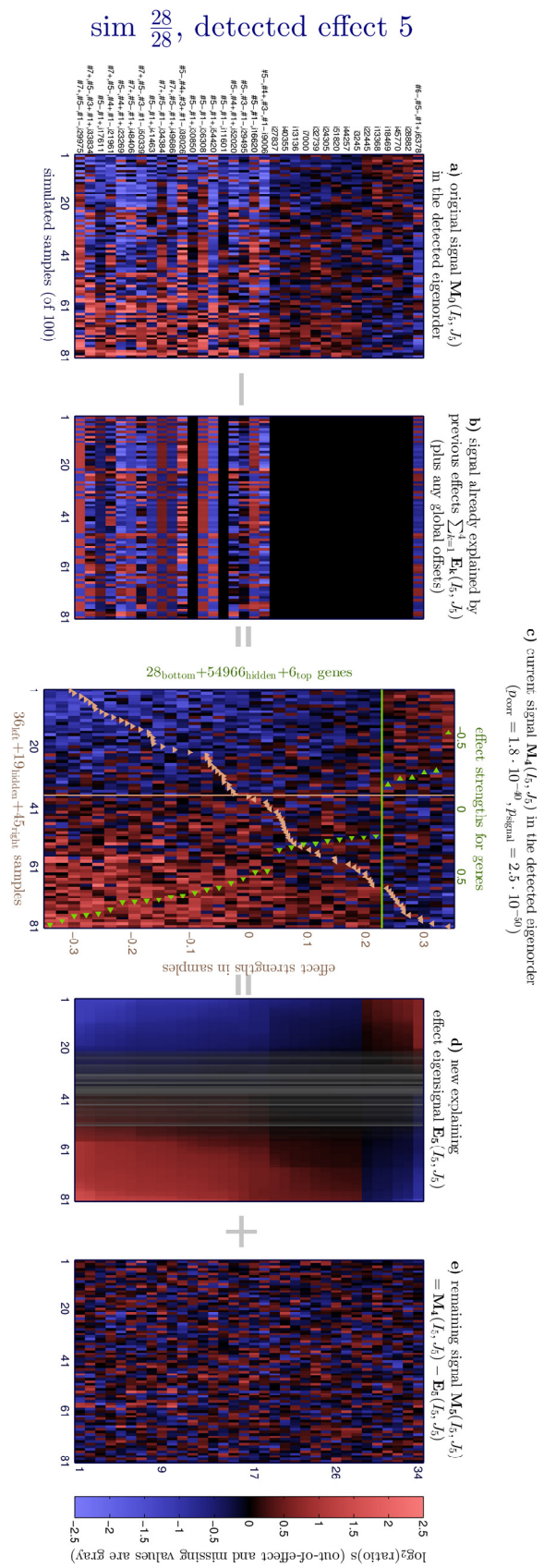


Figure II.6.4.2.c) Versatility test with 7 effects embedded in a noise signal with 55000 genes, detection and dissection of the small pattern #5

However, “small” has to be understood relative to the size of the input signal. Hence, very large input signals may produce larger and more frequent false positive effects.

This is also the reason, why effect size cutoffs used for effect validation have been defined adaptively to the size of the input signal<sup>(cf. II.3.1.8)</sup>.

Naturally, it is increasingly harder for such strong and correlated signals to emerge from pure noise for many genes *and* many samples. Hence, they could be easily excluded by sufficiently high effect size qualification thresholds. Still, this is a tradeoff, as too high thresholds would also exclude small true positives.

sim  $\frac{26}{28}$ , detected effect 8

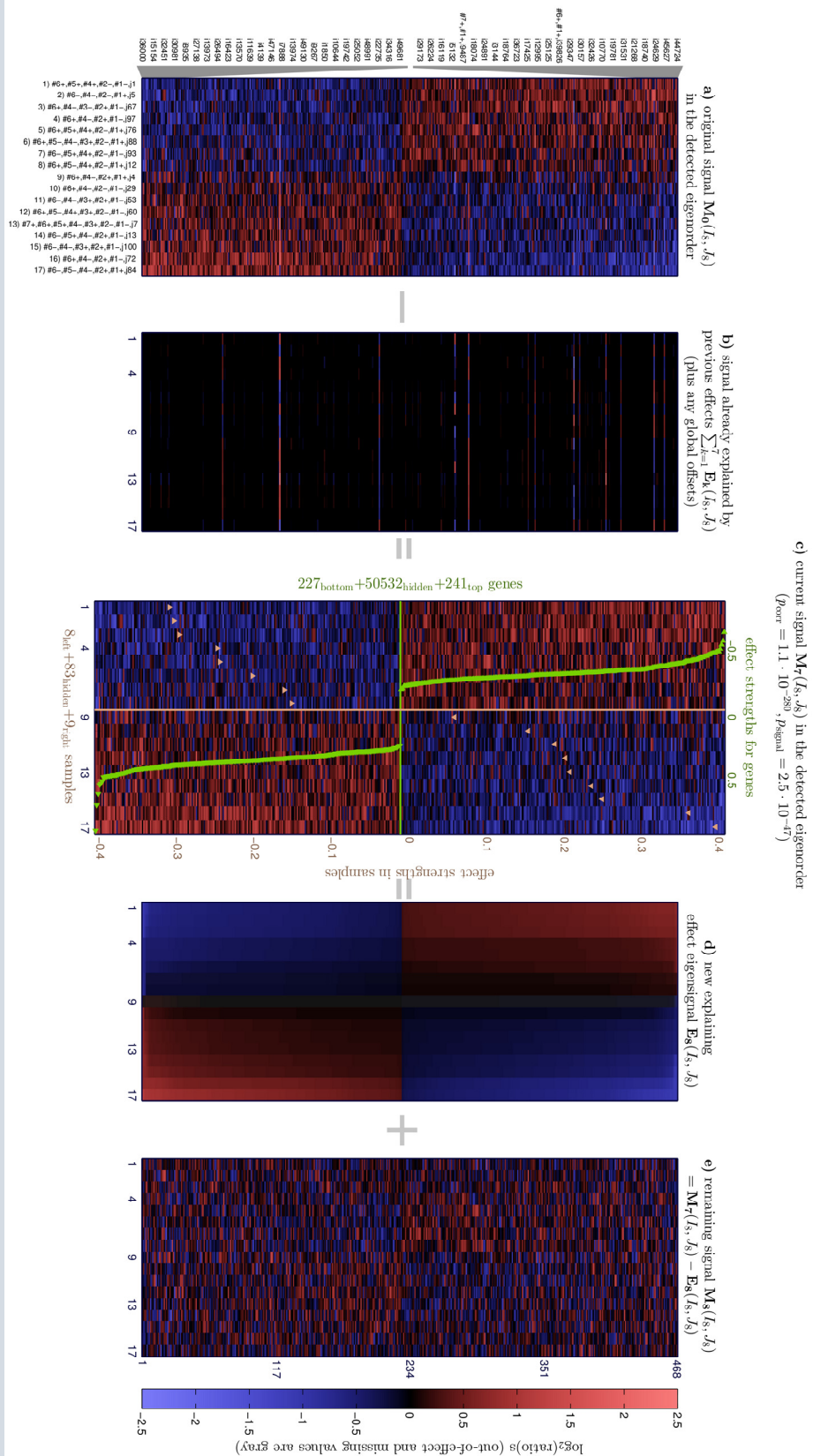


Figure II.6.4.2.d) Example of a false positive discovery (from a versatility test embedded in a noise signal with 51000 genes)

Related to the problem of false positives due to many noise genes<sup>(cf. previous section)</sup>, false positives may also emerge in the limit of few samples. In this limit, genes have a higher chance to be correlated to each other by pure chance. This is similar to false positive genes for narrow effect foci within larger signals<sup>(cf. Figure II.6.4.2.b)</sup>.

To systematically test the detection performance over the number of available samples, the versatility scenario<sup>(cf. Figure II.6.2.1)</sup> has been simulated for 20 to 2000 samples (and again 1000 genes). Detection results and the comparison to PCA results are shown below<sup>(cf. Figure II.6.4.3.a)</sup>. As expected, both methods return more reliable and more robust results for high sample counts. Consistent with previous simulations<sup>(cf. Table II.6.2.5)</sup>, the weak signal pattern #6 was missed by signal dissection for simulations with 766 and 1363 samples, but not so by PCA.

For only 20 samples, the small or weak patterns #5, #6 and #7 are no longer detected by signal dissection. It is possible to detect them with lower significance thresholds for effect qualification, but this would invite unwanted false positives. The simulation for 52 samples demonstrates that the chosen configuration<sup>(cf. II.3.1.8)</sup> is already near the onset of many false positives and should not be lowered much more. PCA still computes principal components that are weakly correlated to patterns #5 respectively #7 for the same 20 samples signal. However, as PCA does neither determine the noise level nor the number of effects in the signal, depicted seven top principal components are not comparable in terms of false positives and detection thresholds.

To illustrate the problem of false positive genes in detected effects for in the few samples limit, the detection iteration for pattern #4 is depicted<sup>(cf. Figure II.6.4.3.b, page 120)</sup>. While top-correlated genes at the bottom are true positives, all anti-correlated genes at the top and several correlated genes in the middle are false positives<sup>(cf. gene labels)</sup>. Results are comparable to false positive genes in the many noise genes limit<sup>(e.g. Figure II.6.4.2.c)</sup>.



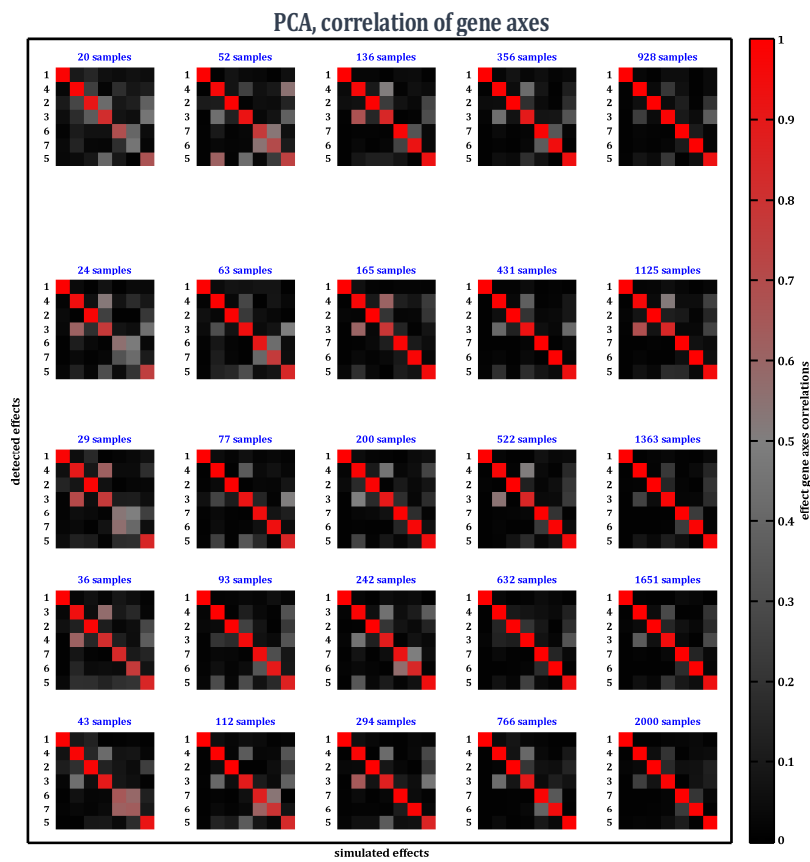
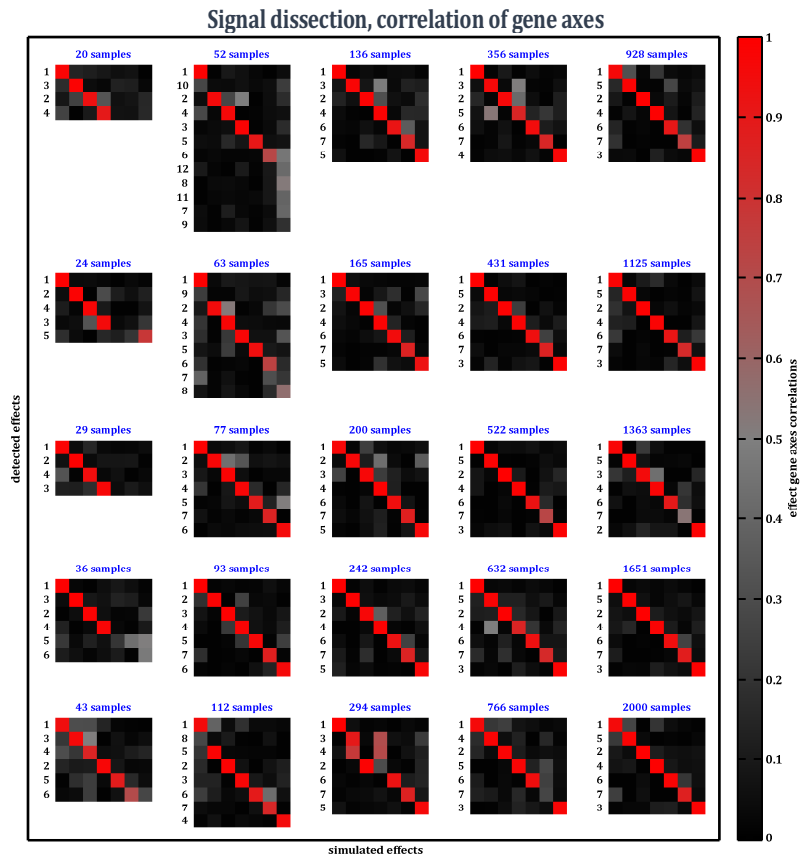


Figure II.6.4.3.a) Versatility test with 7 effects, 20...2000 samples, correlations of gene axes and comparison of signal dissection with PCA

sim  $\frac{1}{25}$ , detected eigenorder 4

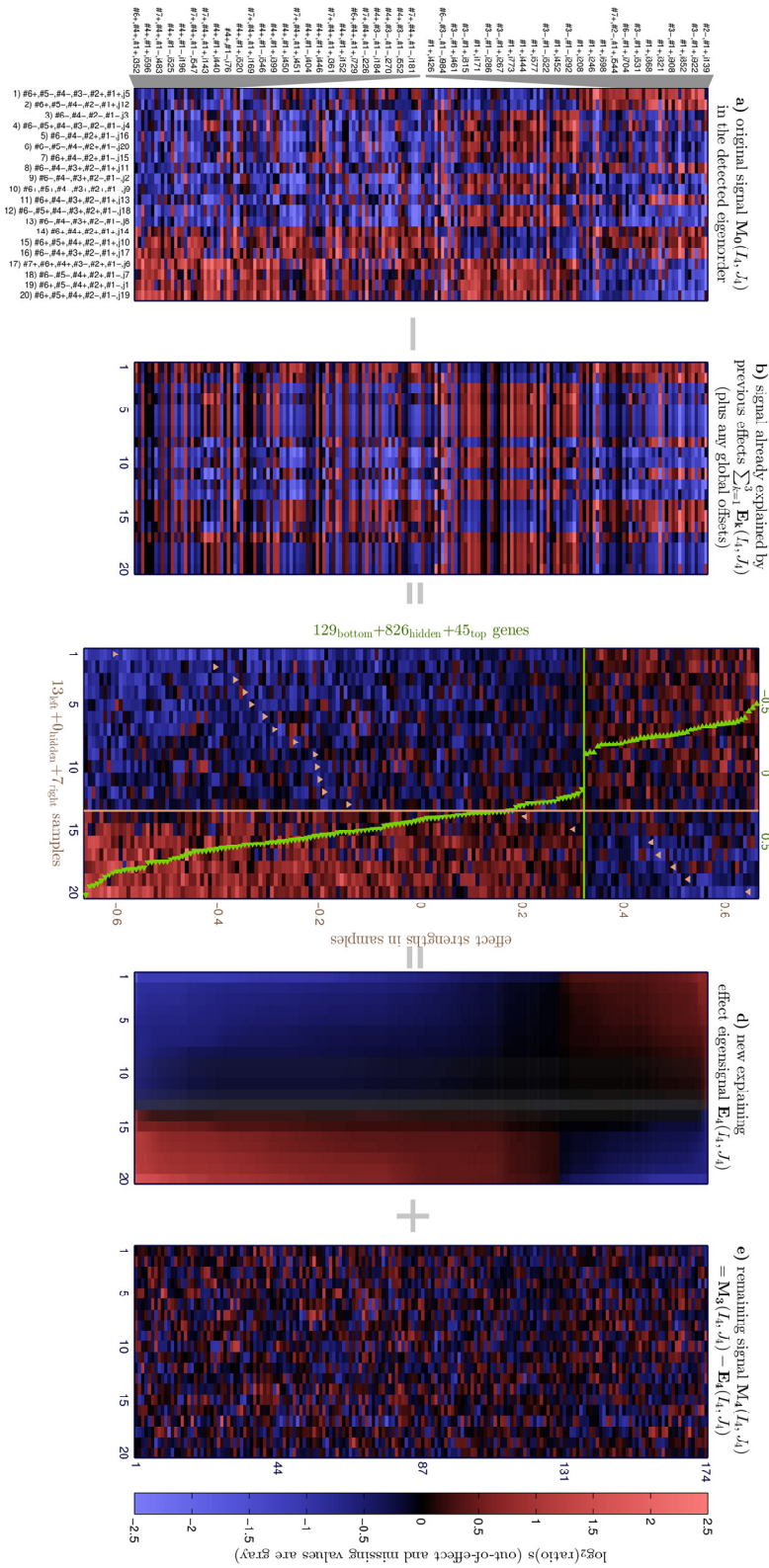


Figure II.6.4.3.b) Versatility test with 7 effects simulated for only 20 samples, detection and dissection of pattern #4 with several false positive genes

An important consideration for practical applications is often that measurement devices do not yield reliable results for some probes. For example, 5% of all measured sequences on a microarray may be not be readable or fail in QC checks. In this case, still 95% are available *per* sample, but only the intersection, i.e. typically considerably less information is be available with robust readouts for *all* samples. Hence, analysis methods that can handle missing values are needed to make use of all non-missing information.

PCA does not support missing values and hence is not compared with signal dissection in this context. In general, an external imputation method for missing values might be utilized to make signals with missing values compatible to PCA. However, global imputation methods (e.g. replacing missing values by zero) might introduce a bias to results.

Correlations underlying signal dissection are weighted<sup>(cf. II.2.3.1)</sup> and thus *natively support missing values*: they are treated identically to non-missing values with zero weight. For example, expressions from genes outside of the effect focus have zero weight and are commonplace for signal dissection already. Furthermore, bimonotonic regression<sup>(cf. II.4.1.3)</sup> is weighted as well and thus supports missing values, too. This regression also effectively *imputes missing values* from neighboring values *in the empirical effect eigenorder*<sup>(cf. II.4.1.2)</sup>. Hence, this *imputation is dependent on the local effect context* and thus does not introduce any bias. In this way, imputed values are effectively inferred from correlations in the non-missing signal. In principle, this might also be used to *predict missing signals*.

To quantify detection performance over missing values, the versatility scenario<sup>(cf. Figure II.6.2.1)</sup> is simulated in its usual size (i.e. for 1000 genes and 100 samples), but simulated (gene, sample) pixels are deleted in steps of 5% from 0% to 100%. Detection results show<sup>(cf. Figure II.6.4.4.a)</sup> that the strong effect pattern #1 is reliably detected even up to 90% missing values. For smaller effects, not enough information is left at such high rates of missing values and thus they cannot be detected. Unexpectedly, all effect patterns are detected for up to 35% of missing values. As expected, the weak signal pattern #6 is lost to missing values first (from a rate of 40% onwards). Effect patterns #2, #3 and #4 that imitate typical biological gene expression effects are *detected for up to 80% of missing values*, again more than expected. The narrow pattern #7 is detected for up to 75% missing values and the small pattern #5 up to approximately 55%.

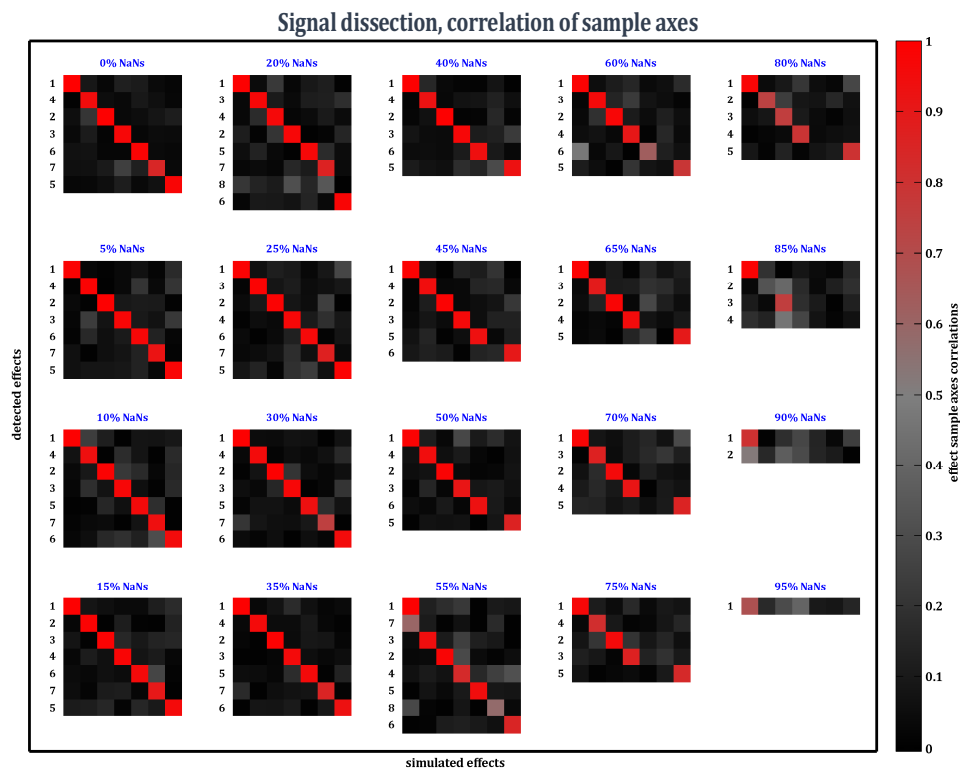
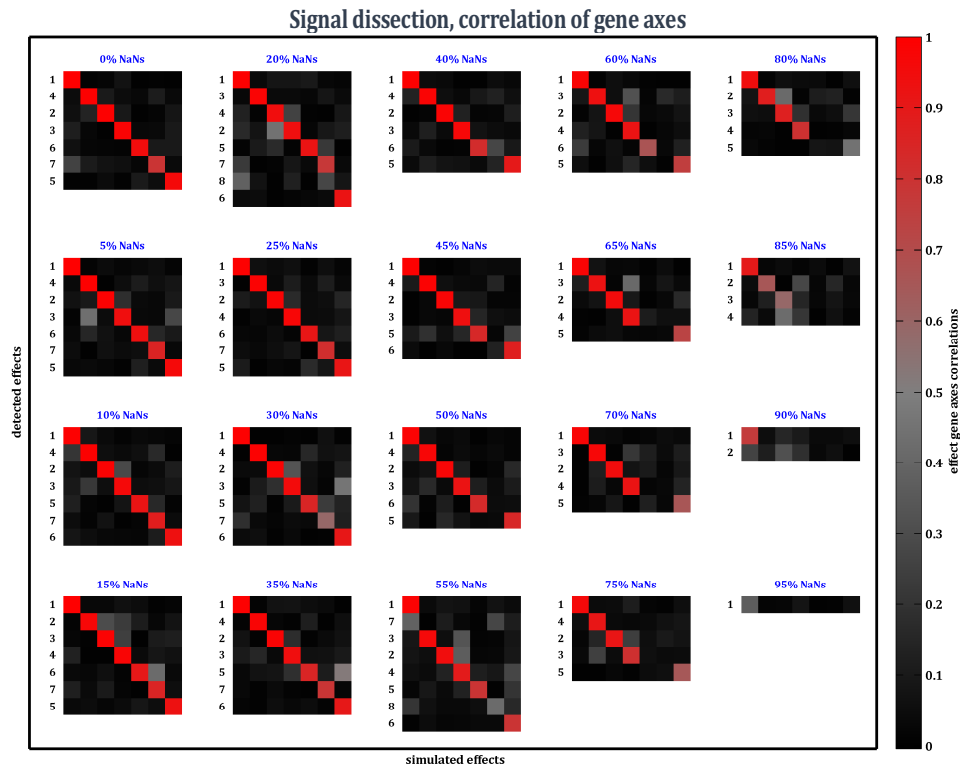


Figure II.6.4.4.a) Versatility test with 7 effects, 0%...100% missing values, correlations of detected gene and sample axes with simulated ones

To demonstrate the imputation of effect signals, a detection and dissection of pattern #3 in context of 80% missing values is presented next (Figure II.6.4.4.b). Despite only relatively few remaining information, the effect is reliably detected with high correlation the simulated law of gene regulation (see the block for 80% NaNs in Figure II.6.4.4.a). Consistently, top and bottom genes in the definition plot are all labelled by “#3+” respectively “#3-”, indicating that they are regulated by the simulated pattern #3. Despite the high rate of missing values, still no dilution by

false positive genes is visible for top correlated ranks. However, the eigensignal (panel d) does not have sharp borders towards the center, as would be normal for this pattern if no values were missing (cf. Figure II.6.2.2.d).

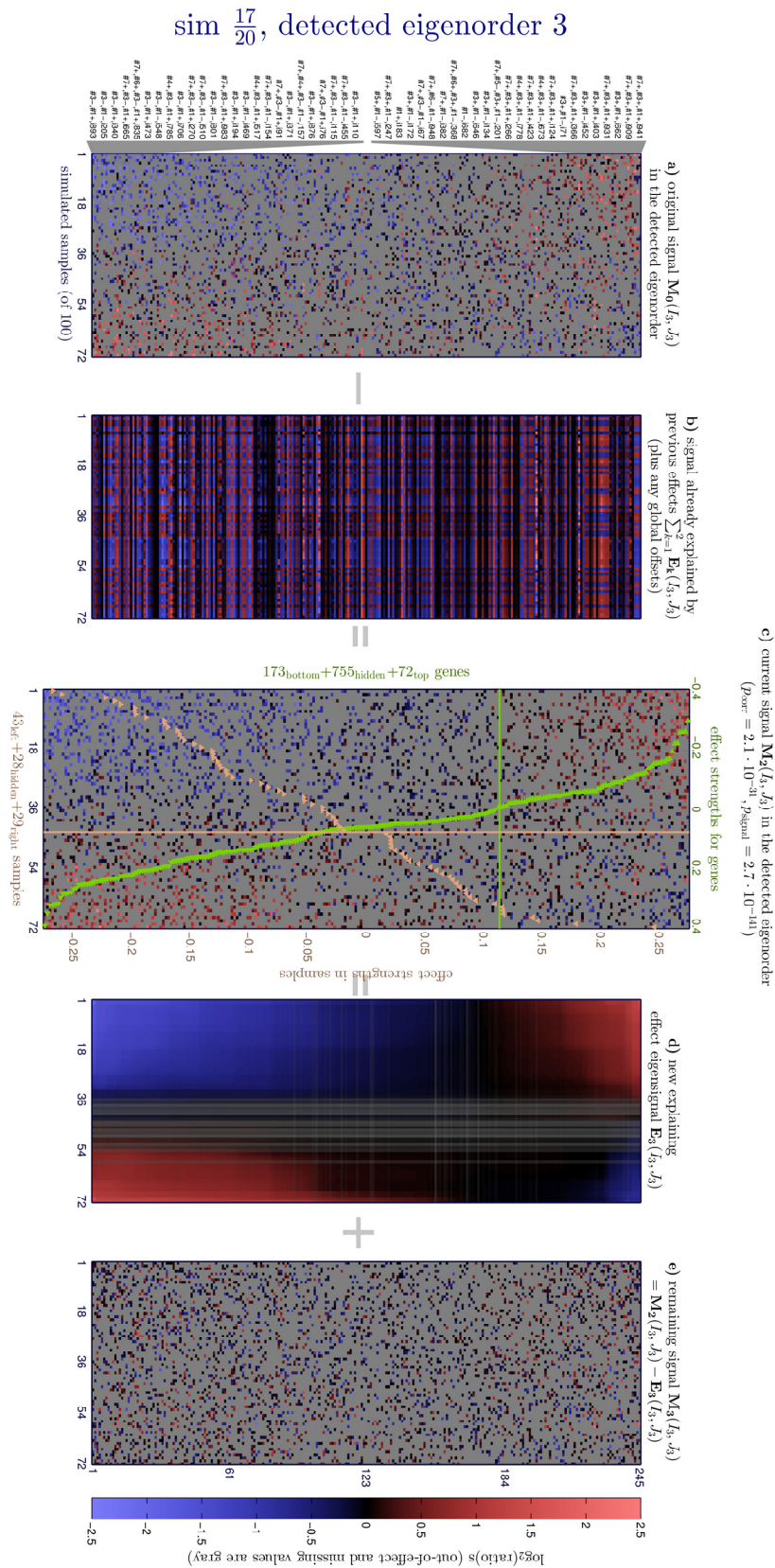


Figure II.6.4.4.b) Detection of effect pattern #3 for the versatility test with 7 effects and 80% missing values

Comparison of all simulated effects(cf. Figure II.6.4.4.c) with discovered and recovered eigensignals(cf. Figure II.6.4.4.d) shows that imputed values can predict many missing value correctly. However and as expected due to 80% missing values and due to simulated measurement noise, this signal reconstruction is not perfect. Still, it is not as worse relative to the same test without missing values(cf. Figure II.6.2.3.c) as could have been expected for 80% missing values. Additionally, neither global imputation methods nor imputation methods using a neighborhood concept (based on gene and sample reference orders) could have predicted such consistent expressions with simulated effects, as these methods do not have any knowledge of actual effects or correlations in the signal.

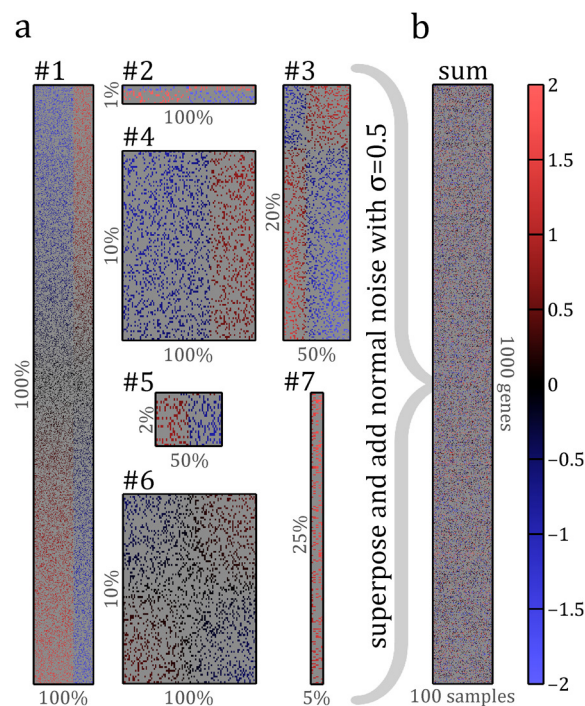


Figure II.6.4.4.c) Versatility test with 80% missing values

Simulation of a versatility scenario with 7 effects for 1000 genes and 100 samples. (a) Seven effects of depicted size and signal shape are simulated for randomly selected genes and samples as before (cf. II.6.2.1). (b) Superposed effects are depicted in a common reference order, including simulated noise ( $\sigma \equiv 0.5$ ). 80% of all (gene, sample) pixels were randomly selected and their signal was deleted.

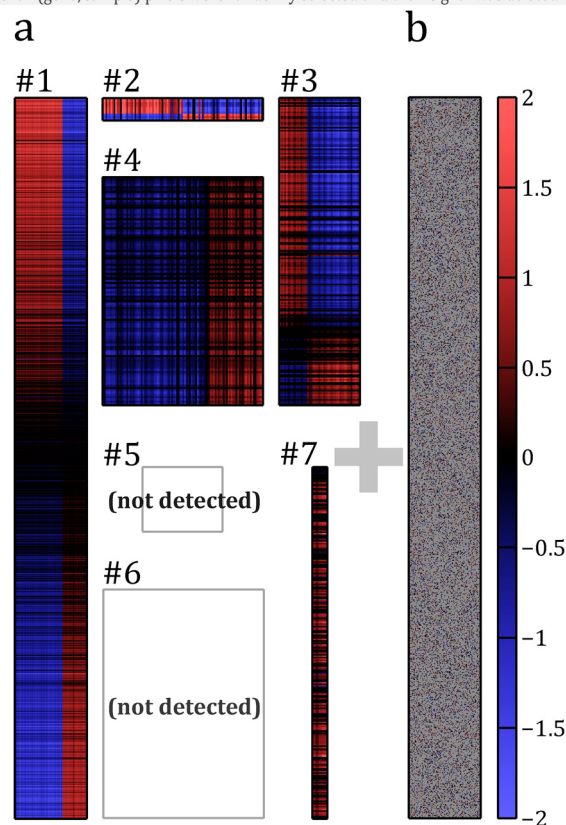


Figure II.6.4.4.d) Imputed eigensignals (versatility test, 80% missing values)

(a) 5/7 simulated effects are detected, despite 80% missing values. Originally simulated effect signals are imputed for missing values. The detected eigensignal for pattern #3 in its empirically determined eigenorder is depicted in Figure II.6.4.4.b, for example. (b) The remaining signal still contains the simulated signals (as depicted in Figure II.6.4.4.c above) for the small pattern #5 and for the weak-signal effect #6.

# Chapter III

---

## Dissecting DLBCL Gene Expressions

*Based on human whole genome profiling, gene expressions of more than 1000 patient samples of Diffuse Large B-Cell Lymphoma are dissected for four independent patient cohorts. Cross-cohort validation yields 135 GEP effects. Each effect can distinguish between patients by significantly differentially expressed genes. To illuminate their biological nature, all are systematically associated with clinical and genomic knowledge.*

*The cell of origin<sup>(COO)</sup> effect that can distinguish between known DLBCL subtypes is rediscovered and redefined in a filtered form. A novel and genetically distinct effect is discovered that can predict DLBCL patient outcome more consistently compared to the COO effect. Multivariate survival analyses reveal novel hierarchical survival dependencies. Additionally, a combination of five GEP effects can predict strong and significant survival differences even within known subtypes and within clinical risk classes by international prognostic index.*





## III.1 Application to DLBCL

---

*Gene expression microarray measurements for samples from four DLBCL patient cohorts have been obtained via NCBI GEO<sup>[94]</sup>. Signal dissection<sup>(Chapter II)</sup> is independently applied to each cohort. Single cohorts may contain lab effects, for example caused by different experimental labelling protocols. Only effects representing molecular mechanisms in cells are of interest here and, therefore, effects are validated across cohorts to filter out non-biological effects.*

*Gene axes of validated effects that were discovered in multiple cohorts are then combined to form consensus gene axes that represent detected effects in their most general and cohort-independent form. Application of these consensus gene axes yields corresponding consensus sample axes that arrange patients according to an effect's impact on them. This could also be utilized to classify patients in future DLBCL cohorts that were not used for effect detection.*

*Backed by successful method validation<sup>(II.6)</sup>, resulting 135 validated effects are expected to represent specific interactions in measured cells that are biologically comprehensible. Comprehensive biostatistical analyses have been computed for each validated effect (available via [DLBCL Master Table 2015, main overview.xlsx](#)). In subsequent subchapters, these effects are systematically associated with patient survival<sup>(III.2)</sup> and top effects are biostatistically evaluated in detail<sup>(III.3)</sup>.*

### III.1.1 Detection in Single Patient Cohorts

---

The dissection algorithm presented in Chapter II is applied to gene expression signals from four cohorts of DLBCL patients. Identical detection settings<sup>(e.g. qualification thresholds; cf. II.3.1.8)</sup> that were used for method validation<sup>(II.6)</sup> are now applied for dissection of these real world signals.

#### III.1.1.1 Available gene expressions cohorts

---

Gene expression microarrays for the following four cohorts and samples have been obtained via NCBI GEO<sup>[94]</sup>: 181 samples from GSE10846.CHOP<sup>[5]</sup>, 166 from GSE4475.CHOP<sup>[95]</sup>, 233 from GSE10846.R-CHOP<sup>[5]</sup> and 498 from GSE31312.R-CHOP<sup>[29]</sup>. (With clinical analyses and validation in mind, GSE10846 has been split in two sub cohorts by the applied chemotherapy<sup>(cf. I.1.2.1)</sup>.)

All GEP measurements took place before start of therapy. GSE31312.R-CHOP is based on formalin-fixed, paraffin-embedded tissue (FFPE), while all other cohorts measure fresh frozen cell material from patients. Generally, gene expressions of FFPE material tend to be noisier because of mRNA degrading effects<sup>[96]</sup>. However, with clinical applications in mind, discovered effects should be robustly identifiable for FFPE based samples as well, as no freezing of patient cell material is available for routine clinical work due to cost and infrastructure reasons. Indeed, all relevant GEP effects discovered by this work exist and validate in this FFPE based cohort. One particularly important survival effect has even been discovered based on this cohort. This also indirectly confirms that HighPure Paraffin RNA Extraction Kits<sup>(Roche Diagnostics, Basle, Switzerland)</sup> utilized for measuring GSE31312 have worked reliably.

Cohorts GSE10846 and GSE31312 were measured with GeneChip® Human Genome U133 Plus 2.0 microarrays (Affymetrix Inc., Santa Clara, California, United States) and GSE4475 with former GeneChip®

Human Genome U133A microarrays. These microarrays are based on probesets that quantify mRNA concentration for genomic clusters. U133 Plus 2.0 arrays contain 54675 probesets for coding and non-coding mRNA sequences, U133A arrays contain a subset of 22283 probesets thereof. Annotations of probeset sequences with 21686 unique human genes are provided by the manufacturer (NetAffx™<sup>[97]</sup> v33). In general, analyzing gene expressions on the probeset level is more specific than analyzing on aggregated gene level. Hence, all dissections are computed for probeset level gene expressions. Sequences of single probes underlying a probeset are available from the manufacturer (they might help to construct specific shRNAs or cDNAs for further experimental investigations of probesets that have been identified by discovered effects and that are of biological interest).


Except for the two sub cohorts of GSE10846, all cohorts were studied and measured by mutually independent teams and labs. Even sub-cohorts in GSE10846 are based on different experimental labelling protocols. Hence, validation of discovered effects across cohorts can be utilized to sort out lab effects and to exclusively keep effects of biological origin.

### III.1.1.2 Dissection overview for single cohorts

---

Raw signal dissection has discovered 221 (GSE10846.CHOP), 82 (GSE4475.CHOP), 161 (GSE10846.R-CHOP) and 105 (GSE31312.R-CHOP) significant gene expression effects before the respective remaining signal has been considered noise<sup>(II.4.3)</sup>.

When compared to GSE10846.R-CHOP, lower numbers of detected effects were expected for the FFPE based cohort (due to the higher noise level) as well as for GSE4475.CHOP (due to the relatively small cohort size). It is surprising that GSE10846.CHOP yields more effects. However, several of these effects are very narrow (with respect to their number of top genes or top samples) and thus are probably lab effects that will be identified as such and sorted out by cross-cohort validation next.

All discovered and dissected effects for raw signals are available in graphical and tabular form in cohort sub folders in  *A=Detection*.

Ideally, GEP effects are independently discovered by signal dissection in all four cohorts. Only very strong and genetically unambiguous effects reach this quad-discovery. To find them, gene axes of all detected effects in all cohorts are systematically compared by correlation. Tuples of highly correlated effects from different cohorts are identified.

Most true positive effects are not independently discovered in all four cohorts, probably because of alternate dissections. However, even an effect that has been discovered in just two cohorts can already be considered as strongly validated, because it was *unsupervisedly* rediscovered in an independent patient cohort.

A weaker yet still sufficient form of GEP effect validation is supervised validation, i.e. an effect has been discovered in just one cohort, but its supervised application to initial signals from other cohorts proves its existence there, too. Alternatively, GEP effects discovered only in single cohorts can be considered validated if they are significantly associated with non-GEP covariates of the same cohort, like patient survival.

Generally, the aim of validation is to extract and validate as many as possible discovered effects and only filter those out that are cohort-specific effects and thus disease-unspecific lab effects.

### III.1.2.1 Comparison between two cohorts by correlation of effect gene axes

Let  $C_1$  and  $C_2$  denote two dissected cohorts and let  $k_1, k_2$  denote indices of detection iterations. Let  $|b_{C_1, k_1}^g\rangle$  and  $|b_{C_2, k_2}^g\rangle$  denote the converged gene axes of corresponding detected effects. Let  $|r_{C_1, k_1}^g\rangle$  and  $|r_{C_2, k_2}^g\rangle$  denote the final gene correlations with the effect's converged sample axis for the respective cohort. Finally, let  $|w_{C_1, k_1}^g\rangle$  and  $|w_{C_2, k_2}^g\rangle$  denote the effect gene weights in the final effect focus. All these vectors are results from effect axes convergence in II.3.2. (The index  $\hat{l}$  that indicated the convergence iteration in II.3.2 is suppressed here for better readability.)

As both the signal strength in effect direction and the consistency of regulations (i.e. correlations) should be similar for discovered effects that originate from the same biological effect, I compare the following product gene scores:

$$||b^g|.r^g\rangle \equiv \sum_{i=1}^m |\langle e_i^g | b^g \rangle| |\langle e_i^g | r^g \rangle| |e_i^g\rangle$$

Eqn. III.1.2.1.a) Gene scores (combining signal strength and correlation information)

I.e.  $||b_{C_1, k_1}^g|.r_{C_1, k_1}^g\rangle$  is compared with  $||b_{C_2, k_2}^g|.r_{C_2, k_2}^g\rangle$ . Notably, these gene scores indirectly contain information from all samples in the respective cohorts, as gene correlations are computed to effect sample axes.

For the comparison of two discovered effects, all genes are relevant that are regulated by the potentially common biological effect. Additionally, both effects should extend to the same set of genes in both cohorts, if they represent the same biological interactions. Hence, the comparison should be computed for the outer effect focus  $|w_{(C_1, k_1; C_2, k_2)}^g\rangle$  defined as the maximum of respective gene weights. Then, the actual comparison score  $r_{(C_1, k_1; C_2, k_2)}$  is again computed by uncentered weighted correlations<sup>(cf. II.2.3.1)</sup>:

$$r_{(C_1, k_1; C_2, k_2)} \equiv \left[ |b_{C_1, k_1}^g|.r_{C_1, k_1}^g \right] \left[ |b_{C_2, k_2}^g|.r_{C_2, k_2}^g \right] |w_{(C_1, k_1; C_2, k_2)}^g\rangle \text{ with } |w_{(C_1, k_1; C_2, k_2)}^g\rangle \equiv \sum_{i=1}^m \max \left( \langle e_i^g | w_{C_1, k_1}^g \rangle, \langle e_i^g | w_{C_2, k_2}^g \rangle \right) |e_i^g\rangle$$

Eqn. III.1.2.1.b) Comparison of effects from two cohorts by correlation

For cross-cohort validation, all pairwise correlations are computed (for all pairs of cohorts and for all discovered effects in these cohorts). Additionally, corresponding  $p_{(C_1, k_1; C_2, k_2)}$ -values for these correlations are computed as before<sup>(cf. II.5.2.1)</sup>.

These computations presuppose that the same gene space is shared by all cohorts. In practice, this is always attainable by aggregating microarray results on gene level. Genes that were not measured by a particular microarray are assigned zero weights. For current cohorts<sup>(cf. III.1.1.1)</sup>, the higher-resolution probeset space of the U133 Plus 2.0 microarray platform with  $m = 54675$  dimensions has been selected. As U133A microarray utilized for GSE4475.CHOP contain a direct subset of these probesets, it was possible to join them by direct probeset ID lookup.

### III.1.2.2 Validation by independent discovery in several cohorts

Four cohorts are available and in the ideal case the same biological effect has been discovered in all cohorts. This corresponds to a 4-tuple of detection ranks  $(k_1, k_2, k_3, k_4)$  such that all pairwise correlations  $r_{(C_i, k_i; C_j, k_j)}$ <sup>(III.1.2.1)</sup> are strong and significant.

For every candidate tuple, pairwise correlations can be collected as a 4x4 matrix  $R_{(k_1, k_2, k_3, k_4)} \equiv \left( r_{(C_i, k_i; C_j, k_j)} \right)_{i, j=1..4}$ . Diagonals equal one (self-comparison) and thus for  $\hat{C}$  cohorts,  $\frac{\hat{C}^2 - \hat{C}}{2}$  cross-cohort correlations remain (the factor  $\frac{1}{2}$  is due to the symmetry of correlations).


Next, the count of significant pairwise correlations is computed for each tuple. More precisely, a four-dimensional count matrix  $\mathcal{C} \in \mathbb{N}^{\hat{k}_1 \times \hat{k}_2 \times \hat{k}_3 \times \hat{k}_4}$  (where  $\hat{k}_C$  denotes the final number of effects discovered for cohort  $C$ <sup>(cf. III.1.1.1)</sup>) is computed such that  $\mathcal{C}(k_1, k_2, k_3, k_4)$  equals the number of pairwise absolute correlations of effects  $(k_1, k_2, k_3, k_4)$  that are stronger than 0.5 and have a  $p$  value  $< 0.001$ . The same pairwise absolute correlations are geometrically averaged for each tuple, resulting in  $\mathcal{R} \in [0, 1]^{\hat{k}_1 \times \hat{k}_2 \times \hat{k}_3 \times \hat{k}_4}$ .

To estimate the count of *additional* cohorts, wherein an effect has been unsupervisedly *rediscovered*, pairwise correlation counts are scaled via  $\mathcal{C} \frac{\hat{C} - 1}{(\hat{C}^2 - \hat{C})/2}$ . For  $\hat{C} = 4$ , the maximum count is three independent rediscoveries and the minimum is zero (if the effect was only discovered in one cohort and no significantly correlated partner effect exists).

To determine a list of validated effect tuples, all tuples are sorted descending by their  $\mathcal{R}$  scores. The list is cut with the same threshold 0.5 that was applied to pairwise correlations above. In principle, an effect from one cohort may be significantly correlated to more than one effect in another cohort. Hence, this list may contain redundancies like  $(k_1, k_2, NaN, k_4)$  and  $(k_1, k'_2, NaN, k_4)$  (i.e. the effect was not detected in cohort three, and it was detected by two alternative effect axes in cohort two, while the same discovered effects are selected for cohorts one and four). Both tuples probably represent the same “outer” effect whose dissection into smaller and more coherent effects may be intrinsically ambiguous. I purposely *permit such redundancies* during validation, i.e. I permit discovered effects from one cohort to be included in more than one tuple. This enables finding different equally valid “views” on a true effect, i.e. views from partially correlated yet not identical consensus gene axes<sup>(cf. III.1.3)</sup>. At this point, it cannot be decided which view is “the best”, because all are statistically significant and have high  $\mathcal{R}$  scores. Later statistical associations e.g. with patient survival may reveal, which views are better, i.e. which are biologically more specific. Here, it is only important to keep all validated alternatives for further analysis.

In total, 133 effect tuples with effects from at least two different cohorts validate in this way. The **validation index**  $\nu$  is used to refer to these effects henceforth.

Another more demanding validation score can be defined based on the renormalized count of significant pairwise correlations times their geometric average:  $\mathcal{V} \equiv \mathcal{R} \cdot \mathcal{C} \frac{\hat{C} - 1}{(\hat{C}^2 - \hat{C})/2}$ . The theoretical maximum for analyzed DLBCL cohorts (i.e. for  $\hat{C} = 4$ ) is still three. The maximal observed validation score was  $\mathcal{V}(8, 124, 9, 17) = 2.84$ , followed by  $\mathcal{V}(3, 4, 6, 6) = 2.64$  for the gender-induced GEP effect<sup>(cf. III.3.2.1)</sup>. However, most real-world biological


effects are not so clearly defined like the gender effect, because their foldings are weaker. Hence, their signal is nearer to the noise level and thus correlations are lower, too. Depending on the chosen cutoff, many effects thus do not validate when requiring highly correlated independent rediscoveries. For example, if one would require a validation score of at least 1, only 37 would be validated (of the above 133 effect tuples with significantly correlated effects from at least two cohorts). These 37 effects are strongly validated, as they have at least one independent rediscovery that is correlated with 1 (which is only theoretically achievable) or have multiple independent rediscoveries with correlations  $0.5 \leq r \leq 1$  (which is always the case in practice). These validation scores are also available in the . While this strong form of unsupervised validation is nice to have, it should not be required. This would *exclude* many discovered effects that can be validated in a supervised form<sup>(cf. III.1.2.3)</sup> and that are actually biologically important, as is evidenced by later biostatistical analyses<sup>(cf. III.3)</sup>.

### III.1.2.3

## Supervised Validation

$|b_{C,k}^g\rangle$  denotes the final gene axis of an effect  $k$  that was discovered when dissecting gene expressions for cohort  $C$ . It suffices for validation to *supervisedly* check the existence of the same effect in other cohorts  $C'$ .

To this end, the same procedure is utilized as for classification of samples by consensus gene effects<sup>(details follow in III.1.4)</sup>. In brief,  $|b_{C,k}^g\rangle$  is accepted as final gene axis, its twin axis is computed, correlations are computed and the effect focus is updated. These vectors replace results from the search strategy<sup>(II.3)</sup> and are directly passed to regression and dissection<sup>(II.4)</sup>. If this leads to a significant<sup>(cf. II.5.1.2 and II.5.2.2)</sup> effect in  $C'$  (e.g. Figure III.3.3.1.b), the effect in  $C$  (e.g. Figure III.3.3.1.a) can be considered validated in  $C'$ . This is a weaker form of validation than unsupervised rediscovery, but effect eigensignals in  $C'$  may still be strong and highly significant.

Alternatively, effects can be supervisedly validated by associating them with non-GEP covariates. For example, detected effects with ranks #27 and #47 in the FFPE cohort GSE31312 were significantly associated with patient survival in the same cohort ( $p_{\#27} = 2.1 \cdot 10^{-5}$  and  $p_{\#47} = 4.2 \cdot 10^{-5}$ , log rank tests between Kaplan-Meier survival estimates for samples partitioned at  $\pm \frac{1}{2}$  standard deviations of effect eigensignal strengths; cf. III.3.1.1). This is also the reason, why I performed biostatistical analyses for these two effects (indexed with  $\nu = 134$  and  $\nu = 135$  respectively) in addition to all 133 effect tuples validated above<sup>(cf. III.1.2.2)</sup>. (Effect  $\nu = 134$  even turned out to have the most consistent association with survival of all effects in all cohorts<sup>(cf. III.2.5.1)</sup>, while  $\nu = 135$  showed still significant yet relatively weak survival association in other cohorts; cf. .)

To define an effect on genomic level, i.e. independent of concrete samples underlying its original discoveries in possibly several different patient cohorts, consensus gene axes are computed. Likewise, consensus gene correlations can be computed. They serve for definition of gene scores used for genomic biostatistical analyses of effects. Additionally, they are the basis for effect application to potentially new cohorts and computation of consensus sample axes<sup>(in III.1.4)</sup>.

## III.1.3.1

## Consensus gene axes

Let  $\nu$  be the index of an effects tuple with validated effects from  $\hat{\tau}$  cohorts<sup>(cf. III.1.2.2)</sup>. In the optimal case, the effect has been rediscovered in all cohorts ( $\hat{\tau} = \hat{C}$ ). For only supervisedly validated effects,  $\hat{\tau}$  equals one.

Let  $\{ |b_{C_i, k_i}^g| \}_{i=1 \dots \hat{\tau}}$  denote the corresponding set of gene axes for effects  $k_i$  in respective cohorts  $C_i$ . Likewise, let  $\{ |r_{C_i, k_i}^g| \}_{i=1 \dots \hat{\tau}}$  denote their gene correlations (with the effect's sample axis in the respective cohort). Their gene weights of the final effect focus are denoted by  $\{ |w_{C_i, k_i}^g| \}_{i=1 \dots \hat{\tau}}$ .

To combine these genomic information over all available cohorts  $C_i$ , weighted arithmetic averages of the above vectors are computed. This results in the consensus gene axis respectively in consensus gene correlations for this set of effects. Pairwise cross-cohort correlations  $r_{(C_i, k_i; C_j, k_j)}$  between effects<sup>(cf. III.1.2.1)</sup> are respected in form of weights when computing these averages. Additionally, they are utilized to align gene axes as they may be anti-correlated (an unaligned summation would cancel signal strengths). As for pairwise correlations<sup>(cf. III.1.2.1)</sup>, the gene focus for the combined effect is again defined as the maximum of respective gene weights.

The **consensus gene axis** for a tuple  $(k_i)_{i=1 \dots \hat{\tau}}$  of effects from  $\hat{\tau}$  cohorts with validation index  $\nu$ <sup>(cf. III.1.2.2)</sup> equals

$$|a_\nu^{g,c}| \equiv \sum_{i=1}^{\hat{\tau}} s_i w_i |b_{C_i, k_i}^g|,$$

where the signs  $s_i \equiv \text{sign}(r_{(C_i, k_i; C_1, k_1)})$  align axes to each other and geometrically averaged

weights  $w_i \equiv \frac{1}{\sqrt{\prod_{i'=1}^{\hat{\tau}} |r_{(C_i, k_i; C_{i'}, k_{i'})}|}}$  respect the correlation of gene axis  $i$  to all others<sup>(cf. III.1.2.1)</sup>.

The **consensus gene correlations** for the same effects tuple is defined likewise as

$$|r_\nu^{g,c}| \equiv \sum_{i=1}^{\hat{\tau}} s_i w_i |r_{C_i, k_i}^g|.$$

Finally, **consensus gene weights** are defined as the maximum of the gene foci in respective cohorts:

$$\langle e_i^g | w_\nu^{g,c} \rangle \equiv \max_{i'=1 \dots \hat{\tau}} \langle e_i^g | w_{C_{i'}, k_{i'}}^g \rangle$$

Together,  $(|a_\nu^{g,c}|, |r_\nu^{g,c}|, |w_\nu^{g,c}|)$  describe a particular **genomic consensus effect**.

Eqn. III.1.3.1) Consensus gene axis, consensus gene correlations and consensus gene weights for a validated effect  $\nu$  describe a genomic consensus effect

The procedure of consensus axis computation can be viewed like a *soft form of intersection* for effects that were detected in multiple cohorts. If a gene is strongly correlated to the common effect in only one of several detection cohorts, it is assigned lower consensus signal strength  $\langle e_i^g | a_\nu^{g,c} \rangle$  and weaker consensus correlation  $\langle e_i^g | r_\nu^{g,c} \rangle$  than genes that are strongly correlated to the same effect in more cohorts. Thus, potential false positive genes (with respect to the common effect) are sorted down and hence true positives (from perspective of all participating cohorts) are sorted to the top. This definition should further increase the biological specificity of multiply rediscovered consensus effects.

In the special case of a supervisedly validated effect that was only discovered in a single cohort<sup>(cf. III.1.2.3)</sup>, its consensus gene axis and its consensus correlations are identical to its gene axis and its gene correlations in its detection cohort. From a processing point of view, this makes no difference. However, relative to effects based on multiple discoveries, top genes are usually less robustly determined here and thus some top genes may still be cohort-specific. Hence, when the effect is applied to other cohorts<sup>(cf. III.1.4)</sup>, some of its top genes may turn out to be false positives, i.e. they are no longer significantly correlated to the effect's sample axis in this other cohort. Still, if the effect has already been supervisedly validated as a whole against the GEP signal of this other cohort<sup>(cf. III.1.2.3)</sup>, only a minority of its top genes can be such false positives.

While all cohorts measure the same genes, they measure different samples. Hence, no common consensus sample axis can be computed here. Notably, consensus sample axes could be computed for data scenarios where the *same* set of samples has been measured several times (preferably by independent labs or teams). Still, cohort-specific sample axes based on consensus gene axes can be and are computed for sample classification by genomic consensus effects<sup>(cf. III.1.4.2)</sup>.

### III.1.3.2 Consensus gene scores and their correlation

While consensus gene axes  $|a_v^{g,c}\rangle$  encode information about folding strengths of genes in effects, consensus gene correlations  $|r_v^{g,c}\rangle$  encode the consistency of gene regulations with the effect's sample axes in the respective cohorts. It is an interesting question, whether genes with a stronger differential signal or with higher correlation to an effect (and thereby on average also to other top genes of this effect) are biologically more relevant. I assume that biologically highly correlated genes may also be important to understand the underlying pathway, even if they have a relatively weak differential signal. Like for effect validation before<sup>(cf. III.1.2.1)</sup>, both sources of information may be important. I assume that the combination of both information sources maximizes biological specificity of gene rankings. Hence, consensus gene scores are defined based on the component-wise product:

$$|a_v^{g,c} \cdot r_v^{g,c}\rangle \equiv \sum_{i=1}^m |e_i^g |a_v^{g,c}\rangle |e_i^g |r_v^{g,c}\rangle |e_i^g\rangle$$

Eqn. III.1.3.2.a) Consensus gene scores for a validated effect  $v$  (combining signal strength and correlation information)

These scores are also used for genomic analyses of consensus effects, in particular for gene set enrichment analyses<sup>(cf. III.3.1.2 and III.1.5)</sup>.

To quantify the genetic similarity of two consensus effects  $v_1$  and  $v_2$ , their consensus gene scores are correlated. Again the maximum effect focus is used for comparison, analogous to cross-cohort comparison of discovered effects for validation<sup>(cf. III.1.2.1)</sup>.

$$r_{(v_1;v_2)}^c \equiv [ |a_{v_1}^{g,c} \cdot r_{v_1}^{g,c} | |a_{v_2}^{g,c} \cdot r_{v_2}^{g,c} | ]_{|w_{(v_1;v_2)}^{g,c}\rangle}, \text{ where}$$

$$|w_{(v_1;v_2)}^{g,c}\rangle \equiv \sum_{i=1}^m \max(\langle e_i^g |w_{v_1}^{g,c}\rangle, \langle e_i^g |w_{v_2}^{g,c}\rangle) |e_i^g\rangle$$

Eqn. III.1.3.2.b) Correlation of consensus gene scores

Pairwise correlations  $r_{(v_1, v_2)}^c$  of consensus gene scores for all pairs  $(v_1; v_2) \in [1, 135]^2$  of validated effects are depicted below:

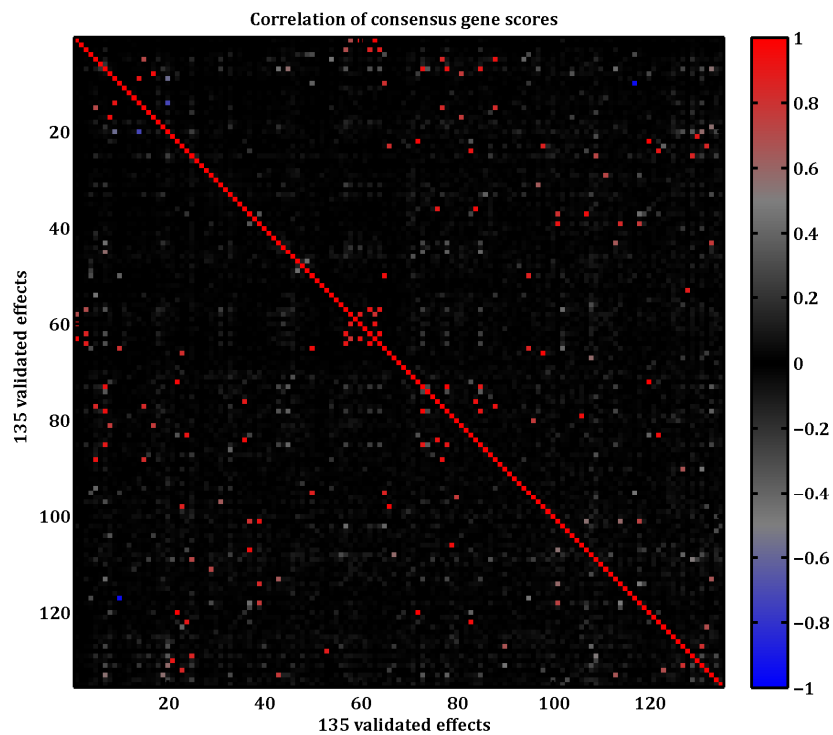


Figure III.1.3.2.a) Pairwise correlation  $r_{(v_1, v_2)}^c$  of consensus gene scores for all validated effects

The permitted redundancies for alternate views<sup>(cf. III.1.2.2)</sup> are visible as strong<sup>(red or blue)</sup> correlations outside of the diagonal. These genetic similarities between some of the consensus effects are respected during biostatistical evaluation<sup>(III.3)</sup>. Sometimes they are utilized to make statements about a group of several equivalent consensus effects at the same time<sup>(e.g. III.3.3.2)</sup>.

Most validated effects have uncorrelated or only partially correlated consensus gene scores, i.e. these effects probably represent biologically unrelated effects, e.g. independent gene regulation networks. (Depicted correlations are also provided in tabular form within [DLBCL Master Table 2015, main overview.xlsx](#).)



While cohorts measure the same genes, they measure different samples. Therefore, consensus gene axes of validated effects are applied to each cohort individually in order to compute effect strengths for their respective samples. Based on these effect strengths, samples can be arranged and classified by a particular effect. This classification can be applied to dissected cohorts in which consensus gene axes have been discovered and learned in the first place, as well as to new patient cohorts (for which the same gene expressions have been measured). Additionally, these classifications are utilized for biostatistical associations of effects with sample covariates (e.g. gender or survival).

To realize this classification, large portions of signal dissection<sup>(Chapter II)</sup> can be reused. Essentially, the search strategy is skipped and replaced by learned consensus gene axes of the respective effect.

## III.1.4.1

## Cleaned signal

All discovered effects<sup>(cf. III.1.1.2)</sup> that were not validated<sup>(cf. III.1.2.2 and III.1.2.3)</sup> are cohort-specific effects. They are considered lab effects that are not of biological origin. To make gene expression signals of different cohorts more comparable, these cohort-specific effects should be removed.

Let  $\widehat{k}_C$  denote the number of effects discovered by dissection of the initial gene expression signal  $M_{C,0}$  of cohort  $C$ . Let  $K_{C,\text{lab}} \subset [1, \widehat{k}_C] \subset \mathbb{N}$  be the index set for all detected effects in this cohort that did not validate<sup>(cf. III.1.2.2 and III.1.2.3)</sup>. Then the cleaned signal for this cohort is obtained by subtraction of all these lab effects<sup>(cf. Eqn. III.1.4.1)</sup>.

$$L_C \equiv \sum_{k \in K_{C,\text{lab}}} E_{C,k}$$

$$M_{C,0}^{\text{cleaned}} \equiv M_{C,0} - L_C$$

Eqn. III.1.4.1) Gene expression signal after dissecting cohort-specific effects

$K_{C,\text{lab}}$  Indices of effects detected in cohort  $C$  that did not validate<sup>(cf. III.1.2.2 and III.1.2.3)</sup>.

$E_{C,k}$  Final eigensignal of effect  $k$  in cohort  $C$ <sup>(cf. II.4.2.1)</sup>.

## III.1.4.2

Application of consensus gene axes  
and sample classification by eigensignal strengths

Application of a validated effect  $\nu$  to a cohort  $C$  computes the eigensignal of this effect in the cleaned signal  $M_{C,0}^{\text{cleaned}}$ <sup>(cf. Eqn. III.1.4.1)</sup> of this cohort. This eigensignal can be utilized for sample classification by the effect. Precise steps are detailed next.

For a consensus gene effect given by its gene axis  $|a_\nu^{g,C}\rangle$  and its gene correlations  $|r_\nu^{g,C}\rangle$ <sup>(cf. Eqn. III.1.3.1)</sup> first the significance of gene correlations is estimated and the gene focus  $|w_\nu^g\rangle$  is updated in the same way as during detection<sup>(cf. Eqn. II.3.1.4.b)</sup>. Then the twin sample axis  $|a_{C,\nu}^{s,C}\rangle$  for cohort  $C$  is computed from the gene axis  $|a_\nu^{g,C}\rangle$  using weights  $|w_\nu^g\rangle$ <sup>(cf. Eqn. II.3.1.3)</sup>, i.e. its components  $\langle e_j^s | a_{C,\nu}^{s,C} \rangle$  equal weighted projections of samples  $\langle s_{C,j} | a_\nu^{g,C} \rangle_{|w_\nu^g}^0$ . Additionally, consensus sample correlations  $|r_{C,\nu}^{s,C}\rangle$  for cohort  $C$  with the gene axis  $|a_\nu^{g,C}\rangle$  are computed using weights  $|w_\nu^g\rangle$ <sup>(cf. Eqn. II.3.1.4.a)</sup>, i.e.  $\langle e_j^s | r_{C,\nu}^{s,C} \rangle = [s_{C,j} | a_\nu^{g,C}]_{|w_\nu^g}$ . Finally, the cohort-specific sample focus  $|w_{C,\nu}^s\rangle$  is computed from these correlations<sup>(cf. Eqn. II.3.1.4.b)</sup>. Together, these effect axes, correlations and weights provide the same information in the same format as determined by the search strategy<sup>(cf. II.3)</sup> for a discovered effect during detection. This format identity is utilized to replace the search strategy. All other parts of signal dissection are identically reused for effect application.

The resulting regressed eigensignal  $E_{C,\nu}^C$ <sup>(cf. II.4.2.1)</sup> for consensus effect  $\nu$  in cohort  $C$  can be utilized to infer sample scores. They can subsequently be used for sample classification purposes or for biostatistical associations of effects with sample covariates. For example, the eigensignal  $E_{\text{GSE10846.RCHOP},134}^C$ <sup>(see Figure III.3.3.1.b, panel d)</sup> can be utilized to quantify sample involvements in this effect for subsequent survival analyses.

To this end and more precisely, I define sample eigensignal strengths as column averages of these eigensignal matrices  $E_{C,v}^c$ , weighted with the consensus gene focus  $|w_v^g\rangle$  of the effect<sup>(Eqn. III.1.4.2)</sup>, and using correlation signs to sum correlated and anti-correlated genes constructively.

$$|u_{C,v}^s\rangle \equiv \sum_{j=1}^n \langle \text{sign}(r_v^{g,c}) \cdot w_v^g | E_{C,v}^c(I_0, j) \rangle |e_j^s\rangle$$

Eqn. III.1.4.2) Sample eigensignal strengths

Other than sample effect strengths used for regression<sup>(cf. II.4.1.1)</sup>, sample eigensignal strengths are relatively robust against outliers in raw sample signals and do no longer contain signal parts of partially correlated foreign effects (as dissection strengths are zero for them). Hence, I assume that eigensignal based sample scores are the purest with respect to an effect. I also utilize them for sample partitioning for *clinical association analyses and survival analyses*. (This assumption has been supported by experience with several alternate sample scores and with multivariate survival analyses during development: Sample orders based on sample eigensignal strengths resulted in the lowest  $p$  values for survival associations.)

### III.1.5

## Overview of Scores


To clarify gene and sample scores utilized for biostatistical analyses of consensus effects and to compare them with scores used for other purposes in signal dissection, the following table provides an overview of all gene and sample scores that are available for validated effects.

| Gene scores                             |   |
|---|---|
| $ a_v^{g,c}\rangle$                     | The cohort-independent <b>consensus gene axis</b> of validated effect $v$ is a weighted average of gene axes in underlying detection cohorts <sup>(cf. Eqn. III.1.3.1)</sup> . Components can be interpreted as the <i>differential expression strengths</i> of individual genes (in $\log_2$ (ratios) units) between samples of patients that were ordered and distinguished by corresponding effects in underlying detection cohorts <sup>(cf. III.1.2.2)</sup> .   |
| $ r_v^{g,c}\rangle$                     | Likewise, the cohort-independent <b>consensus gene correlations</b> of validated effect $v$ are based on weighted averages of gene correlations in underlying detection cohorts <sup>(cf. Eqn. III.1.3.1)</sup> . Components can be interpreted as the <i>consistency</i> of an individual gene's expressions with sample eigenorders of the effect in underlying detection cohorts <sup>(cf. III.1.2.2)</sup> . The nearer a gene's correlation is to $\pm 1$ , the more likely it is that the gene is regulated by the underlying biological program that is represented by this effect. The nearer towards zero, the less likely it is that expressions of the gene can be explained by the current effect.  |
| $   a_v^{g,c}  \cdot r_v^{g,c} \rangle$ | <b>Consensus gene scores</b> are defined as the gene-wise product of the effect's absolute consensus gene axis with its consensus gene correlations <sup>(cf. Eqn. III.1.3.2.a)</sup> . Hence, top genes when ranking by these scores are <i>both strong and consistent</i> with respect to effect $v$ . Using both information is assumed to result in <i>biologically most specific effect genes</i> at top and bottom. Therefore, this ranking is <i>utilized as basis for all genomic analyses</i> of effects (i.e. primarily for gene set enrichment analyses and for gene ontology analyses).   |
| $ u_{C,v,i}^g\rangle$                   | During application of a consensus gene effect to a specific cohort $C$ , signal dissection computes <b>gene effect strengths</b> for regression iterations $i$ as during detection <sup>(cf. II.4.1.1)</sup> . These gene effect strengths are utilized to determine the gene eigenorder for the next bimonotonic regression iteration and to determine the final gene eigenorder. In effect definition plots, final gene effect strengths $ u_{C,v,i}^g\rangle$ are depicted in the center panel.  |
| $ u_{C,v}^g\rangle$                     | After determining the final eigensignal $E_{C,v}^c$ for validated effect $v$ in cohort $C$ <sup>(cf. III.1.4.2)</sup> , <b>gene eigensignal strengths in cohort <math>C</math></b> can be computed to quantify the cohort-specific gene involvement in the effect, using sample correlation signs and the effect's sample focus as weights: $ u_{C,v}^g\rangle \equiv \sum_{i=1}^m \langle \text{sign}(r_{C,v}^{s,c}) \cdot w_{C,v}^s   E_{C,v}^c(i, J_{C,0}) \rangle  e_i^g\rangle$ . Other than gene effect strengths $ u_{C,v,i}^g\rangle$ before, gene eigensignal strengths are relatively robust against outliers in raw signals of single genes $ g_{C,i}\rangle$ and do no longer contain signal parts of partially correlated foreign effects (as dissection strengths are zero for them). They would also be a justifiable candidate for genomic analyses of effects, but they are only available in a cohort-dependent context, requiring a method to merge scores from different cohorts. In contrast, consensus gene scores above are already based on merged consensus axes and correlations. |

| Sample scores           |  |
|-------------------------|--|
| $ a_{C,v}^{s,c}\rangle$ | The <b>consensus sample axis</b> for validated effect $v$ in cohort $C$ is the twin axis of $ a_v^{g,c}\rangle$ (cf. Eqn. III.3.1.3). It is computed as weighted projections of all samples in $C$ on the consensus gene axis, i.e. its components equal $\langle s_{c,j}   a_v^{g,c} \rangle_{w_v^g}$ (cf. III.1.4.2). They can be interpreted as the <i>differential expression</i> of samples (in $\log_2$ (ratios) units) between co- and anti-regulated genes in the effect.  |
| $ r_{C,v}^{s,c}\rangle$ | <b>Consensus sample correlations</b> for validated effect $v$ in cohort $C$ , i.e. components equal $\langle e_i^g   r_{C,v}^{s,c} \rangle = \langle s_{c,j}   a_v^{g,c} \rangle_{w_v^g}$ (cf. III.1.4.2).<br>For interpretation, see $ r_v^{g,c}\rangle$ above.   |
| $ u_{C,v,i}^s\rangle$   | During application of a consensus gene effect to a specific cohort $C$ , signal dissection computes <b>sample effect strengths</b> for regression iterations $i$ as during detection (cf. II.4.1.1). These sample effect strengths are utilized to determine the sample eigenorder for the next bimonotonic regression iteration and to determine the final sample eigenorder. In effect definition plots, final sample effect strengths $ u_{C,v,i}^s\rangle$ are depicted in the center panel.   |
| $ u_{C,v}^s\rangle$     | After determining the final eigensignal $E_{C,v}^c$ for validated effect $v$ in cohort $C$ (cf. III.1.4.2), <b>sample eigensignal strengths in cohort <math>C</math></b> can be computed to quantify the cohort-specific sample involvement in the effect, using gene correlation signs and the effect's gene focus as weights: $ u_{C,v}^s\rangle \equiv \sum_{j=1}^n \langle \text{sign}(r_v^{g,c}) \cdot w_v^g   E_{C,v}^c(I_0, j) \rangle  e_j^s\rangle$ (Eqn. III.1.4.2). Other than sample effect strengths $ u_{C,v,i}^s\rangle$ before, sample eigensignal strengths are relatively robust against outliers in raw signals of single samples $ s_{c,j}\rangle$ and do no longer contain signal parts of partially correlated foreign effects (as dissection strengths are zero for them).<br>Hence, I assume that eigensignal based scores are the purest with respect to an effect and utilize them <i>for sample partitioning for all clinical association analyses and survival analyses</i> . (This assumption has been supported by experience with several alternate sample scores and with multivariate survival analyses during development: Sample orders based on sample eigensignal strengths resulted in the lowest $p$ values for survival associations.) |

Table III.1.5) Overview of available gene and sample scores for validated effects

Several gene and sample scores are available for validated consensus effects. For genomic analyses, only consensus gene axes and consensus gene correlations are needed (the same global gene space is shared by all cohorts). For ordering and classifying samples in a concrete cohort, an effect is applied to it (cf. III.1.4.2), resulting in cohort-dependent sample eigensignal strengths for clinical association analyses. This overview lists all available scores and explains their differences.

Different scores are utilized for different types of effect presentation. Effect definitions in tabular form are available in  *DLBCL Master Table 2015, gene orders.xlsx*; they show all genes and provide columns for  $|a_v^{g,c}\rangle$ ,  $|r_v^{g,c}\rangle$  and  $|a_v^{g,c} | r_v^{g,c}\rangle$ . Effect overview plots (e.g. Figure III.3.2.1) are always presented in the empirical eigenorder (cf. II.4.1.2), for which the respective eigensignal is regressed (i.e. they are ordered by final effect strengths  $|u_{C,v,i}^g\rangle$  and  $|u_{C,v,i}^s\rangle$  respectively). Additionally, plots apply a filter to exclude strongly folded yet uncorrelated genes from the view. This filter demands that absolute consensus correlations  $|\langle e_i^g | r_v^{g,c} \rangle|$  are  $\geq 50\%$  of their maximum. Gene tables for biostatistical evaluations (III.3) are presented in descending order of the combined scores  $|a_v^{g,c} | r_v^{g,c}\rangle$ , as these scores are also utilized for all genomic analyses. Hence, these gene tables list genes at top ranks that have both strong folding and high correlation (if available for the respective effect). For abbreviation, let  $a_{v,i}^{g,c} \equiv \langle e_i^g | a_v^{g,c} \rangle$  and  $r_{v,i}^{g,c} \equiv \langle e_i^g | r_v^{g,c} \rangle$ ; consequently  $|a_{v,i}^{g,c} | r_{v,i}^{g,c}\rangle = \langle e_i^g | |a_v^{g,c} | r_v^{g,c} \rangle$ . (The gene index  $i$  may be suppressed in the context of an arbitrary gene, e.g. for gene table headers. Or it may be replaced by the gene name in the text for clarity.)

## III.2 Multivariate Survival Prediction

---

*Gene expression effects associated with patient outcome are probably most informative towards understanding the molecular pathogenesis of DLBCL. To reveal these effects, multivariate Cox survival models are trained. All validated GEP effects are systematically analyzed for their individual and combined predictive performance.*

*Many GEP effects can explain significant portions of patient outcome. Several genetically distinct effects can explain the same portions. Some effects can explain hierarchical survival dependencies after patients have been stratified by other effects. In principle, all of these effects could contain genes that are causal to the disease or its subtypes, reflecting the genetic heterogeneity and complexity of DLBCL.*

*Results reveal that a particular unsupervisedly discovered GEP effect<sup>(index  $v=134$ , evaluated in III.3.3.1)</sup> can predict patient outcome exceptionally well with  $p = 4.5 \cdot 10^{-17}$ , while the unsupervisedly rediscovered cell-of-origin induced effect<sup>( $v=129$ , evaluated in III.3.2.2)</sup> follows only at rank two<sup>(Figure III.2.5.1.a)</sup> with  $p = 1.1 \cdot 10^{-12}$ . Additionally, another GEP effect can improve predictions on top of effect  $v=134$  in a hierarchical manner, while the same is not possible on top of the COO induced effect. This might indicate that genes in  $v=134$  are a more natural choice to identify subtypes of DLBCL. So far, many genes that have been investigated with respect to their contribution to DLBCL lymphomagenesis belong to the COO induced effect. Given that the discovered effect  $v=134$  is not only more consistently associated with patient outcome but also has fewer top genes with higher correlation<sup>(cf. Figure III.3.3.1.a)</sup> relative to the larger and less specific COO induced effect<sup>(cf. Figure III.3.2.2.b)</sup>, its top genes might be interesting for further experimental investigation.*

### III.2.1 Survival Model and Effect Selection

---

Before fitting concrete models and presenting predictor results, a brief recap for Cox proportional hazard models is provided here.

Available data and follow-up types are listed, the choice of sample scores utilized for predictions is motivated and the iterative selection procedure for GEP effects that significantly explain patient outcome is clarified.

#### III.2.1.1 Cox Proportional Hazard Models

---

All fits in this subchapter are based on Cox proportional hazard survival models<sup>[98]</sup>. For a selected training set of samples, these models test the association of survival data with a selection of explanatory variables  $X_v$ . These  $X_v$  correspond to measured information about training samples, for example their average expression of validated GEP effects. Cox models<sup>(Eqn. III.2.1.1)</sup> estimate a baseline hazard function  $h_0(t)$  from survival data, and

regress on coefficients  $\beta_v$  for selected explanatory effects  $X_v$ . Fitted estimates  $\hat{\beta}_v$  utilize selected explanatory variables to predict patient risk ratios relative to baseline risk.

$$h_X(t) \equiv h_0(t) \cdot \exp\left(\sum_{v \in V} \beta_v X_v\right)$$

Eqn. III.2.1.1) Cox proportional hazard models

|                                  |   |
|----------------------------------|---|
| $v, V$                           | Explanatory effect index and the selected set of indices. All $v \geq 1$ correspond to indices of selected validated GEP effects <sup>(cf. III.1.2)</sup> , while $v \leq 0$ correspond to prescribed temporary model variables like patient age <sup>(cf. III.2.1.4)</sup> . |
| $X_v$                            | Explanatory variable with index $v$ . (For example, the average gene expressions of a validated GEP effect for all samples.)  |
| $\beta_v$                        | Coefficients. Numeric values are obtained by regression and are denoted as $\hat{\beta}_v$ . In case of standardized variables $X_v$ , $\beta_v$ can be compared directly to estimate the relative impact of explanatory variables on survival.                               |
| $h_0(t)$                         | Baseline hazard function. Estimated empirically by the model. High hazards are equivalent to high risk and correspond to adverse outcome.   |
| $h_X(t)$                         | Hazard rate at $X \equiv (X_v)_v$ at time $t$ . Substitute $X_v$ by numeric values for concrete samples $j$ to predict their hazard rate. Ratios of hazard rates quantify how much more probable an event is for one sample compared to the other.                            |
| $\exp(\beta_v X_{v,j})$          | The predicted hazard ratio relative to baseline for patient $j$ due to its value $X_{v,j}$ for effect $v$ .   |
| $\sum_{v \in V} \beta_v X_{v,j}$ | The predicted $\log(\text{hazard ratio})$ for patient $j$ based on its values $X_{v,j}$ for all effects in the model.   |

The aim is to find gene expression effects that allow the prediction of significantly different hazard ratios between patients. Generally, high hazard rates, i.e. high risks correspond to adverse patient outcome. Predictions are computed relative to estimated baseline hazards and are described by  $\log(\text{hazard ratios})$ , similar to using  $\log_2(\text{ratios})$  when comparing sample gene expressions to cohort average expressions. Kaplan Meier survival estimates for risk intervals of  $\log(\text{hazard ratios})$  will be utilized to visualize survival differences predicted by particular trained Cox models. All fits of Cox models are computed with the **coxphfit** function of the MATLAB® Statistics toolbox.

Conceivable nonlinear dependencies (for example, an effect might only influence outcome if expressed above a threshold) cannot be fitted exactly by Cox models, but are linearly approximated in the exponential. Symmetric dependencies (e.g. if average expression of an effect maps to baseline survival, but both upregulation and downregulation cause adverse outcome) cannot be fitted or revealed by such models.

### III.2.1.2 Available survival data and its limited information

Survival data from patients for samples  $j$  are available in form of follow-up times  $t_j$  after diagnosis and Boolean flags  $b_j$  that indicate whether an event (a death or a progression) was observed ( $b_j = 1$ ) or if a patient was lost to follow-up before it was observed ( $b_j = 0$ , right censoring). In total, 947 DLBCL samples with follow-up events are available: 181/181 in GSE10846.CHOP<sup>[5]</sup>, 76/166 in GSE4475.CHOP<sup>[95]</sup>, 220/233 in GSE10846.R-CHOP<sup>[5]</sup> and 470/498 in GSE31312.R-CHOP<sup>[29]</sup>.

Observed deaths are used when estimating overall patient survival. These events are different from observed disease progressions that do not necessarily lead to deaths, but are significantly associated with these later events. Progression events have the statistical advantage that they usually can be observed even within relatively short follow-up studies, while deaths may appear as relatively uninformative censored events, especially if the follow-up is too short. Therefore, progression events can potentially convey more information about survival than the usually fewer death events for the same follow-up duration; this is especially useful for cohort GSE10846.R-CHOP due to its relatively short follow-up. While death events were recorded for all cohorts, only in R-CHOP treated cohorts progression events are available.

It should be illuminated before fitting any survival model that except for time information and even when combining all four cohorts, survival data comprises **just 947 bits of information**. Their primary use is to serve as a biologically independent validation of *already discovered and validated* effects on GEP level. If only survival information was used as primary information source, for example to supervisedly find genes with

expressions associated with these bits<sup>(as done for signature definition in [5])</sup>, this could result in many false positive genes, since the information base is too low. Even if these genes are arranged and filtered via hierarchical clustering to find related subsets of genes<sup>(as also done in [5])</sup>, the information base is much smaller than for unsupervised dissection of the full GEP signal; hence, I utilize survival data *only for independent validation rather than for direct supervised learning* of interesting genes or GEP effects.


For the same reason, survival data should not be used to exclusively select only those GEP effects that are most consistently associated with DLBCL patient outcome. That would exclude many also significantly associated GEP effects that might contain biological true positive genes that are causal to the disease. While it is necessary to make a selection of effects for predictor construction (and only the most consistent effects are selected to this end), it should be kept in mind that in principle *all* effect candidates with significant survival association might describe important and causal parts of this genetic disease.

A third problem tied to having 947 bits of information is the problem of “overlearning”: It is relatively easy to inadvertently learn 947 bits “by heart” with a model based on many variables; such a predictor would produce perfect results for the 947 samples it was trained with, but would not generalize well or at all to *new* patients. Herein this study this is less problematic, as predictors are based only on already validated GEP effects that are biological true positive effects and are usually supported by many correlated genes (or probesets). Still, the construction procedure for predictors should and will be validated to quantify and prove their predictive performance.

### III.2.1.3

## Choosing sample scores

Given a validated GEP effect  $v$ , several sample scores<sup>(cf. III.1.5)</sup> could be utilized for  $X_v$ , for example correlations of samples to the effect’s consensus gene axis. Alternatively, sample projections on its consensus gene axis, its consensus gene weights or effect strengths based on its eigensignal in the respective cohort could be utilized. While all are based on the validated gene order from the effect’s consensus gene axis, only eigensignal strengths  $|u_{v,C}^s|$ <sup>(cf. Table III.1.5)</sup> are also based on bimonotonic regression in the respective cohort. This may be the reason why multivariate survival analyses performed during development indicated that sample orders based on sample eigensignal strengths resulted in the lowest  $p$  values for survival associations. I therefore consider them the highest-quality sample scores available and choose them for fitting all survival predictors in this subchapter.

Visually and for direct interpretation, a component  $(e_{j,C}^s | u_{v,C}^s)$  is simply a constructively weighted average expression of sample column  $j$  in the eigensignal heatmap of the respective effect application plot<sup>(e.g. Figure III.3.3.1.a, panel d)</sup>. The selected sample score allows for later practical applications of predictors, as consensus gene weights  $|w_v^g|$  needed for its computation have already been learned and are readily available for each validated GEP effect in . Additionally, the effect’s eigensignal can be computed from learned and available consensus gene axes even for future DLBCL cohorts<sup>(cf. III.1.4)</sup>.

Each sample of every cohort should have the same weight when training survival predictor models. To this end and to compensate for scaling differences between cohorts, I additionally standardize eigensignal strengths  $S_C \equiv \{(e_{j,C}^s | u_{v,C}^s) | j = 1 \dots n_C\}$  by dividing with their standard deviation in the respective cohort  $C$ . I.e. the final predictor variables are  $X_{v,C} \equiv \{(e_{j,C}^s | u_{v,C}^s) / \hat{\sigma}(S_C) | j = 1 \dots n_C\}$ , where  $n_C$  is the number of samples in

cohort  $C$ . This standardization also allows resulting  $\hat{\beta}_v$  of fitted Cox models to be *compared directly*, as they all multiply effect variables distributed with *standard deviation one*.

Sample scores are not shifted by their cohort mean before predictor training (as in standard  $z$  transformations), as zero already represents the theoretical global baseline (zero  $\log_2(\text{ratio})$ s indicate cohort-average gene expressions). Hence, samples with zero eigensignal (i.e. baseline expression) for a particular effect also map to baseline survival of fitted Cox models, since their products  $\hat{\beta}_v X_v(j)$  are zero, irrespective of fitted  $\hat{\beta}_v$ .

#### III.2.1.4 Correcting for survival factors that are not specific for DLBCL

Patient age is known to influence patient outcome, but it is not specific for DLBCL. Thus, assignment of significant explanatory value to GEP effects that are accidentally related to age should be avoided. To this end, I temporarily prescribe centered patient age as first explanatory variable  $X_0$  during GEP factor selection.

For a maximal training base, I merge all available survival events across all four patient cohorts. To this end, another factor should be prescribed: Two cohorts were treated by the current standard therapy of R-CHOP, whereas the other two were treated by the former standard of CHOP. These therapies are associated with different average survival. In addition, follow-up event types are different for the same pairs of cohorts. To prevent selection of GEP effects that are accidentally related to therapy or follow-up type, I prescribe  $X_{-1}$  as centered binary grouping variable for R-CHOP and CHOP events, whenever samples from both groups are used for predictor training.

After having selected all significant explanatory GEP effects, factors for age and therapy are removed again from predictor models to only keep factors that are specific for DLBCL.

#### III.2.1.5 Effect selection procedure and likelihood ratio tests

The procedure for selecting explanatory GEP effects is iterative. For each effect selection iteration, each of the 135 validated GEP effects is candidate-added to the current Cox model and it is quantified, how well it can explain survival *on top of* already explained dependencies, initially only on top of age and (if needed) therapy dependent outcome.

This can be computed via *likelihood ratio tests*, comparing the larger  $\{\text{age, effect}\}$  model with just the  $\{\text{age}\}$  model: Let  $l_2$  be the log-likelihood of the larger model and  $l_1$  be the log-likelihood of the smaller nested model (both are determined by the fitting procedure `coxphfit` from the MATLAB® Statistics toolbox). Then  $-2 \cdot (l_1 - l_2)$  is asymptotically  $\chi^2$  distributed with one degree of freedom as per Wilks' theorem<sup>[99]</sup>. Hence,  $p$  values readily can be approximated by integrating the respective upper tail of the  $\chi^2$  distribution.

GEP effects are filtered by their  $p$  values with respect to a tightly chosen error threshold of  $\alpha \equiv 10^{-4}$ . If significant effects remain, the one with the best  $p$  value is selected and added to the Cox model.

In the next iteration, again all GEP effects are candidate-added to the now larger model. Already selected effects can no longer provide any significant additional explanatory value; hence no effect can be selected twice. Again, the most explanatory effect is added to the model. This selection procedure continues as long as increasing model size by a particular GEP effect can fit survival data significantly better than the preceding smaller Cox model.

To prevent and quantify the problem of overlearning, trained Cox proportional hazard models should be validated with other observed survival data.

Three validation scenarios will be applied: first, a model is trained with R-CHOP-treated patients and validated with CHOP-treated ones. Secondly, a model is trained with CHOP-treated patients and validated with R-CHOP-treated ones. Besides these two classical validation schemes, thirdly a leave-one-out validation will be employed to estimate generalization performance of a model trained with all available samples.

Validation of fitted Cox models is only required to test their *generalization performance* and to confirm that the predictor did not learn observed survival data “by heart”. This model validation should not be confused with validation of effects on GEP level (either by applying them to other cohorts or even by their independent unsupervised detection in multiple cohorts). It is also different from validating discovered single GEP effects by their significant association with patient outcome.

Both the genetic heterogeneity of DLBCL and the low information base of only 947 bits cause that many GEP effects can explain significant portions of observed survival. Different GEP effects may appear as “the best ones”, given a particular set of training samples. In principle, *every significant GEP effect* has the potential to contain biological true positive genes that are causal to the disease’s pathogenesis.

To represent the most important ambiguities, several predictors are trained and validated:

- A bivariate predictor resulting from training with only R-CHOP-treated patients will be validated in the CHOP treated subset<sup>(III.2.2)</sup>.
- For comparison, another bivariate predictor will be presented that uses the rediscovered COO induced effect as primary explanatory variable. It also emerges from training with R-CHOP-treated patients, but only if manually excluding the strongest survival effect<sup>(III.2.3)</sup>.
- Similarly, a different bivariate predictor results from training with CHOP-treated patients only; it will be validated in the R-CHOP-treated subset<sup>(III.2.4)</sup>.
- Using the same significance thresholds for effect selection, additionally and finally a five-effect predictor is obtained when training with all samples<sup>(III.2.5)</sup>.

The number of effects that could contain causal genes for observed differences in patient outcome following chemotherapy can be reduced from 135 to approximately 20 (cf. Figure III.2.5.1.a). Of these 20, some are genetically correlated, and some explain only weak survival dependencies, probably affecting only few patients. Nine effects are used in the various predictor models and can explain significant portions of observed survival, albeit in part they explain the same portions. This cannot be further reduced because of the genetic heterogeneity that is intrinsic to DLBCL (several smaller patient subsets seem to show different outcome depending on their gene expressions in different effects).

All GEP effects selected for predictors will be presented and biostatistically evaluated in III.3 towards finding out the biologically most relevant gene expression differences and to sort out indirect effects about the tumor microenvironment.



For later comparison, survival differences between known subtypes ABC DLBCL and GCB DLBCL are of interest. The cell of origin induced GEP effect that underlies classification into ABC DLBCL and GCB DLBCL has been unsupervisedly rediscovered in a filtered form (cf. III.3.2.2), but here and for reference survival differences based on published sample classifications are depicted (Figure III.2.1.8) for each of the four analyzed patient cohorts (cf. III.1.1.1).

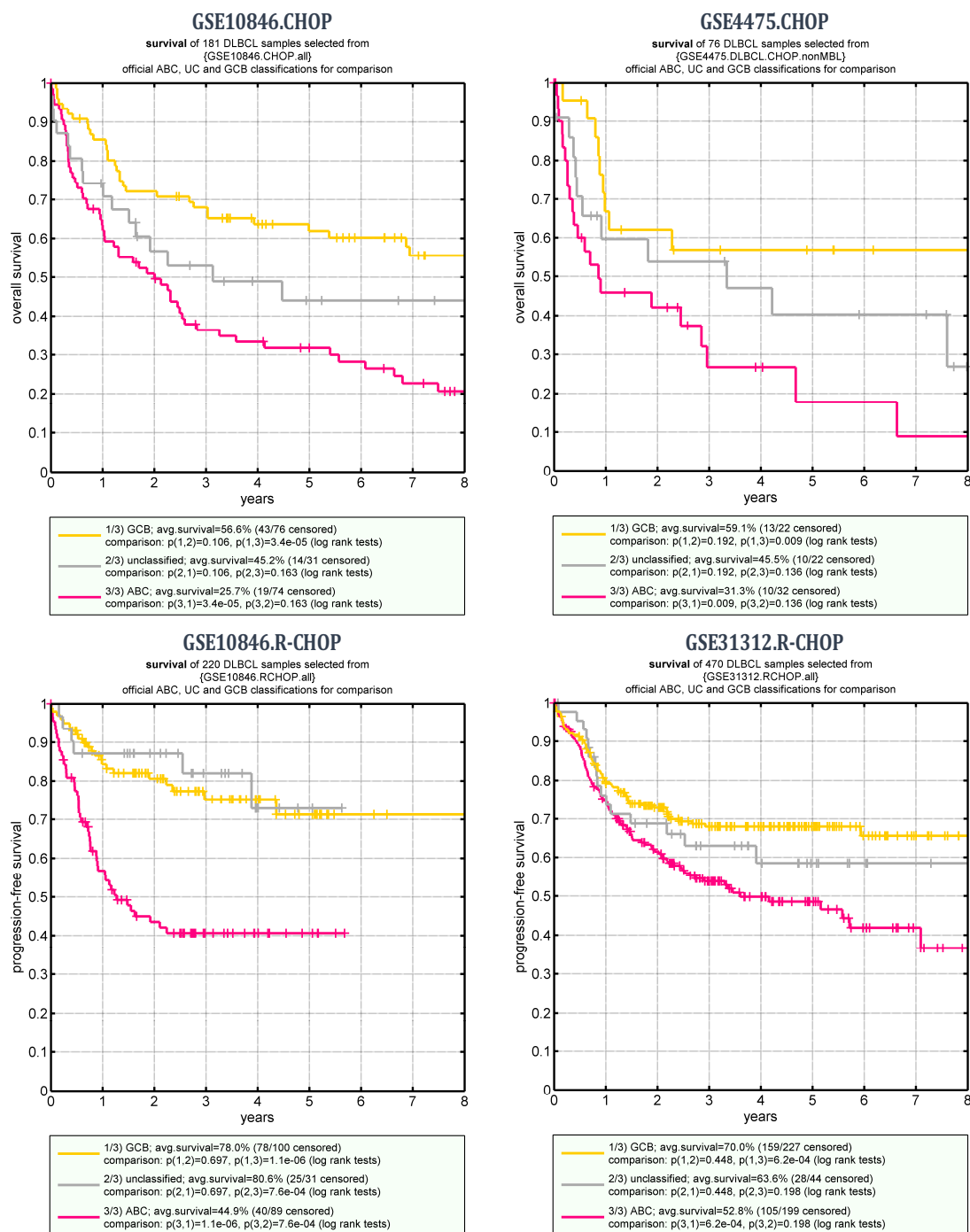


Figure III.2.1.8) Survival spread for standard classifications in ABC DLBCL, unclassified and GCB DLBCL in all four analyzed patient cohorts for comparison

Predictors should ideally show a survival spread at least as strong as between these known subtypes. Additionally, it would be interesting if predictors could reveal significant survival dependencies *within* ABC DLBCL or *within* GCB DLBCL.

Differences in the survival spread between ABC DLBCL and GCB DLBCL for the two R-CHOP treated cohorts demonstrate that the current classification procedure either does not generalize well or that it is not robust when transferring it from a cohort based on frozen cell material like GSE10846 to a FFPE based cohort like GSE31312.

### III.2.1.9 Revisiting binary subtype classifications and associated cutoffs

---

The established Bayes classification for ABC DLBCL and GCB DLBCL<sup>(as defined in [83] and applied to GSE31312, cf. Figure II.2.1.2)</sup> implicitly *assumes* that every patient is either of the ABC DLBCL or of the GCB DLBCL subtype. Only patients that cannot be significantly assigned to either group under this assumption are called unclassified. Thus, the noise level in the data and the chosen error cutoff determine the size of the unclassified group. If for example only two thirds of all patients were truly either ABC-like or GCB-like, while another third was *neither*, this assumption would be violated and the classifier would tend to underestimate the number of unclassified patients with respect to the COO induced effect, i.e. it would produce false positive classifications.

Average GEP effect strengths or  $\log(\text{hazard ratios})$  are able to encode more information about samples compared to binary subtype classifications. In particular, they represent the *gradual nature* of observed survival dependencies<sup>(see also II.1.2.6)</sup>. Given the straight-forward interpretation of hazard ratios for patients relative to baseline risk and with clinical precision in mind, it is tempting to replace hard-cut clinical subtypes by gradual GEP effect strengths in the midterm towards precision medicine, rather than forcing patients into a two-class model. The importance of a more accurate description than binary subtype classifications is confirmed by the existence of several genetically distinct GEP effects that are all significantly associated with DLBCL subtypes<sup>(cf. III.3.2.2)</sup>, but have strongly varying predictor capabilities with respect to patient outcome.

For this reason, hard cutoffs during effect detection are not used and all effects are defined in a cutoff-free way with correlations and weights for all measured genes. I only utilize cutoffs for presenting top-correlated genes or for visualization of predicted survival spreads via Kaplan Meier estimates. Even here, cutoffs are based on correlations with the effect respectively based on risk intervals for predicted  $\log(\text{hazard ratios})$ . Consequently, all samples with predicted hazards near baseline are combined in the same Kaplan Meier curve.

Thus, if a predictor can only predict strong survival dependencies for a few patients, most patients are in the baseline Kaplan-Meier curve, rather than trying to minimize the size of the effect's "unclassified" group over the limit that is justifiable based on GEP regulation strengths. Additionally, these hazard-based cuts can visualize predictor performance more naturally than, for example, quartile Kaplan Meier plots. This is especially true in presence of small patient subgroups that highly express a given effect and show significantly different outcome relative to other patients.

Using progression events from cohorts GSE10846.R-CHOP and GSE31312.R-CHOP as training base, two GEP effects are iteratively identified and selected that can explain significant survival dependencies. Validation in both CHOP-treated cohorts is successful. As selected GEP effects are genetically highly specific and distinct from the larger COO induced effect, these effects could yield novel biological insights into molecular causes of DLBCL.

The second explanatory variable ( $v = 127$ ) is shown to affect only patients in the lower risk partition with respect to the first effect ( $v = 134$ ), i.e. in terms of standard subtypes, GCB DLBCL patients are affected predominantly, while ABC DLBCL patients seem not to be affected by its underlying pathway. Because of an anti-aligned survival trend, effect  $v = 127$  is not significant if selected as primary effect, but highly significant as secondary effect, i.e. a multivariate analysis was necessary to reveal it.

### III.2.2.1 Selection of validated GEP effects as predictor variables

Centered age is prescribed as initial factor in this Cox model to prevent potential selection advantages for GEP effects that are related to patient age; this factor has  $p = 0.004$  over all R-CHOP-treated patients.

Next, each of the 135 validated GEP effects is candidate-added to the model as described (III.2.1) and likelihood ratio tests are utilized to quantify significance of their added explanatory value. With  $\alpha = 10^{-4}$  as tight significance threshold, eight GEP effects (Figure III.2.2.1.a) qualify as primary predictor variable:

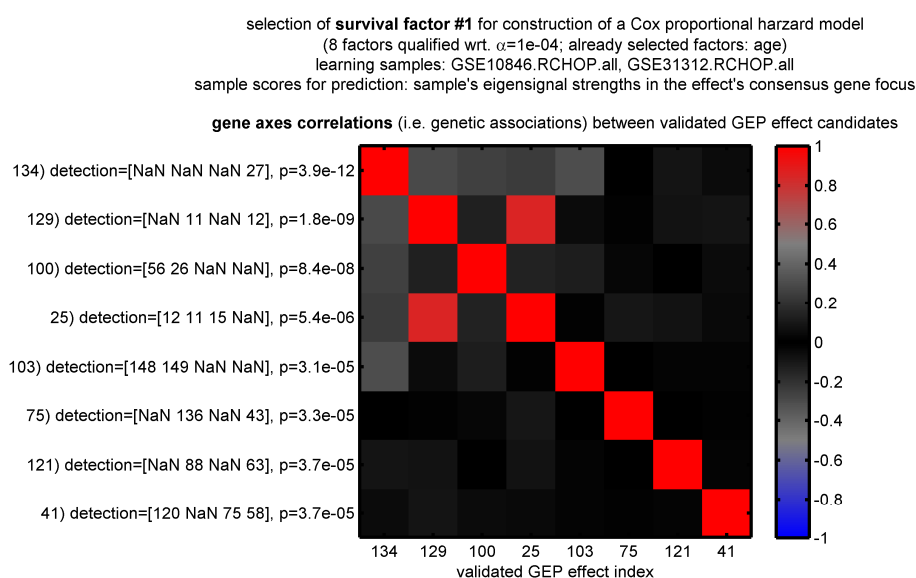


Figure III.2.2.1.a) Selection of the best primary GEP effect for survival prediction based on all 690 available R-CHOP-treated patients

Shown are all 8 GEP effects with a  $p$  value  $\leq 10^{-4}$  for their additional explanatory value of observed patient outcome (likelihood ratio test). The matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2). Indices of pre-consensus effects detected in the four source cohorts GSE10846.CHOP, GSE10846.R-CHOP, GSE4475.CHOP.nonMBL and GSE31312.R-CHOP are displayed on the left in square brackets; NaNs indicate that no sufficiently correlated gene axis was used for dissection of the respective cohort's signal.

With  $p = 3.9 \cdot 10^{-12}$ , effect  $v = 134$  can explain survival most consistently, followed by the COO induced effect  $v = 129$  on rank #2 with  $p = 1.8 \cdot 10^{-9}$ . Genetically,  $v = 134$  is weakly correlated to all four following effects, but only has 69 top genes (unique gene IDs with  $|\langle e_i^g | r_{134}^{c,g} \rangle| \geq 0.4$ ), while following top effects are larger (151 genes for  $v = 129$ , 95 for  $v = 100$  and 435 for  $v = 25$ , using same gene correlation cutoffs). This might indicate that it can capture survival-relevant genes with a higher biological specificity.

During detection, dissection of GSE31312.R-CHOP revealed  $v = 134$  in iteration #27, while no highly correlated gene axes were used for dissection of the other three cohorts. The reason for this cannot be an

ambiguous dissection of the larger COO induced effect, since it has also been rediscovered in dissection iteration #12 in GSE31312.R-CHOP. Probably different cell material (GSE31312 is FFPE based, while all three other cohorts are based on frozen cell material) makes an alternate dissection of signal parts more probably that remain *after* the COO induced effect has been dissected. Anyway,  $\nu = 134$  has been supervisedly validated, i.e. it has been shown to exist on GEP level in all four cohorts irrespective of cell material by applying its consensus gene axis<sup>(cf. III.3.3.1)</sup>.

On rank #4 ( $\nu = 25$ ), an alternate view onto the standard COO induced effect follows; its consensus gene scores are correlated to rank #2 with  $r_{(25;129)}^c = 0.85$ <sup>(cf. Eqn. III.1.3.2.b)</sup>. From GEP validation perspective,  $\nu = 25$  would have been preferred over  $\nu = 129$ , as it was independently discovered in three cohorts rather than in just two cohorts. This proves that the decision was right to allow redundancies in validation<sup>(III.1.2)</sup> in order to let survival (or other covariates) decide which alternate high-dimensional gene axis can represent true biological effects best.

After addition of  $\nu = 134$  as explanatory variable to the Cox model, only three effects<sup>(Figure III.2.2.1.b)</sup> remain that can explain *additional* significant survival dependencies:

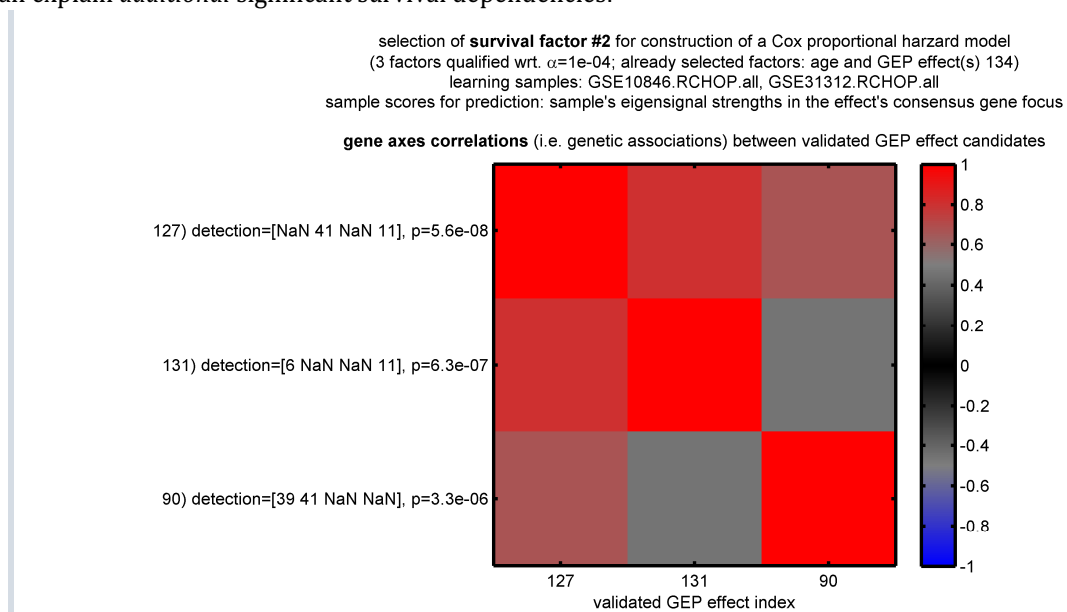


Figure III.2.2.1.b) Selection of the best secondary GEP effect for survival prediction based on all 690 available R-CHOP-treated patients  
 Three GEP effects can explain significant additional survival dependencies ( $p$  value  $\leq 10^{-4}$ , likelihood ratio test). The matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2).

All three qualified effects are genetically similar; the first two have a consensus gene scores correlation of  $r_{(127;131)}^c = 0.80$ , i.e. their top genes overlap strongly. Effect  $\nu = 127$  can explain remaining survival slightly better ( $p_{\nu 127} = 5.6 \cdot 10^{-8}$  opposed to  $p_{\nu 131} = 6.3 \cdot 10^{-7}$ ) and is therefore selected here as second GEP effect for the predictor model.

With only  $p = 7.4 \cdot 10^{-3}$  the COO induced effect  $\nu = 129$ <sup>(III.3.2.2)</sup> follows at rank #13 here (not shown in the plot, as this is no longer significant with respect to  $\alpha = 10^{-4}$ ). This indicates that  $\nu = 134$  can explain most of the survival dependencies explained by  $\nu = 129$ . Vice versa, this is not the case<sup>(cf. Figure III.2.3.1)</sup>.

To find a potential third explanatory effect, again all 135 validated GEP effects are candidate-added to the model, but none is able to significantly explain additional survival dependencies with respect to error threshold  $\alpha = 10^{-4}$ .

Taken together, two validated GEP effects  $v \in \{134, 127\}$  have been selected as predictors and can explain survival independently and on top of age. Removing the DLBCL-unspecific age variable again and fitting a bivariate Cox proportional hazard model  $h_0(t) \cdot \exp(\sum_{v \in \{134, 127\}} \beta_v X_v)$  for described sample scores<sup>(III.2.1.3)</sup> results in listed final statistics<sup>(Table III.2.2.2)</sup> for the two selected GEP effects. (Final  $p$  values are not directly comparable to  $p$  values used during effect selection, since the latter were based on likelihood ratio tests to measure the *additional* explanatory value of an increased model size, rather than absolute predictive capability in presence of all selected predictor variables.)

| GEP effect          | $v = 134$            | $v = 127$           |
|---------------------|----------------------|---------------------|
| $\hat{\beta}_v$     | -0.658               | 0.395               |
| Hazard ratios       | [0.52, 1.93]         | [0.67, 1.48]        |
| $p$ values          | $1.4 \cdot 10^{-16}$ | $1.2 \cdot 10^{-9}$ |
| links to evaluation | III.3.3.1            | III.3.3.2           |

Table III.2.2.2) Bivariate Cox model  $v \in \{134, 127\}$ , final statistics for the R-CHOP training set

$\hat{\beta}_v$  are the fitted Cox coefficients of the log(hazard ratios) for the explanatory variables  $v \in \{134, 127\}$ . Hazard ratio intervals equal  $[\exp(-|\hat{\beta}_v|), \exp(|\hat{\beta}_v|)]$ . They are useful for a comparison of the relative impact on survival explained by different GEP effects. Cox  $p$  values are for individual GEP effects in the final model (not directly comparable to likelihood ratio based  $p$  values for effect selection).

Kaplan-Meier survival estimates for risk intervals based on predicted log(hazard ratios) are used to visualize a spread from 36% to 86% average survival<sup>(Figure III.2.2.2)</sup> in the R-CHOP training set. As expected, this is highly significant ( $p = 3.1 \cdot 10^{-14}$  between the first and last risk interval, log rank test, 100+109 patients). Application of this bivariate predictor to the CHOP-treated validation set proves its generalization capabilities (survival spread from 9.7% to 65% with  $p = 5.2 \cdot 10^{-7}$  between the first and last risk interval, log rank test, 31+31 patients). Compared to standard DLBCL subtypes<sup>(Figure III.2.1.8)</sup>, explanatory GEP effects  $v = 134$  and  $v = 127$  can predict wider survival spreads.

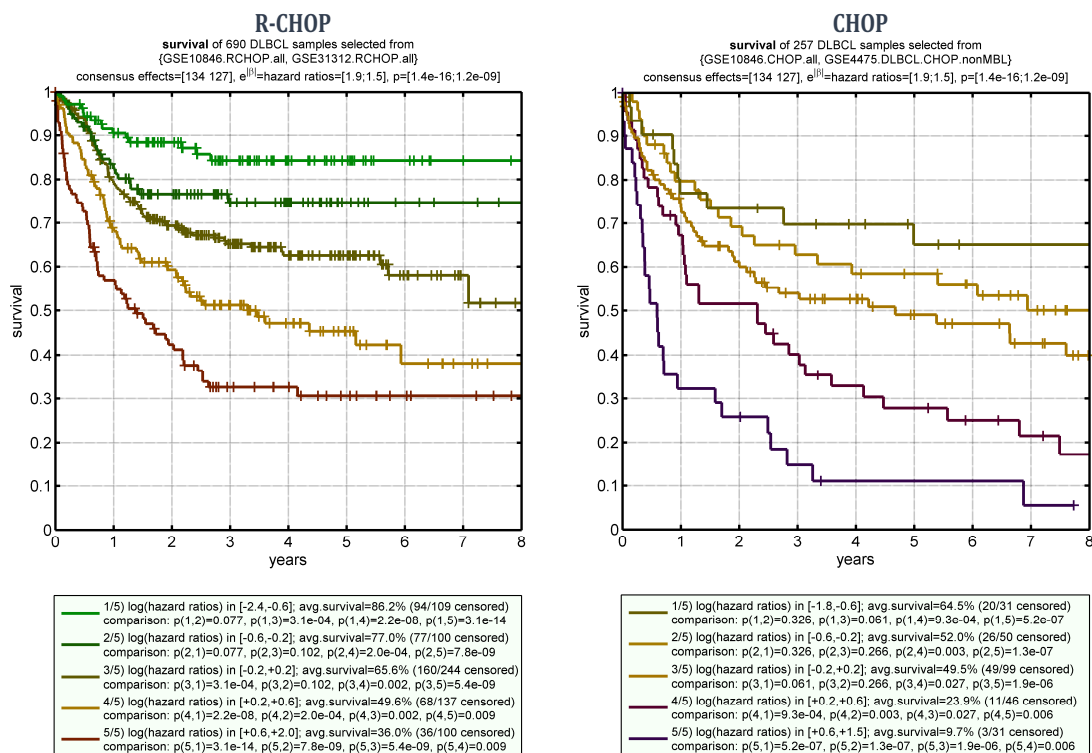


Figure III.2.2.2) Bivariate predictor  $v \in \{134, 127\}$ ; performance in the R-CHOP training set and generalization to the CHOP validation set

Survival predictions for 690 R-CHOP-treated patients (left, training set) and 257 CHOP-treated patients (right, validation set). Chosen split points to present the survival spread in five risk intervals equal multiples -1.5, -0.5, 0.5 and 1.5 of a log(hazard ratio of 150%).

If effect  $\nu = 127$  is used alone in a univariate Cox model, it also shows a “higher expression is better” trend like effect  $\nu = 134$  before, albeit relative weakly and inconsistently with  $p$  only 0.009 (for the R-CHOP training set). Hence, it was no candidate as primary predictor variable with respect to  $\alpha = 10^{-4}$ . Interestingly, after  $\nu = 134$  has been added to the Cox model,  $\nu = 127$  is able to explain much of the observed survival on top of that with  $p = 5.6 \cdot 10^{-8}$  (likelihood ratio test) and with *anti-aligned survival trend* ( $\beta_{134} < 0$  and  $\beta_{127} > 0$ ). Only a multivariate survival analysis can reveal such dependencies.

To elucidate this hierarchical survival dependency, I partition R-CHOP treated patients by their expression of  $\nu = 134$  into negative log(hazard ratios) (i.e. lower risk and relatively favorable outcome) and positive log(hazard ratios) (i.e. higher risk and relatively adverse outcome). Then the explanatory value of adding  $\nu = 127$  to the model is *separately* quantified for each risk partition by  $\nu = 134$ . Again  $\nu = 127$  conveys significant explanatory value for the lower risk partition, even with  $p = 1.0 \cdot 10^{-11}$  despite the reduced sample set (this is better than  $p = 5.6 \cdot 10^{-8}$  before for the full R-CHOP training set). On the other hand, it is not significant for the higher risk partition ( $p = 0.15$ ). This indicates that  $\nu = 134$  can *partition patients cleanly into two biologically distinct phenotypes*.

Fitting a univariate Cox model based only on  $\nu = 127$  to the lower risk partition of  $\nu = 134$  results in  $\hat{\beta}_{127} = 0.59$  ( $p = 1.9 \cdot 10^{-10}$ ); applying this predictor to both partitions visualizes the one-sidedness of this effect (Figure III.2.2.3.a).

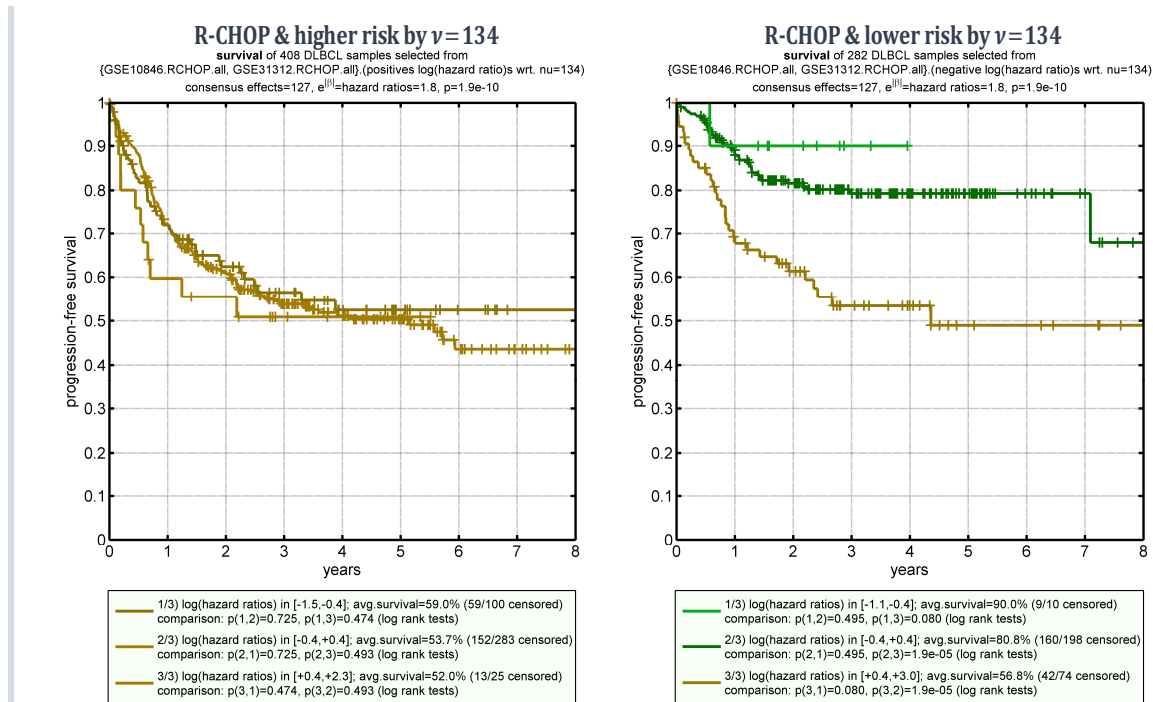


Figure III.2.2.3.a) Univariate predictor based on  $\nu=127$  applied to risk partitions of  $\nu=134$

Survival predictions for 408 R-CHOP-treated patients in the higher risk partition of  $\nu=134$  and for 282 R-CHOP-treated patients in the lower risk partition of  $\nu=134$ . Chosen split points to present the survival spread in three risk intervals equal  $\pm$  log(hazard ratios of 150%).

For a better comparability with previously defined subtypes in DLBCL, I also applied the univariate  $v = 127$  predictor separately to ABC DLBCL and GCB DLBCL subsets of all R-CHOP treated patients (Figure III.2.2.3.b). Here, GCB DLBCL roughly corresponds to the lower risk partition of  $v = 134$ .

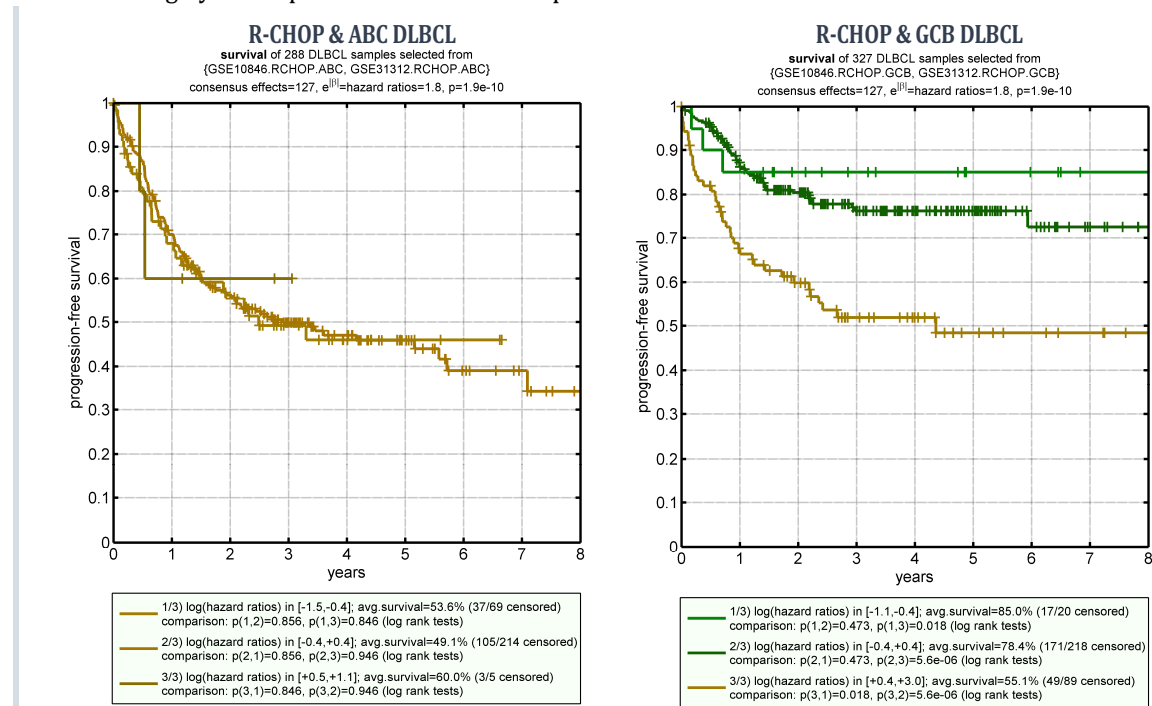


Figure III.2.2.3.b) Univariate predictor based on  $v=127$  separately applied to GCB DLBCL and ABC DLBCL

Survival predictions for 288 R-CHOP-treated patients classified as ABC DLBCL and for 327 R-CHOP-treated patients classified as GCB DLBCL. Again, chosen split points to present the survival spread in three risk intervals equal  $\pm \log(\text{hazard ratios of } 150\%)$ .

As expected, effect  $v = 127$  can explain significantly different survival within GCB DLBCL, while it does not have any significant additional explanatory value for the ABC DLBCL subtype.

#### III.2.2.4

### Predictions within risk partitions of $v = 134$

Knowing that  $v = 127$  only applies to one class of patients and that it displays an opposite survival trend relative to the primary effect, it might further increase predictive performance when applying the bivariate predictor *separately* to these primary classes. (Otherwise expressions of  $v = 127$  for patients in the higher risk partition of  $v = 134$  might wash out their predicted  $\log(\text{hazard ratios})$ .) Partitioning by  $v = 134$  does not make the primary predictor variable completely superfluous, because it is still able to predict some survival dependencies within each partition (a binary partitioning cannot capture its full explanatory value). Still, remaining survival dependencies on  $v = 134$  are weaker after partitioning, and sample counts in the low risk predictor interval for the high risk partition of  $v = 134$  as well as in the high risk predictor interval for the low risk partition of  $v = 134$  are thinned out.

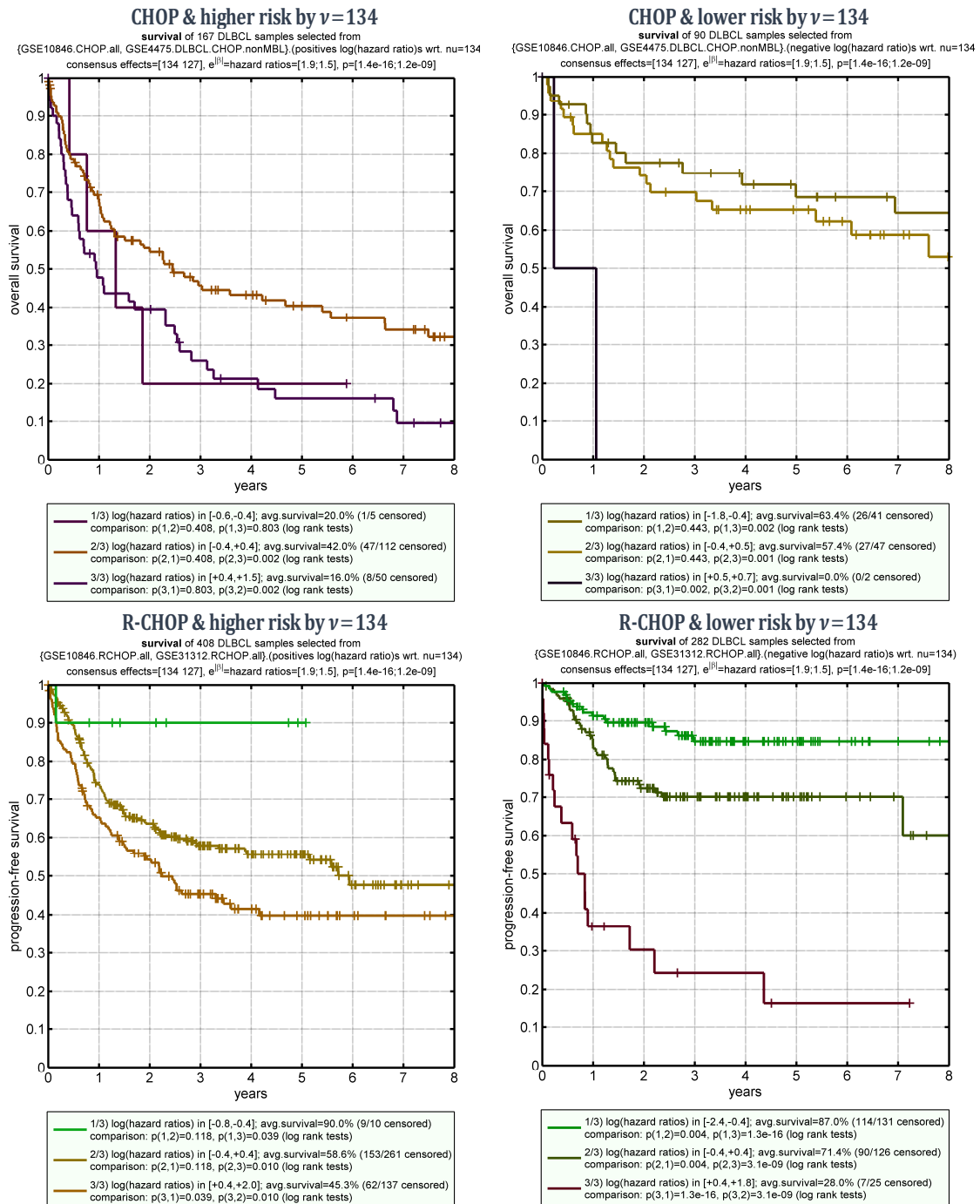


Figure III.2.2.4) Bivariate predictor  $v \in \{134, 127\}$ , predictor performance within risk partitions of  $v=134$

Survival predictions for higher and lower risk partitions of  $v=134$ , separately for R-CHOP and CHOP cohorts. Chosen split points to present the survival spread in three risk intervals equal  $\pm \log(\text{hazard ratios of } 150\%)$ .

Significant survival dependencies remain within both partitions, but most show only a weak spread. In the lower risk partition of  $v = 134$  for R-CHOP therapy however, a subset of DLBCL patients with significantly adverse outcome because of their expression of effect  $v = 127$  exists, despite showing relatively low risk based on  $v = 134$  alone.

This is not clearly reflected following CHOP therapy, which is consistent with univariate survival analyses in GCB DLBCL for  $v = 127$  in individual cohorts. They neither showed significance for the former standard CHOP therapy, but were significant in both R-CHOP treated cohorts(cf. III.3.3.2).



For a better comparability with previously defined subtypes, the bivariate predictor is also applied separately to ABC DLBCL and GCB DLBCL. ABC DLBCL roughly corresponds to the higher risk partition of  $v = 134$  and GCB DLBCL to its lower risk partition.

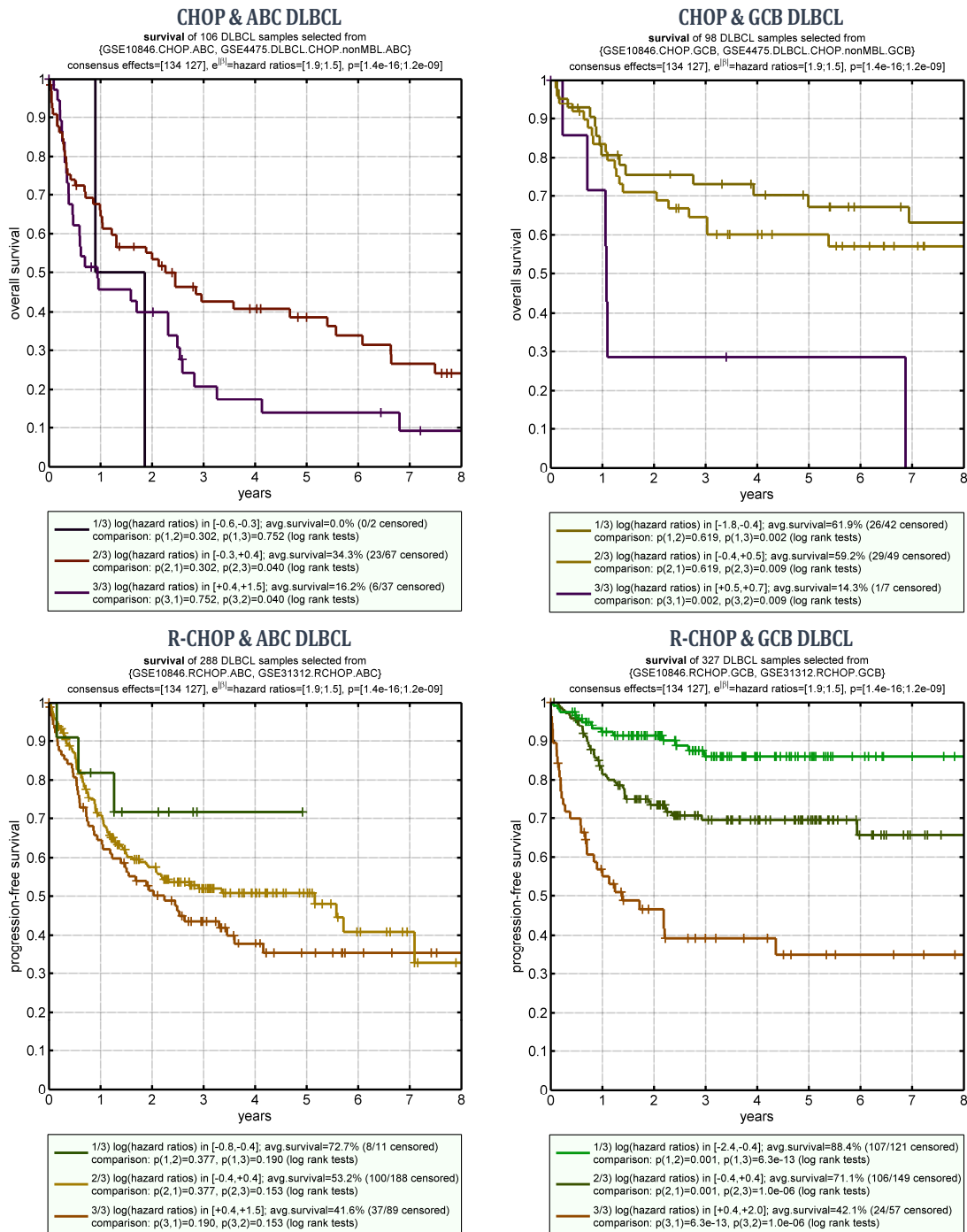


Figure III.2.2.5) Bivariate predictor  $v \in \{134, 127\}$ , performance in subtypes ABC DLBCL and GCB DLBCL

Survival predictions for ABC DLBCL and GCB DLBCL, separately for R-CHOP and CHOP cohorts. Chosen split points to present the survival spread in three risk intervals again equal  $\pm \log(\text{hazard ratios of } 150\%)$ .

While a strong survival dependency in R-CHOP treated GCB DLBCL exists again, survival prediction tends to be more consistent for risk partitions of  $v = 134$  (cf. Figure III.2.2.4) (for example, it displays a wider survival spread compared to R-CHOP & GCB DLBCL and reached significance between neighboring risk intervals when comparing to R-CHOP & ABC DLBCL).

### III.2.3 Bivariate Model for the COO Induced Effect

For a comparison, the bivariate analysis<sup>(III.2.2)</sup> is repeated, this time excluding the top GEP effect  $\nu = 134$  as primary explanatory factor. In this case, the rediscovered COO induced effect  $\nu = 129$ <sup>(cf. III.3.2.2)</sup> is selected. Again, one secondary effect qualifies and no tertiary, using progression events from GSE10846.R-CHOP and GSE31312.R-CHOP as training base.

This alternative bivariate survival model can also explain significant portions of observed patient outcome. However, both models are based on genetically only partially correlated GEP effects. This ambiguity demonstrates that survival dependencies can only be used to filter out unassociated GEP effects, but *several equally valid GEP effects remain* for explanation. In principle, all these effects could contain genes and represent pathways that might impact outcome differences following current therapy and for the molecular pathogenesis of DLBCL. Survival data alone cannot exclusively pinpoint a single GEP effect that is causal for certain. This also implies that the COO induced effect is just one among many<sup>(also see Figure III.2.5.1.a)</sup>.

Unlike before, the best secondary survival effect for the alternate model is shown to affect both GCB DLBCL and ABC DLBCL. This might indicate that the primary COO induced effect is not able to separate patients into biologically distinct survival subtypes as clear as effect  $\nu = 134$ .

#### III.2.3.1 Selection of validated GEP effects as predictors

Again, centered age is prescribed as initial factor in the Cox model to prevent advantages for GEP effects that are related to patient age by accident. As  $\nu = 134$  is excluded,  $\nu = 129$  with  $p = 1.8 \cdot 10^{-9}$  is selected as primary explanatory GEP effect<sup>(cf. Figure III.2.2.1.a)</sup>.

Only three effects<sup>(Figure III.2.3.1)</sup> show additional significant explanatory value and thereby qualify as secondary predictor variable:

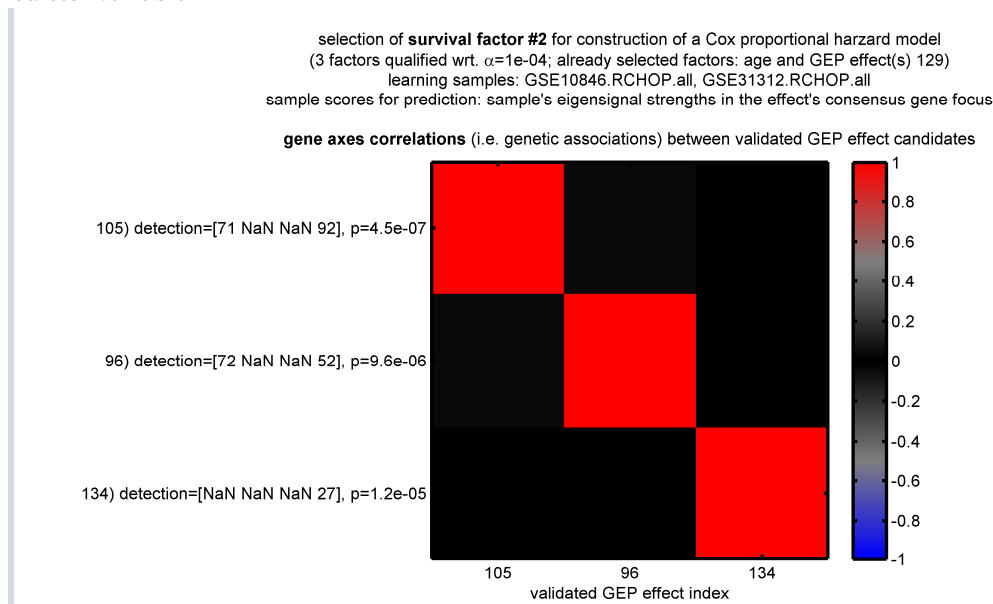


Figure III.2.3.1) Selection of the best secondary GEP effect for survival prediction based on all 690 available samples from R-CHOP-treated patients

Three GEP effects can explain significant additional survival dependencies ( $p$  value  $\leq 10^{-4}$ , likelihood ratio test). The matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2).

The first two effects are dominated by different single genes,  $\nu = 105$  by KIAA1217 and  $\nu = 96$  by FCRL5. Consequently, their consensus gene scores are perpendicular to each other with  $r_{(105;96)}^C = -0.03$ <sup>(cf. Eqn. III.1.3.2.b)</sup> and to the multi-gene effect  $\nu = 134$ .

Even after explaining survival by the COO induced effect  $\nu = 129$ , survival effect  $\nu = 134$  can still explain significant additional survival dependencies with  $p = 1.2 \cdot 10^{-5}$  (rank #3). This was not the case vice versa (cf. Figure III.2.2.1.b). This difference in predictive capabilities is even more pronounced when learning with all samples, rather than only with samples from R-CHOP-treated patients (cf. III.2.5.1). Effect  $\nu = 127$ , i.e. the secondary explanatory variable for the primary bivariate model (cf. III.2.2.2), follows at rank #5 with  $p = 4.65 \cdot 10^{-4}$ , i.e. just over the prescribed error threshold of  $\alpha = 10^{-4}$ .

Effects that are dominated by single gene effects have the advantage of being very specific and the disadvantage of not having the support of a broad co-regulated genetic network. Still, as validated GEP effects they do not represent only noise, as there are several independently measured and correlated probesets pointing to the same gene; furthermore both  $\nu = 105$  and  $\nu = 96$  were unsupervisedly discovered in two independent cohorts.

Consistently applying the selection procedure, I select the most explanatory effect for the model. However, in terms of  $p$  values both single gene effects provide nearly equal additional explanatory value here, i.e. they are practically both equally valid.

None of the 135 validated GEP effects qualifies for addition as tertiary predictor variable on top of  $\nu \in \{129, 105\}$  for the R-CHOP-treated training set (error threshold  $\alpha = 10^{-4}$ ).

### III.2.3.2 Fit results, prediction performance and validation

Taken together, this procedure selects two validated GEP effects  $\nu \in \{129, 105\}$  that can explain survival independently and on top of age. Fitting a bivariate Cox proportional hazard model  $h_0(t) \cdot \exp(\sum_{\nu \in \{129, 105\}} \beta_\nu X_\nu)$  for described sample scores (III.2.1.3) results in final statistics for the two selected GEP effects (Table III.2.3.2).

Comparison with final fit statistics for  $\nu \in \{134, 127\}$  (Table III.2.2.2) already shows that the COO based predictor has larger  $p$  values and can only predict weaker survival dependencies (lower  $|\hat{\beta}_\nu|$ ). To visualize this, again Kaplan-Meier survival estimates for risk intervals based on predicted  $\log(\text{hazard ratios})$  are used (Figure III.2.3.2).

The survival spread following R-CHOP reaches from 45% to 92% ( $p = 3.1 \cdot 10^{-11}$ , log rank test, 84+82 patients) and from 21% to 58% in the CHOP validation set ( $p = 0.001$ , log rank test, 28+33 patients). Though still significant, both survival spreads are more narrow compared to before and prediction is less homogeneous for the inner risk intervals (compare Figure III.2.2.2). However, the top R-CHOP curve has a bit higher average survival compared to before (92.5% instead of 89%; not significant).

| GEP effect          | $\nu = 129$          | $\nu = 105$         |
|---------------------|----------------------|---------------------|
| $\hat{\beta}_\nu$   | 0.525                | -0.321              |
| Hazard ratios       | [0.59, 1.69]         | [0.73, 1.38]        |
| $p$ values          | $2.1 \cdot 10^{-14}$ | $3.0 \cdot 10^{-7}$ |
| links to evaluation | III.3.2.2            | III.3.3.3           |

Table III.2.3.2) Bivariate Cox model  $\nu \in \{129, 105\}$ , R-CHOP training set

$\hat{\beta}_\nu$  denote fitted Cox coefficients of the  $\log(\text{hazard ratios})$  for the explanatory variables  $\nu \in \{129, 105\}$ . Hazard ratio intervals equal  $[\exp(-|\hat{\beta}_\nu|), \exp(|\hat{\beta}_\nu|)]$ . They are useful for a comparison of the relative impact on survival explained by different GEP effects. Cox  $p$  values are for individual GEP effects in the final model (not directly comparable to likelihood ratio based  $p$  values for effect selection).

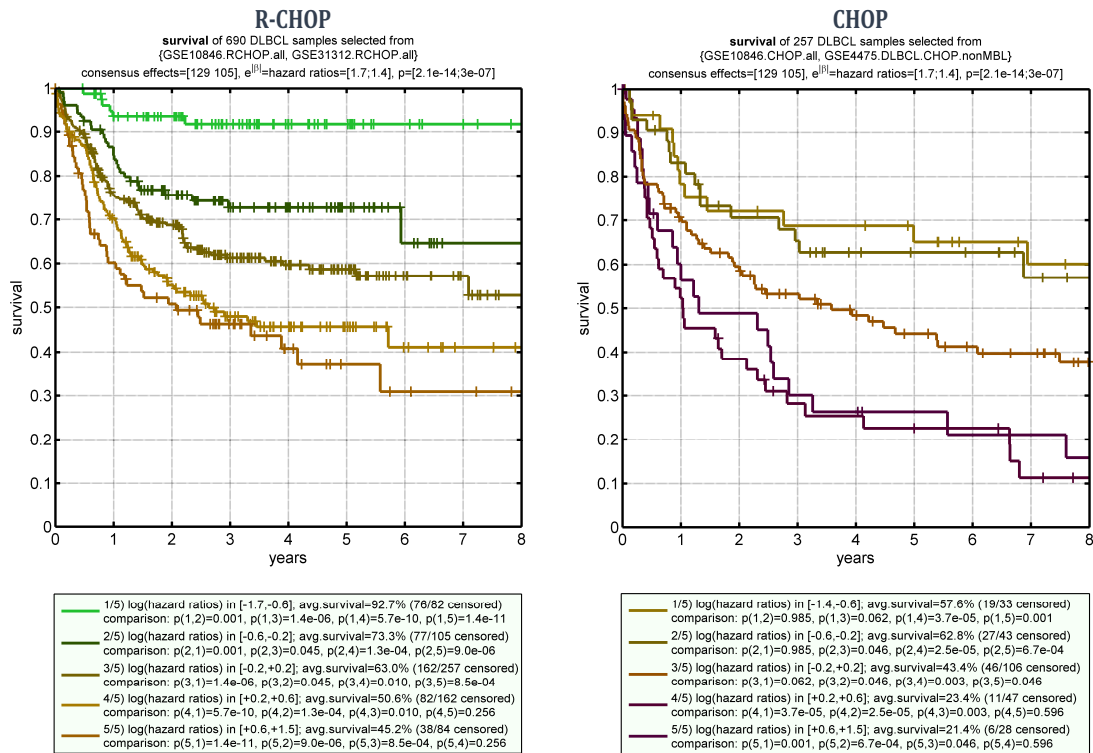


Figure III.2.3.2) Bivariate predictor  $v \in \{129, 105\}$ , performance in the R-CHOP training set and generalization to the CHOP validation set

Survival predictions for 690 R-CHOP-treated patients (left, training set) and 257 CHOP-treated patients (right, validation set). Chosen split points to present the survival spread in five risk intervals equal multiples -1.5, -0.5, 0.5 and 1.5 of a log(hazard ratio of 150%).

### III.2.3.3

## Subtype-specific analysis of $v = 105$

To investigate whether  $v = 105$  only influences one risk partition of  $v = 129$  in a hierarchical fashion (like  $v = 127$  did before<sup>(III.2.2.3)</sup> with respect to  $v = 134$ ), I trained another univariate predictor for  $v = 105$  based on R-CHOP-treated patients from the lower risk partition of  $v = 129$  (i.e. predominantly based on GCB DLBCL samples). This resulted in  $p = 9.7 \cdot 10^{-3}$  and rank #12 for  $v = 105$  only. Training with the higher risk partition resulted still in rank#1 for  $v = 105$ , but only  $p = 1.6 \cdot 10^{-4}$ . This is a weaker association than when training with the full R-CHOP set ( $p = 4.5 \cdot 10^{-7}$ ). In contrast, significance for effect  $v = 127$  increased when fitting in the lower risk partition with respect to  $v = 134$ , despite the reduction in sample size.

Hence, no clear-cut hierarchical survival dependency as seen for risk partitions of  $v = 134$ <sup>(cf. III.2.2.3)</sup> does exist here and outcome for both ABC DLBCL and GCB DLBCL depends on the selected secondary predictor variable  $v = 105$ . With the assumption that patients from the same true biological phenotype show identical survival dependencies over their average gene expressions, this result is another indication besides lower  $p$  values that the predictor based on  $v \in \{134, 127\}$ <sup>(III.2.2)</sup> is a more natural fit of survival data and for subtypes within DLBCL compared to predictors based on the COO classification

Still, as  $v = 105$  is a highly specific effect based exclusively on KIAA1217 and as it can reveal a significant survival dependency on top of the standard COO induced effect, it will also be evaluated in detail<sup>(cf. III.3.3.3)</sup>.

So far, fitted models were trained with samples from R-CHOP treated patients. Alternatively, samples from patients treated with the former standard CHOP therapy can be utilized. Again two effects iteratively qualify as significant explanatory variables with respect to  $\alpha = 10^{-4}$ . While GEP effects selected as primary explanatory variables before are still in the front ranks for the CHOP training set as well, genetically distinct effects have higher explanatory value here.

As all GEP measurements occurred before treatment, it can be assumed that they are similar to the R-CHOP-treated cohorts. Hence, differences in explanatory value of GEP effects are probably caused by the retrospective follow-up information. This indicates that Rituximab did not increase survival uniformly for all DLBCL patients. It also shows again that several genetically distinct GEP effects exist that all may contain genes that are causal for the disease or maybe important to understand its consequences, for example for the tumor microenvironment.

## III.2.4.1

## Selection of validated GEP effects as predictors

Again, centered age is prescribed as initial factor in the Cox model to prevent advantages for GEP effects that are related to patient age by accident. While previously selected primary effects  $v = 134$  and  $v = 129$  can also explain CHOP survival significantly with  $p = 1.3 \cdot 10^{-6}$  respectively  $p = 7.6 \cdot 10^{-5}$ , a group of four genetically highly correlated GEP effects occupies the top ranks, with  $v = 5$  in the lead ( $p = 2.2 \cdot 10^{-8}$ ):

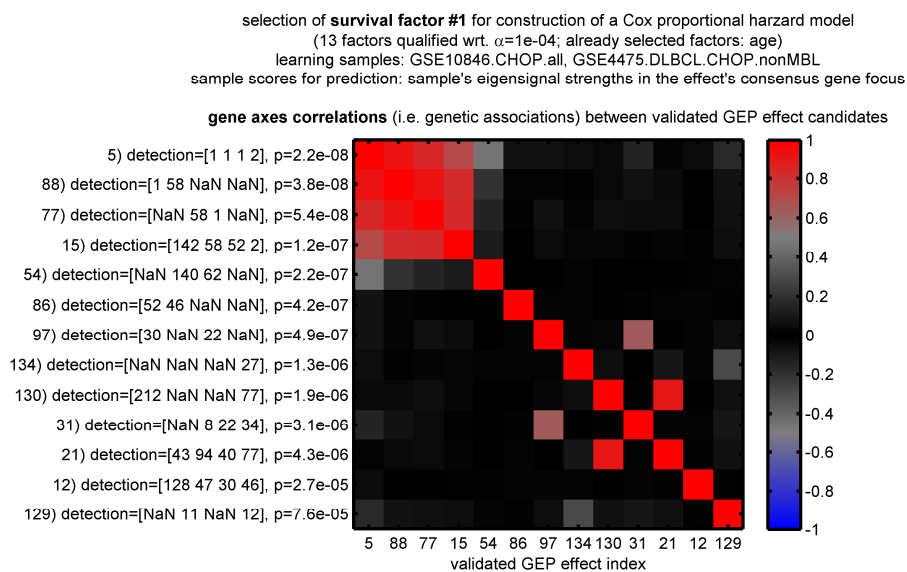


Figure III.2.4.1.a) Selection of the best primary GEP effect for survival prediction based on all 257 available samples from CHOP-treated patients

In total 13 GEP effects can explain significant survival dependencies on top of patient age ( $p$  value  $\leq 10^{-4}$ , likelihood ratio test). The matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2).

Effect  $v = 5$  is interesting, because it contains many genes with strongly differential signal and is detected at rank one in all three cohorts with frozen cell material and at rank two in the FFPE cohort; its biostatistical evaluation revealed<sup>(III.3.3.5)</sup> that it is significantly related to the extracellular matrix and collagen biosynthesis. However, this may also indicate that it is no direct measure of expressions of DLBCL cells.

Only two more GEP effects qualify as secondary explanatory variable for patient outcome following CHOP therapy *after* incorporating  $\nu = 5$  in the model:

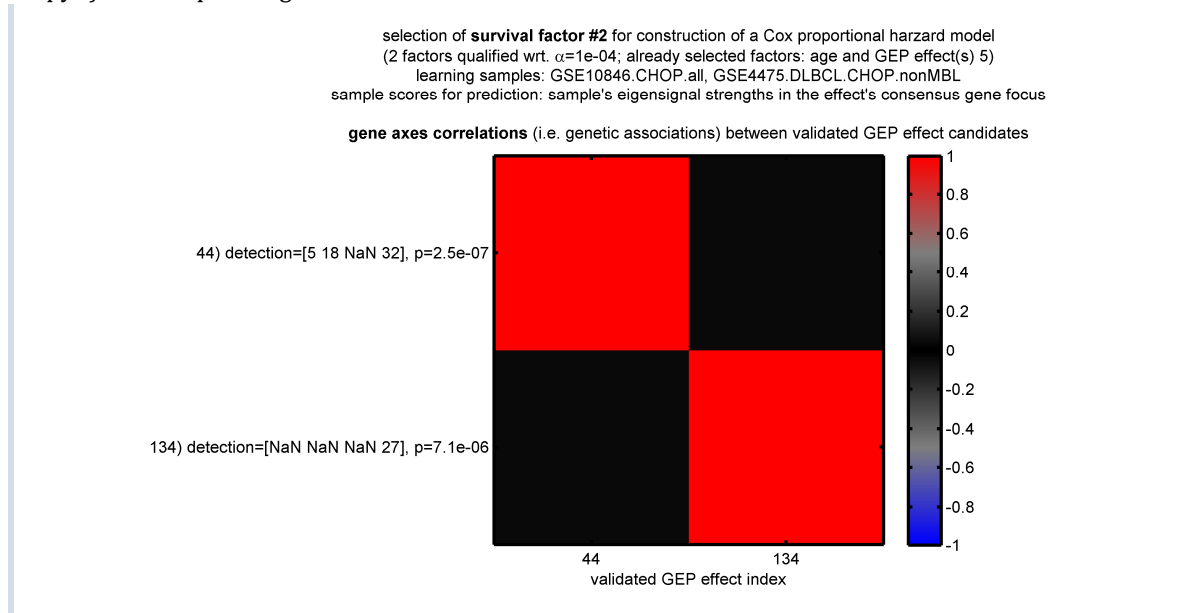


Figure III.2.4.1.b) Selection of the best secondary GEP effect for survival prediction based on all 257 available samples from CHOP-treated patients  
 Two GEP effects can explain significant additional survival dependencies ( $p$  value  $\leq 10^{-4}$ , likelihood ratio test). The matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2).

The primary effect in the best-performing predictor model for R-CHOP ( $\nu = 134$ ) is still significant for CHOP treated patients on top of effect  $\nu = 5$  (on rank #2 with  $p = 7.1 \cdot 10^{-6}$ ); this is another indirect validation of  $\nu = 134$ . In contrast, the COO induced effect  $\nu = 129$  cannot explain any significant additional survival dependencies here ( $p = 0.025$  and rank #32 only).

Effect  $\nu = 44$ <sup>(III.3.3.5)</sup> provides only slightly better explanatory value (with  $p = 2.5 \cdot 10^{-7}$ ) compared to  $\nu = 134$ . Sticking to the selection procedure, it becomes the secondary explanatory variable here.

After incorporating  $\nu = 44$  into the model, no tertiary effect qualifies for this CHOP based predictor.

### III.2.4.2 Fit results, prediction performance and validation

Taken together, two validated GEP effects  $\nu \in \{5, 44\}$  were selected that can explain CHOP survival independently and on top of age. Fitting a bivariate Cox proportional hazard model  $h_0(t) \cdot \exp(\sum_{\nu \in \{5, 44\}} \beta_\nu X_\nu)$  for described sample scores<sup>(III.2.1.3)</sup> results in final statistics for the two selected GEP effects<sup>(Table III.2.4.2)</sup>.

For visualization, again Kaplan-Meier survival estimates for risk intervals based on predicted

| GEP effect          | $\nu = 5$            | $\nu = 44$          |
|---------------------|----------------------|---------------------|
| $\hat{\beta}_\nu$   | -0.723               | 0.475               |
| Hazard ratios       | [0.48, 2.07]         | [0.62, 1.61]        |
| $p$ values          | $9.0 \cdot 10^{-12}$ | $2.6 \cdot 10^{-7}$ |
| links to evaluation | III.3.3.4            | III.3.3.5           |

Table III.2.4.2) Bivariate Cox model  $\nu \in \{5, 44\}$ , CHOP training set

$\hat{\beta}_\nu$  are the fitted Cox coefficients of the log(hazard ratios) for the explanatory variables  $\nu \in \{5, 44\}$ . Hazard ratio intervals equal  $[\exp(-|\hat{\beta}_\nu|), \exp(|\hat{\beta}_\nu|)]$ . They are useful for a comparison of the relative impact on survival explained by different GEP effects. Cox  $p$  values are for individual GEP effects in the final model (not directly comparable to likelihood ratio based  $p$  values for effect selection).

log(hazard ratios) are used. They show (Figure III.2.4.2) a spread from 17% to 63% ( $p = 2.2 \cdot 10^{-9}$ , log rank test, 53+46 patients) of predicted survival in the CHOP training set and from 38% to 77% in the R-CHOP validation set ( $p = 8.3 \cdot 10^{-10}$ , log rank test, 106+106 patients):

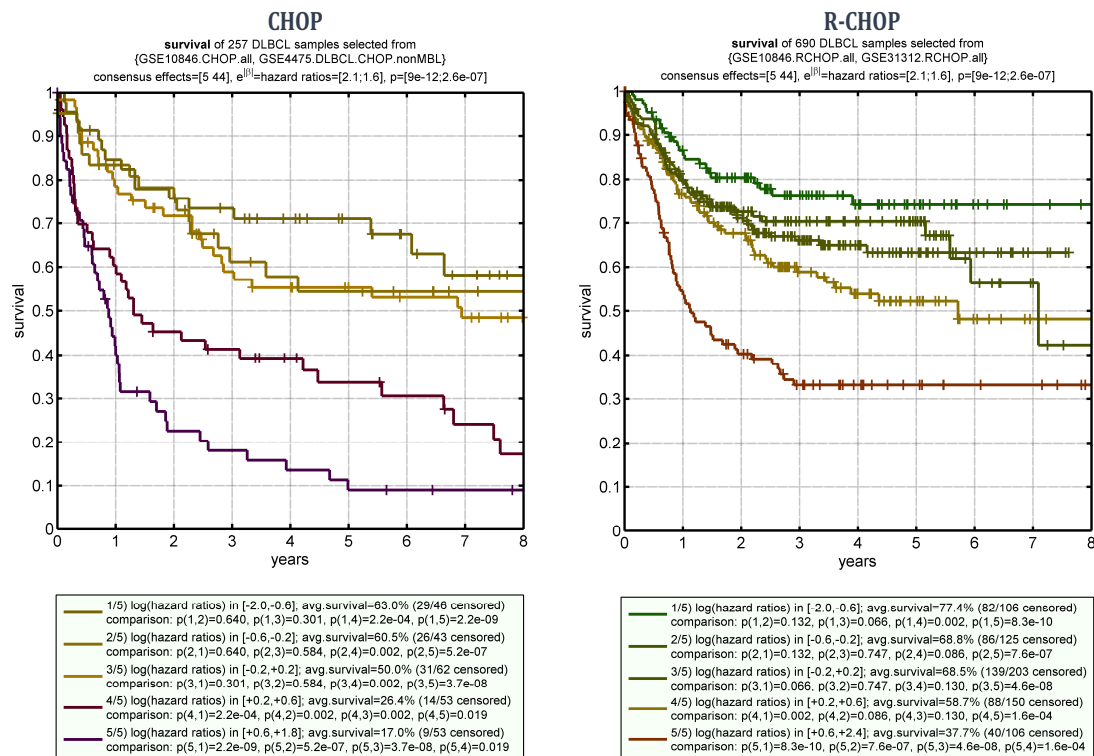


Figure III.2.4.2) Bivariate predictor  $\nu \in \{5,44\}$ ; performance in the CHOP training set and its generalization to the R-CHOP validation set

Survival predictions for 257 CHOP-treated patients (left, training set) and 690 R-CHOP-treated patients (right, validation set). Chosen split points to present the survival spread in five risk intervals again equal multiples -1.5, -0.5, 0.5 and 1.5 of a log(hazard ratio of 150%).

The CHOP-trained predictor clearly validates for samples from R-CHOP treated patients. The highest risk interval containing 106 patients is strongly separated in the R-CHOP validation cohort ( $p = 1.6 \cdot 10^{-4}$  to the neighboring lower risk interval). However, survival of most R-CHOP treated patients cannot be resolved as homogeneously by the CHOP-trained predictor as by the best R-CHOP trained predictor (cf. Figure III.2.2.2). The survival spread in the CHOP *training* cohort is even slightly lower compared to the predicted spread for the same patients by the best-performing R-CHOP-trained predictor (for identical risk intervals).

Taken together, the best R-CHOP trained predictor (III.2.2) works approximately as well for CHOP as the CHOP-trained predictor, but is more powerful and resolves survival more homogeneously for the current standard therapy. While significant GEP effects that qualified for the CHOP predictor may also contribute insightful information about molecular causes of DLBCL, the R-CHOP trained model should be preferred for predictive purposes. In part, this could be expected, as the R-CHOP training set consists of 690 events opposed to only 257 available events for CHOP-treated patients.

### III.2.4.3

## Hierarchical survival analysis of $\nu = 44$

Partitioning CHOP-treated patients by the selected primary predictor variable  $\nu = 5$  (at baseline eigensignal), shows that effect  $\nu = 44$  can only predict significant additional survival dependencies for the high risk partition of  $\nu = 5$  (still at rank #1 with  $p = 1.24 \cdot 10^{-6}$ ), while for the lower risk partition of  $\nu = 5$  it only ranks #12<sup>th</sup> with  $p = 0.024$ .

Similar to analyzing  $\nu = 127$  in context of  $\nu = 134$  (cf. Figure III.2.2.3.a), this hierarchical survival dependency can be visualized by Kaplan Meier survival estimates for risk intervals of effect  $\nu = 44$  in risk partitions by  $\nu = 5$ . This also confirms the hierarchical survival dependency in the R-CHOP validation set (Figure III.2.4.3).

Counting patients outside of the respective baseline risk interval, approximately 36/131 $\approx$ 27% of CHOP-treated and 98/384 $\approx$ 26% of R-CHOP-treated patients in the high risk partition of  $\nu = 5$  seem to be influenced by this hierarchical effect:

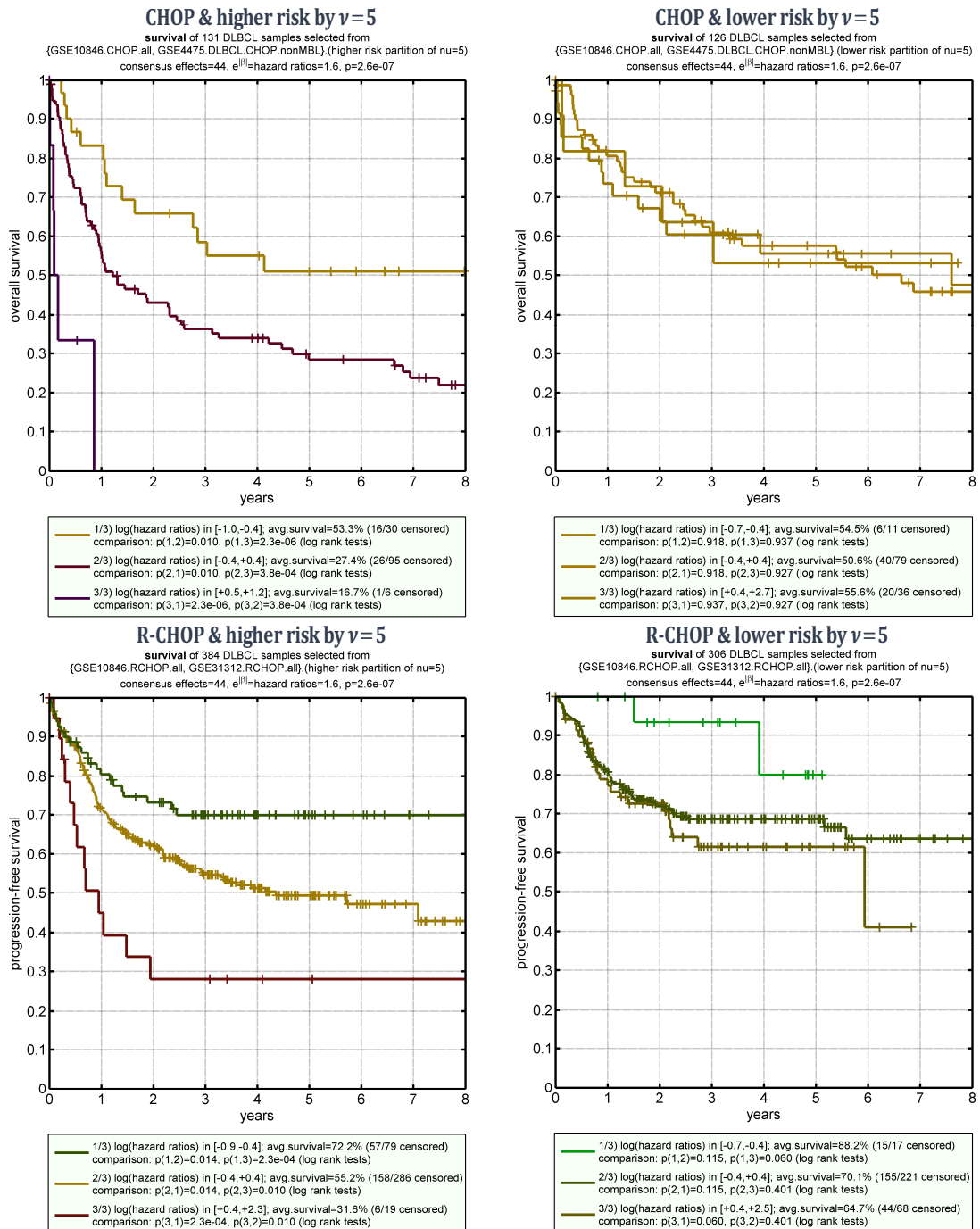


Figure III.2.4.3) Univariate predictor based on  $\nu=44$  applied to risk partitions of  $\nu=5$

For CHOP-treated patients (upper panels), survival predictions show a dependency that exclusively exists in the higher risk partition of effect  $\nu=5$ . This is confirmed in the R-CHOP-treated validation set (lower panel).

Chosen split points to present the survival spread in three risk intervals equal  $\pm \log(\text{hazard ratios of } 150\%)$ .



The majority of patients in the higher risk partition of  $\nu = 5$  are classified as ABC DLBCL. Hence the same hierarchical survival dependency might still exist when applying the univariate predictor based on  $\nu = 44$  to DLBCL subtypes separately.

While the trend is still noticeable (Figure III.2.4.4), survival differences are washed out. Hence,  $\nu = 44$  probably does not stand in a biological hierarchical relation to the COO induced effect, but only to effect  $\nu = 5$ .

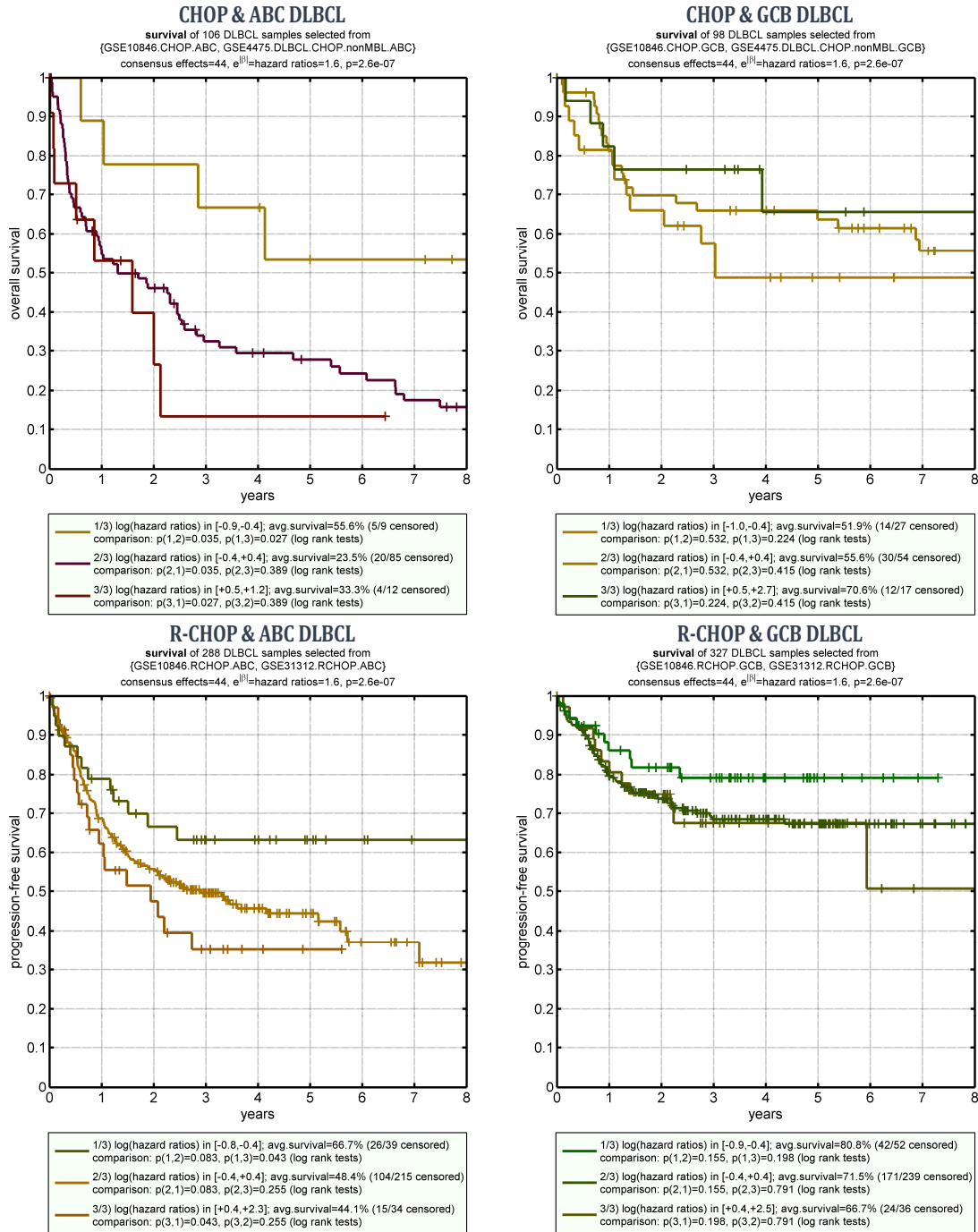


Figure III.2.4.4) Univariate predictor based on  $\nu = 44$  applied to DLBCL subtypes

The hierarchical survival dependency (cf. III.2.4.3) is still visible when splitting by subtypes, because ABC DLBCL patients are overrepresented in the higher risk partition of  $\nu = 5$ ; however, it is washed out. Chosen split points to present the survival spread in three risk intervals again equal  $\pm \log(\text{hazard ratios of } 150\%)$ .

To extract the maximum information from available data, a Cox model is trained now with all available samples. For the same significance threshold  $\alpha = 10^{-4}$ , five genetically independent GEP effects qualify for selection and can explain different dependencies of observed patient outcome. The resulting quinvariate model for survival prediction is validated by leave-one-out validation.

The model confirms the predictive advantage of  $v = 134$  (rank #1 as primary factor,  $p = 4.5 \cdot 10^{-17}$ ) over the COO induced effect ( $v = 129$ , rank #2 as primary factor,  $p = 1.1 \cdot 10^{-12}$ ).

Prediction performance based on leave-one-out validation shows a homogeneous spread from 29% to 89% average survival<sup>(cf. Figure III.2.5.3.b)</sup>. Comparison with outcome differences predicted by standard DLBCL subtypes<sup>(III.2.1.8)</sup> shows an obvious and strong increase in predicted survival spreads, both for CHOP and R-CHOP therapy. Significant survival differences are also predicted *within* standard subtypes<sup>(cf. Figure III.2.5.6)</sup> and within risk partitions of the primary survival effect  $v = 134$ <sup>(cf. Figure III.2.5.7)</sup>. Additionally, significantly different survival within risk classes by international prognostic index<sup>(cf. Figure III.2.5.8)</sup> is predicted.

Together, these results indicate that GEP effects selected as predictor variables in the quinvariate model probably contain novel and not yet molecularly understood mechanisms with significant impact on DLBCL patient outcome and that known standard DLBCL subtypes are intrinsically heterogeneous. Therefore, all selected GEP effects will be biostatistically evaluated in III.3.3.

To prevent selection advantages for GEP effects that are accidentally associated with DLBCL-unspecific factors, centered age is prescribed as first predictor variable again; it explains survival dependencies with  $p = 5.5 \cdot 10^{-7}$  over all patients. Additionally and as two cohorts were treated with the former standard CHOP chemotherapy, therapy is prescribed as second predictor variable ( $p = 4.1 \cdot 10^{-4}$  over all patients). This also prevents finding GEP effects that are accidentally associated with the type of follow-up event.

For  $\alpha = 10^{-4}$ , 21 GEP effects show<sup>(Figure III.2.5.1.a)</sup> significant additional explanatory value. Like for the best-performing bivariate predictor that was only trained with R-CHOP events<sup>(III.2.2)</sup>, effect  $v = 134$ <sup>(III.3.3.1)</sup> is again the most explanatory variable for observed patient outcome ( $p = 4.5 \cdot 10^{-17}$ ). On rank #2 and with a distance, the well-known COO induced effect<sup>(III.3.2.2)</sup> follows ( $p = 1.1 \cdot 10^{-12}$ ). The top effect for the CHOP-trained bivariate model ( $v = 5$ ) follows at rank #3 ( $p = 2.6 \cdot 10^{-9}$ ). Other effects already seen during factor selection for CHOP-trained<sup>(Figure III.2.4.1.a)</sup> and R-CHOP-trained<sup>(Figure III.2.2.1.a)</sup> models follow, plus some effects that have not been revealed when learning with a reduced sample base.

It should be stressed again that *all* these unsupervisedly detected GEP effects are hereby validated on survival level, i.e. in principle all of them could contain genes that are causal for the disease, not just the strongest ones selected for the predictor model.

selection of **survival factor #1** for construction of a Cox proportional hazard model  
 (21 factors qualified wrt.  $\alpha=1e-04$ ; already selected factors: age, therapy)  
 learning samples: GSE10846.CHOP.all, GSE10846.RCHOP.all, GSE4475.DLBCL.CHOP.nonMBL, GSE31312.RCHOP.all  
 sample scores for prediction: sample's eigensignal strengths in the effect's consensus gene focus

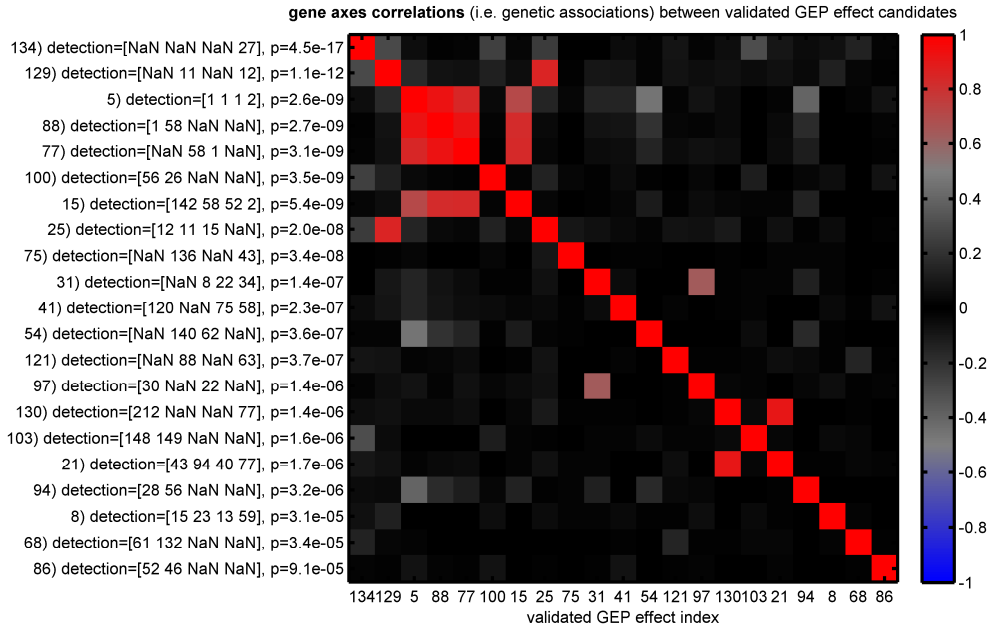


Figure III.2.5.1.a) Selection of the primary GEP effect for survival prediction based on all 947 available events  
 In total, 21 GEP effects can explain significant survival dependencies on top of age and therapy ( $p$  value  $\leq 10^{-4}$ , likelihood ratio test). The matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2).

With  $v = 134$  as primary explanatory variable in the predictor, 16 GEP effects can provide significant explanatory value on top of that (Figure III.2.5.1.b). As before (Figure III.2.2.1.b), the COO induced effect  $v = 129$  is no longer significant as secondary factor ( $p = 5.3 \cdot 10^{-3}$  and rank #35 only), as  $v = 134$  can already explain most survival dependencies explained by it. Vice versa, if  $v = 129$  was selected as primary factor here (not plotted),  $v = 134$  would still be significant as secondary explanatory variable (with  $p = 1.4 \cdot 10^{-7}$  on rank #2), again demonstrating that it is the *more natural choice* as primary DLBCL survival effect.

selection of **survival factor #2** for construction of a Cox proportional hazard model  
 (16 factors qualified wrt.  $\alpha=1e-04$ ; already selected factors: age, therapy and GEP effect(s) 134)  
 learning samples: GSE10846.CHOP.all, GSE10846.RCHOP.all, GSE4475.DLBCL.CHOP.nonMBL, GSE31312.RCHOP.all  
 sample scores for prediction: sample's eigensignal strengths in the effect's consensus gene focus

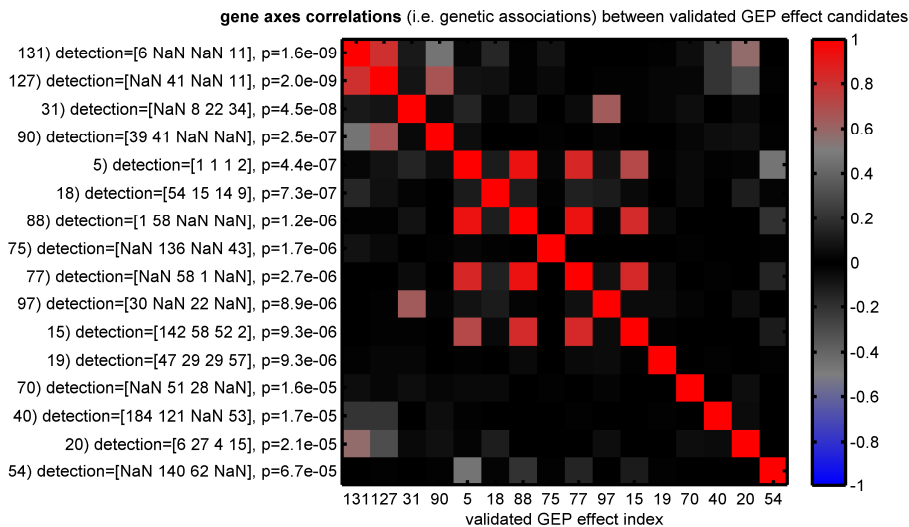


Figure III.2.5.1.b) Selection of the secondary GEP effect for survival prediction based on all 947 available events  
 Training with all samples, 16 validated GEP effects can explain significant additional survival dependencies ( $p$  value  $\leq 10^{-4}$ , likelihood ratio test). The matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2).

As for the R-CHOP-based model (Figure III.2.2.1.b), the genetically correlated effects  $\nu = 131$  and  $\nu = 127$  occupy ranks #1 and #2 for the secondary explanatory variable, but this time  $\nu = 131$  is slightly in the lead ( $p = 1.6 \cdot 10^{-9}$  instead of  $p = 2.0 \cdot 10^{-9}$ ). The correlation of their consensus gene scores equals  $r_{(127;131)}^c = 0.80$  (cf. Eqn. III.1.3.2.b), i.e.  $\nu = 127$  and  $\nu = 131$  are genetically similar and also share many top genes like IRF4 and BATF, as they are based on the same discovered effect in one of the four DLBCL cohorts. Consistently applying the selection procedure,  $\nu = 131$  becomes the secondary explanatory variable for the predictor trained with all samples.

Like before, two effects can explain the strongest survival trends. However, using all samples as learning set has the power to reveal some additional genetically independent effects with significant explanatory value, albeit their influence on survival is not as strong as for the first two variables. Probably they play a role only relatively small patient subsets. The best tertiary explanatory variable is the quad-discovered effect  $\nu = 19$  (with  $p = 7.5 \cdot 10^{-5}$ ). Effect  $\nu = 75$  follows at rank #2 (with  $p = 9.1 \cdot 10^{-5}$ ):

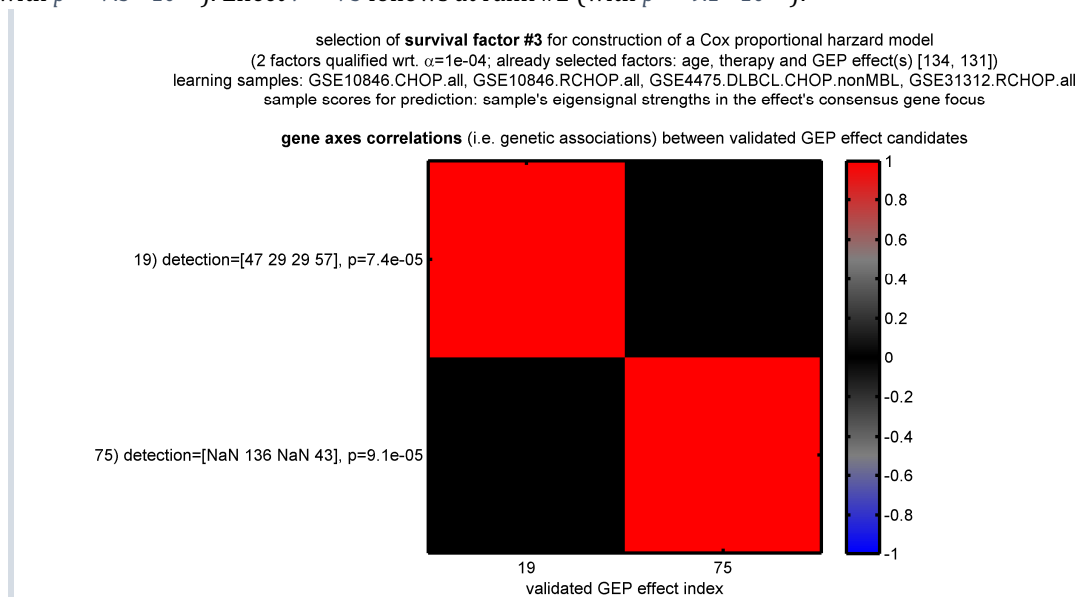


Figure III.2.5.1.c) Selection of the tertiary GEP effect for survival prediction based on all 947 available events

Shown are two validated GEP effects with a  $p$  value  $\leq 10^{-4}$  (likelihood ratio test) for their additional explanatory value of observed patient outcome. Again, the matrix shows whether they are genetically associated with each other (correlations of their consensus gene scores; cf. III.1.3.2).

In selection iteration four,  $\nu = 75$  is still significant and now at rank #1 (with  $p = 3.6 \cdot 10^{-5}$ ). This indicates that it explains different survival dependencies compared to the last selected effect  $\nu = 19$ ; otherwise, it would no longer be significant. The  $p$  value for  $\nu = 75$  even decreases a bit after adding  $\nu = 19$  to the model (this can happen if already selected effects can stratify samples in a way such that additional independent survival dependencies are revealed, as has been seen (III.2.2.1) for  $\nu = 127$  on top of  $\nu = 134$ ).

In selection iteration five, another quad-discovered effect qualifies ( $\nu = 3$  with  $p = 2.5 \cdot 10^{-5}$ ).

After selecting it, no more GEP effect can convey any additional significant explanation of observed patient outcome (relative to the chosen significance threshold of  $\alpha = 10^{-4}$ ).

### III.2.5.2

### Fit results

Taken together, the selection procedure yields five validated GEP effects  $\nu \in \{134, 131, 19, 75, 3\}$  that can explain survival independently and on top of DLBCL-unspecific factors age and therapy. All of them will be

biostatistically evaluated and genetically discussed in detail<sup>(III.3.3)</sup>. Fitting a Cox proportional hazard model  $h_0(t) \cdot \exp(\sum_{v \in \{134, 131, 19, 75, 3\}} \beta_v X_v)$  with all five selected effects for described sample scores<sup>(III.2.1.3)</sup> results in the following final statistics:

| GEP effect         | $\nu = 134$          | $\nu = 131$         | $\nu = 19$          | $\nu = 75$          | $\nu = 3$           |
|--------------------|----------------------|---------------------|---------------------|---------------------|---------------------|
| $\hat{\beta}_\nu$  | -0.739               | 0.421               | -0.287              | 0.232               | 0.257               |
| Hazard ratios      | [0.48, 2.09]         | [0.65, 1.52]        | [0.75, 1.33]        | [0.79, 1.26]        | [0.77, 1.29]        |
| $p$ values         | $2.8 \cdot 10^{-22}$ | $1.1 \cdot 10^{-9}$ | $2.5 \cdot 10^{-7}$ | $9.1 \cdot 10^{-7}$ | $6.3 \cdot 10^{-6}$ |
| link to evaluation | III.3.3.1            | III.3.3.2           | III.3.3.6           | III.3.3.7           | III.3.3.8           |

Table III.2.5.2) Quivariate Cox model, final statistics for the complete training set based on available samples and follow-up data for all DLBCL patients

$\hat{\beta}_\nu$  are the fitted Cox coefficients of the log(hazard ratios) for the five explanatory variables. Hazard ratio intervals equal  $[\exp(-|\hat{\beta}_\nu|), \exp(|\hat{\beta}_\nu|)]$ . They are useful for a comparison of the relative impact on survival explained by different GEP effects. Cox  $p$  values are for individual GEP effects in the final model (not directly comparable to likelihood ratio based  $p$  values for effect selection).

Like before<sup>(Table III.2.2.2)</sup>, the first two effects are associated with opposing survival trends, as  $\text{sign}(\hat{\beta}_{134}) = -\text{sign}(\hat{\beta}_{131})$ . As consensus gene scores of effects  $\nu = 131$  and  $\nu = 127$  are strongly correlated ( $r_{(127,131)}^c = 0.80$ ), effect  $\nu = 131$  stands in a similar hierarchical context<sup>(cf. III.2.2.3)</sup> to  $\nu = 134$ . Signs of  $\hat{\beta}_\nu$  are also different for other variates and further hierarchical dependencies might exist (for example, effects  $\nu \in \{19, 75, 3\}$  might only affect some 2D risk partitions by both  $\nu = 134$  and  $\nu = 131$ ). One additional hierarchical dependency of  $\nu = 3$  in partitions by  $\nu \in \{19, 75\}$  is illuminated in III.3.3.8.

### III.2.5.3 Leave-one-out cross-validation and predictor performance

As the model has been trained using all available samples, no validation set remains. Therefore, leave-one-out cross-validation is employed to test the predictive performance of the quivariate model.

For the five selected GEP effects, I fit  $j = 1 \dots 947$  models of type  $h_{0j}(t) \cdot \exp(\sum_{v \in \{134, 131, 19, 75, 3\}} \beta_{jv} X_{jv})$ , based on 946 samples each. For example  $X_{1,134}$  are eigensignal strengths in effect  $\nu = 134$  for all patients  $j = 2 \dots 947$ , except for the first one. Each fit is then used to predict the log(hazard ratio) for the left-out sample only. I.e. for  $j = 1$  the prediction reads  $\sum_{v \in \{134, 131, 19, 75, 3\}} \hat{\beta}_{1v} x_{1v}$ , where  $\hat{\beta}_{1v}$  are the fitted coefficients without using  $j = 1$  and  $x_{1v}$  are the eigensignal strengths in all five GEP effects for the left-out patient  $j = 1$ .

Taken together, this procedure results in predictions for every sample, but *never uses a particular sample for its own prediction*. Resulting distributions of  $\hat{\beta}_\nu$  over all 947 fits are tight, which already indicates an effective generalization performance:

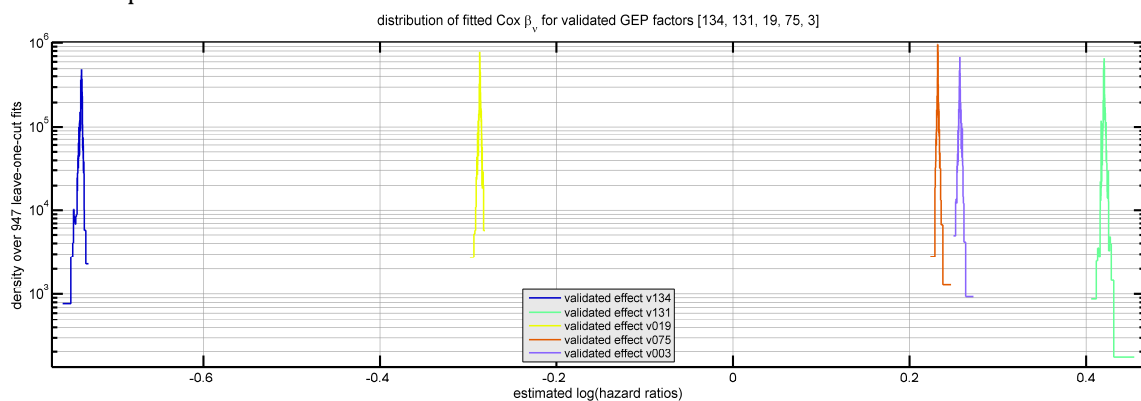


Figure III.2.5.3.a) Distributions of  $\hat{\beta}_\nu$  over all 947 leave-one-out fits

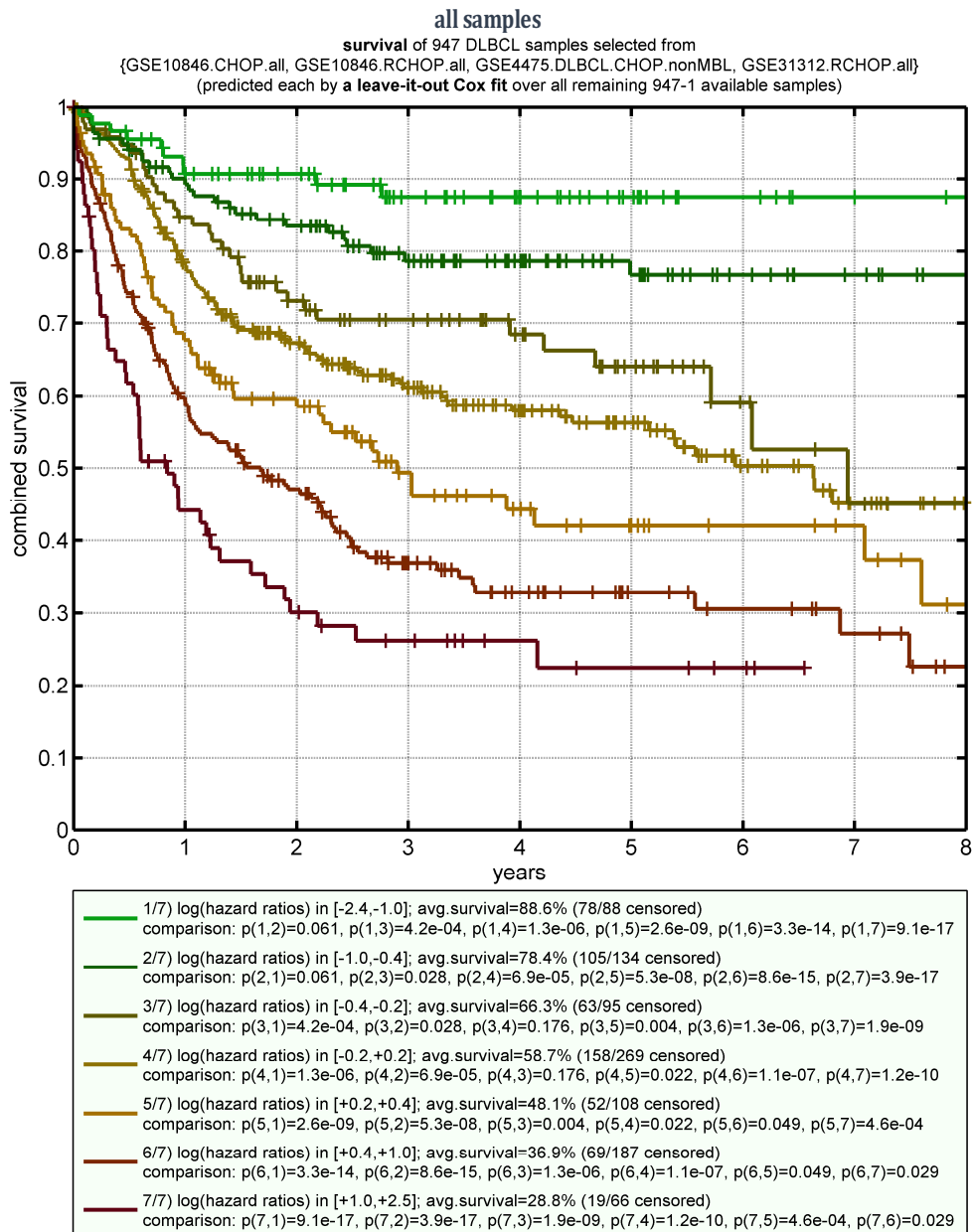


Figure III.2.5.3.b) Quinvariate predictor performance for all 947 DLBCL samples based on leave-one-out validation

Leave-one-out survival predictions for all available 947 available patients using the quinvariate model. Chosen split points to present the survival spread in seven risk intervals equal multiples -2.5, -1, -0.5, 0.5, 1 and 2.5 of a log(hazard ratio of 150%).

To test and quantify predictive performance, seven risk intervals based on predicted  $\log(\text{hazard ratios})$  from leave-one-out validation have been visualized by Kaplan-Meier survival estimates and compared via log rank tests. Resulting survival curves (Figure III.2.5.3.b) show that it is possible to predict a survival spread from approximately 29% to 89% with the quinvariate model and to resolve survival homogeneously in-between. Log rank test  $p$  values in the legend compare all seven risk intervals to each other; top and bottom risk intervals are highly separate with  $p_{1,7} = 9.1 \cdot 10^{-17}$ , thoroughly validating the predictive performance of the quinvariate GEP based predictor.

As samples originate from four independent patient cohorts, it is also safe to assume that no systematic survival bias can exist that is specific only to all four analyzed DLBCL cohorts, but not to DLBCL in general. Hence, this leave-one-out-validated quinvariate predictor shows a generalization performance that is very promising for its application to future DLBCL cohorts.

To double-check that the predictor does not just predict survival differences because of therapy, and to show that it is applicable to both CHOP- and R-CHOP-treated patients as well as to both overall and progression-free survival, I also applied it separately to the CHOP- and R-CHOP-treated cohorts (Figure III.2.5.4). Clearly, predicted outcome for patients is significantly different again; top and bottom survival curves are separated with  $p_{1,5} = 1.1 \cdot 10^{-8}$  for 257 CHOP-treated patients and  $p_{1,3} = 6.3 \cdot 10^{-16}$  for 690 R-CHOP-treated patients, respectively.

Comparing this result with survival dependencies predicted by standard subtypes (III.2.1.8) shows an obvious and strong increase in predicted survival spreads, both for CHOP and R-CHOP. (This could in part be expected, as subtype classification is based on just one relatively heterogeneous GEP effect (see e.g. [29], figure 3 for GSE31312), rather than being based on five genetically distinct correlation-based effects.)

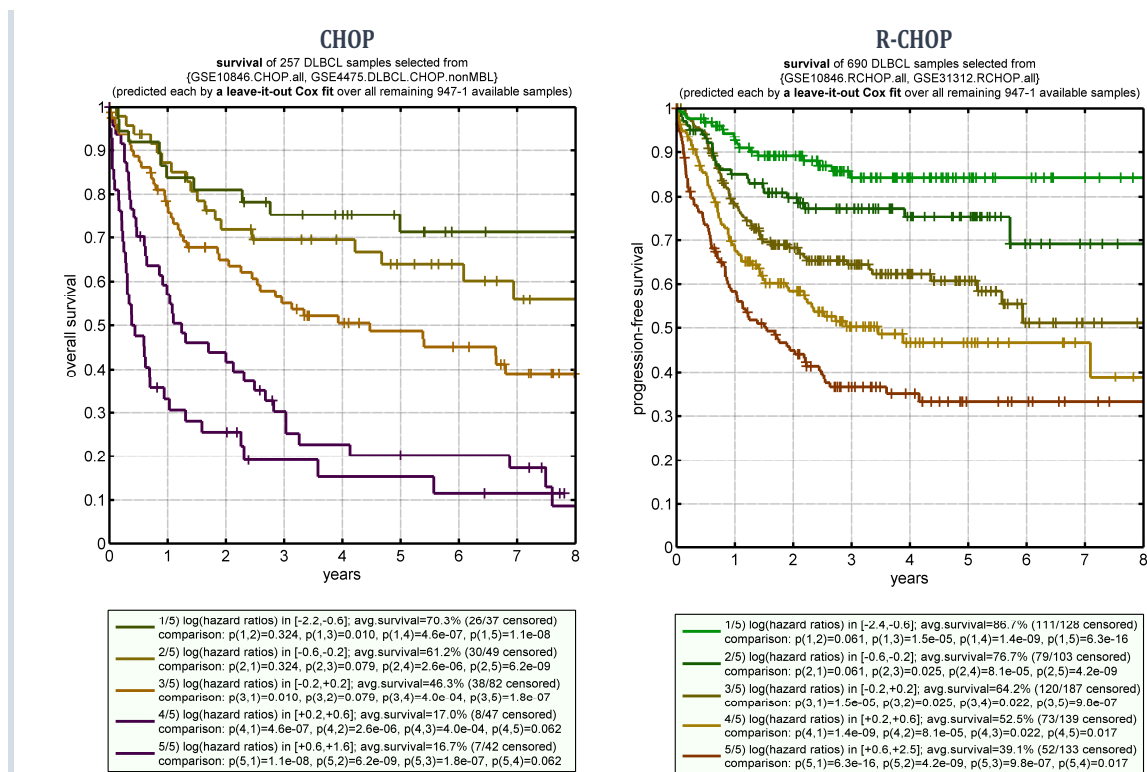


Figure III.2.5.4) Quinvariate predictor performance for all CHOP-treated and R-CHOP-treated cohorts

Quinvariate leave-one-out survival predictions for 257 CHOP-treated patients (left) and 690 R-CHOP-treated patients (right). Chosen split points to present the survival spread in five risk intervals equal multiples -1.5, -0.5, 0.5 and 1.5 of a log(hazard ratio of 150%).

Interestingly, the survival spread between ABC DLBCL and GCB DLBCL was considerably lower for GSE31312 as for GSE10826 when comparing both R-CHOP treated cohorts (cf. Figure III.2.1.8).

One difference between these cohorts is that GSE31312 is based on FFPE samples (rather than based on fresh frozen cell material). Hence, it is conceivable that this difference might cause that the standard classification into subtypes is less strongly associated with outcome in GSE31312. However, this is probably not the case, because survival spreads between top and bottom risk intervals that were predicted by the quivariate model are very similar for both R-CHOP treated cohorts (Figure III.2.5.5). This might indicate that GSE10846-based gene signatures that are currently used for subtype classification (cf. Figure III.3.2.2.c) are too cohort-specific do not generalize well, making subtype classifications difficult to reproduce for new DLBCL cohorts (also read III.3.2.2), irrespective of their underlying cell material.

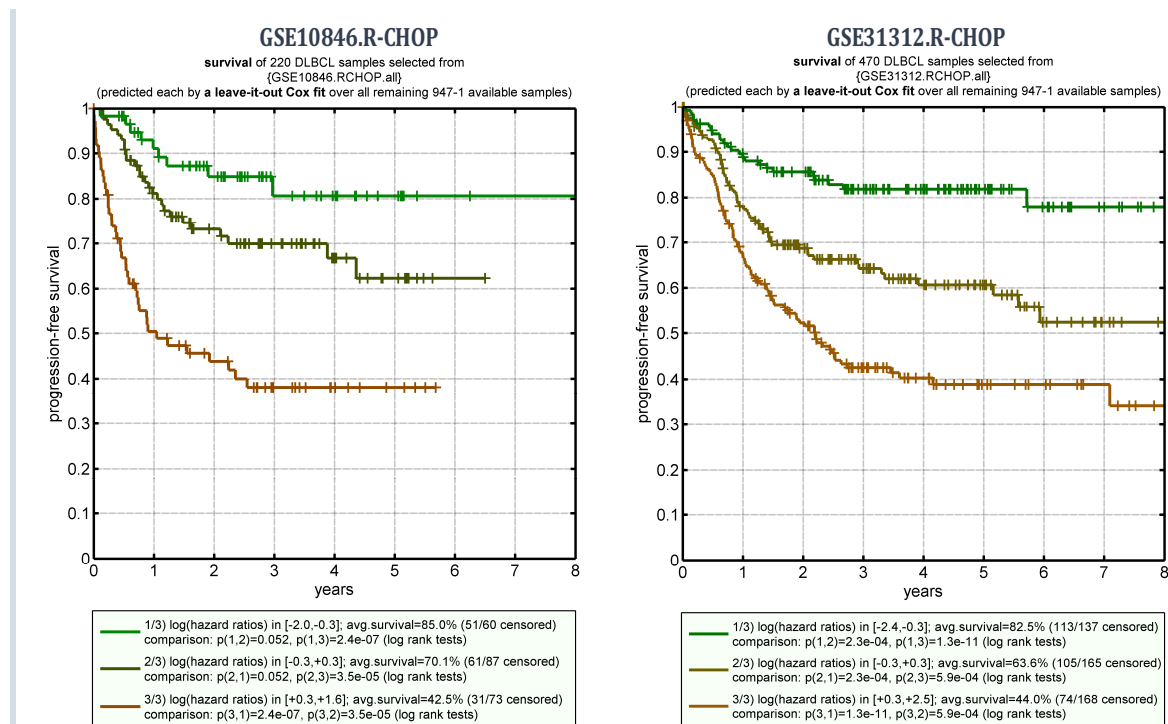


Figure III.2.5.5) Quivariate predictor performance, separately for frozen cell material and FFPE based R-CHOP cohorts

Quivariate leave-one-out survival predictions for 220 patients from GSE10846.R-CHOP (frozen cell material, left) and 470 patients from GSE31312.R-CHOP (FFPE, right). Chosen split points to present the survival spread in three risk intervals equal  $\pm \log(\text{hazard ratios of } 133\%)$ .



To demonstrate that the quivariate predictor indeed reveals survival dependencies beyond those already known by standard subtype classification, I also applied it separately to ABC DLBCL samples and GCB DLBCL samples (Figure III.2.5.6). As expected, sample counts in risk intervals with negative  $\log(\text{hazard ratios})$  are thinned out for ABC DLBCL and sample counts in risk intervals with positive  $\log(\text{hazard ratios})$  are thinned out for GCB DLBCL, reflecting the known difference in average subtype survival. Still, significant survival dependencies remain on top of this, for both ABC DLBCL and GCB DLBCL and within both CHOP and R-CHOP treated cohorts:

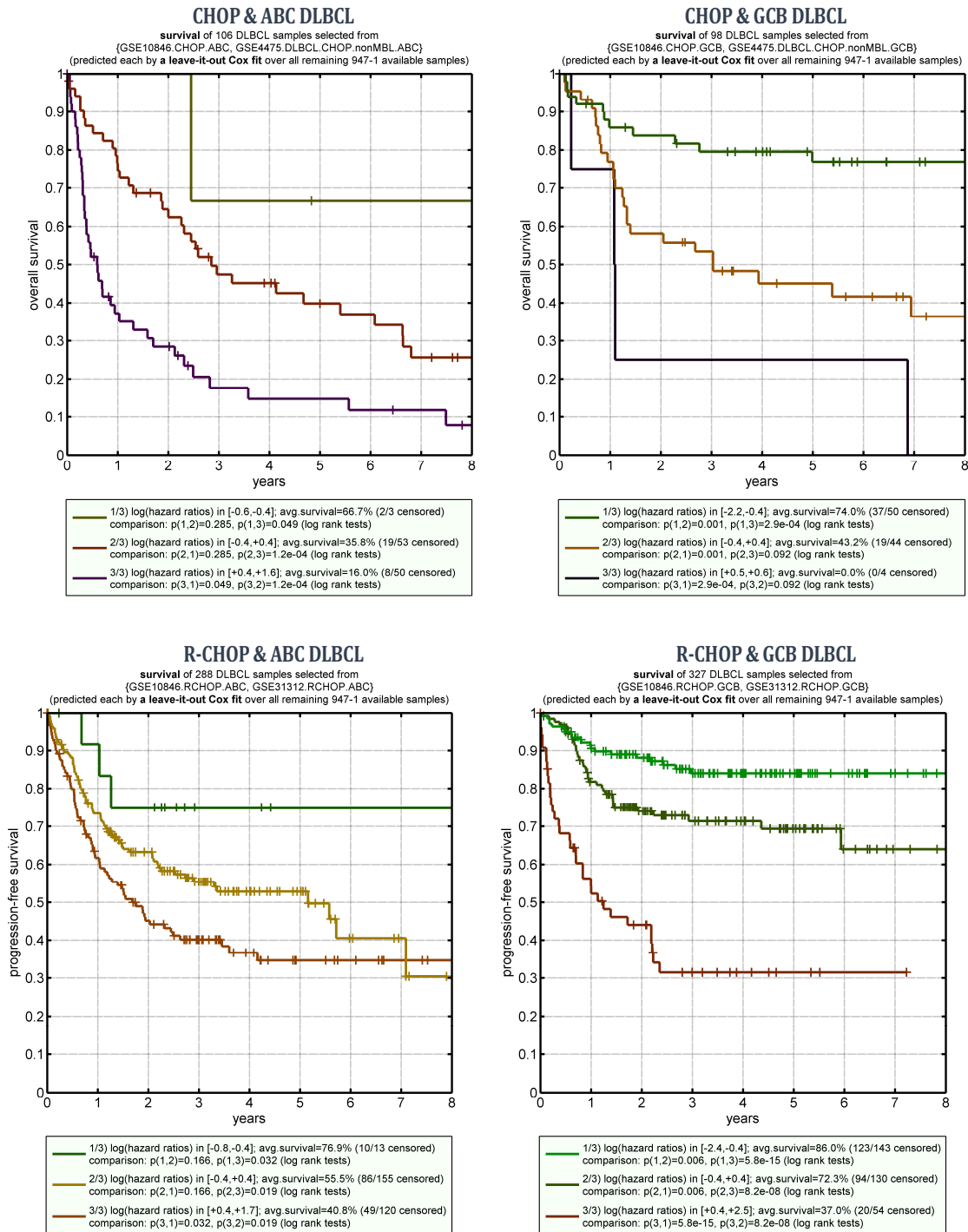


Figure III.2.5.6) Quivariate model, predicted survival dependencies within standard ABC- and GCB-like DLBCL subtypes

Quivariate leave-one-out survival predictions shown separately for standard DLBCL subtypes and for CHOP and R-CHOP therapies. Chosen split points to present the survival spread in three risk intervals equal  $\pm \log(\text{hazard ratios of } 150\%)$ .

Effect  $\nu = 127$  and hence its genetically highly correlated partner effect  $\nu = 131$  influence survival *hierarchically*<sup>(III.2.2.3)</sup>, i.e. it only affect patients in the lower risk partition of the primary effect. For this reason and as before<sup>(cf. III.2.2.4)</sup>, predictions are additionally analyzed for each risk partition of  $\nu = 134$  separately<sup>(Figure III.2.5.7)</sup> rather than for all samples simultaneously.

The lower risk partition of  $\nu = 134$  roughly corresponds to GCB DLBCL and its higher risk partition roughly corresponds to ABC DLBCL, but no samples are excluded as “unclassified” by partitioning at zero effect eigensignal<sup>(cf. III.2.1.3)</sup>. Similar to splitting into standard subtypes<sup>(Figure III.2.5.6)</sup> and as expected by average survival,

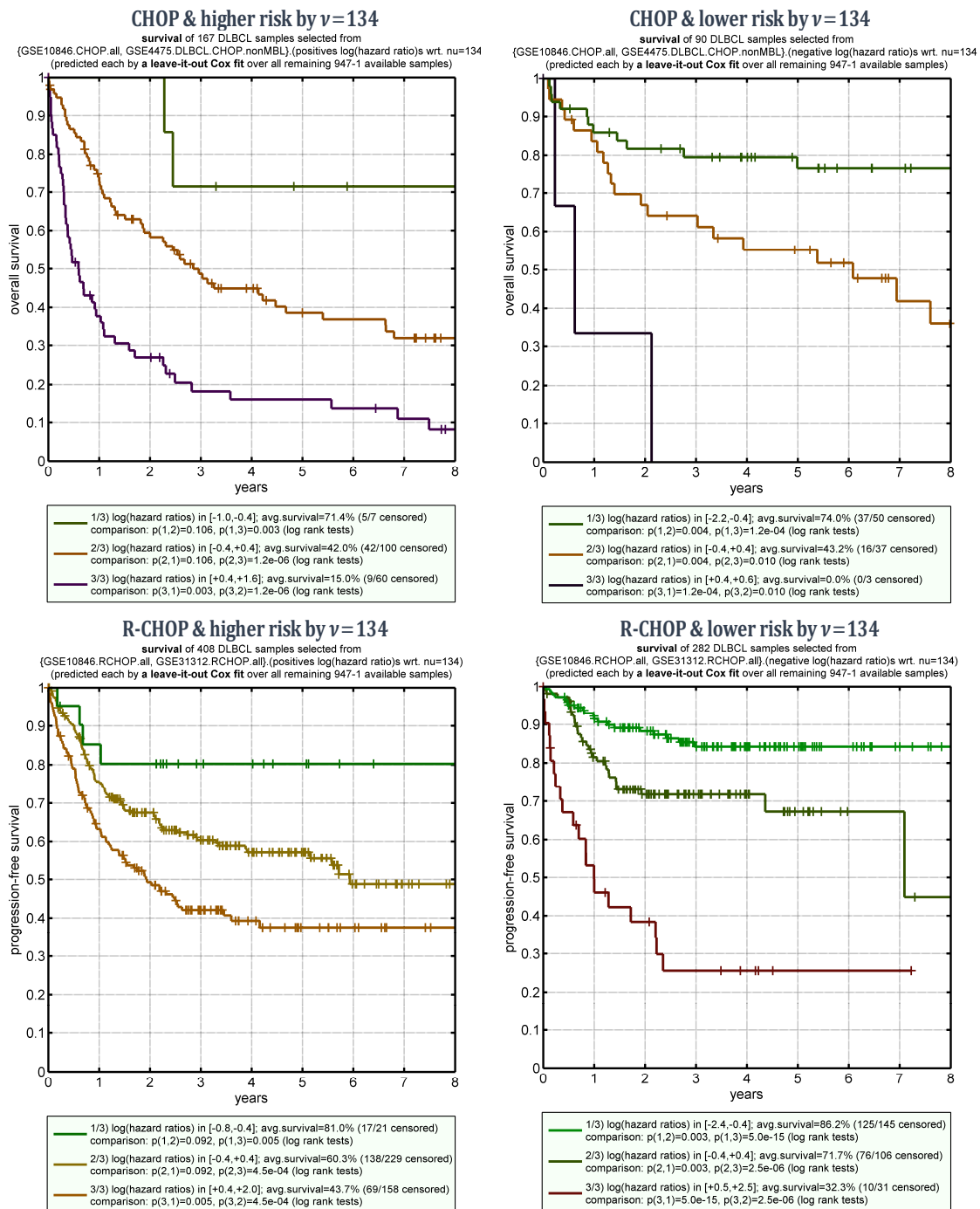


Figure III.2.5.7) Quinvariate model, predicted survival differences within risk partitions of  $\nu = 134$

Quinvariate leave-one-out predictions shown separately for risk partitions of  $\nu = 134$  and for CHOP and R-CHOP therapies. Chosen split points to present the survival spread in three risk intervals equal  $\pm \log(\text{hazard ratios of } 150\%)$ .

samples in lower risk predictor intervals are thinned out in the higher risk partition of  $\nu = 134$  and samples in higher risk predictor intervals are thinned out in its lower risk partition.

As with the bivariate model<sup>(III.2.2.4, III.2.2.5)</sup>, splitting into risk partitions of  $\nu = 134$  allows for a cleaner separation by survival with lower  $p$  values compared to splitting into standard subtypes<sup>(Figure III.2.5.6)</sup>, despite including all unclassified samples that usually have average outcome and wash out predicted survival spreads. For example, the two lower risk intervals for ABC DLBCL in R-CHOP are separated only with  $p = 0.019$  only, but corresponding survival curves in the higher risk partition of  $\nu = 134$  are already separated by  $p = 4.5 \cdot 10^{-4}$ . This is consistent with other indications that partitioning by  $\nu = 134$  seems to be the more natural choice for identification of DLBCL subtypes. This is especially interesting, because the GEP effect  $\nu = 134$  is genetically rather distinct from the rediscovered COO induced effect  $\nu = 129$  and it may point to genes that are potentially causal to the disease, but so far biologically unexplored in DLBCL context<sup>(cf. III.3.3.1)</sup>.

### III.2.5.8 Predictions within risk classes by International Prognostic Index

---

The international prognostic index<sup>[100]</sup> (IPI) summarizes macroscopic clinical data for survival prediction. In brief, the IPI score for a patient is the sum of following general conditions: IPI = (age  $\geq 60$ ) + (Ann Arbor stage  $\geq 3$ ) + (LDH ratio  $\geq 1$ ) + (# extranodal sites  $\geq 2$ ) + (ECOG performance score  $\geq 2$ ). It can also predict a strong survival spread, but it cannot give any hints to molecular causes of the disease, nor could it robustly recommend therapies, once therapies that are specific for molecular subtypes exist. Still, for clinical relevance of predictors it is important to know, whether molecular effects can predict significant survival differences *within* IPI risk classes. Clinical data for IPI scores were available for the majority of patients in both R-CHOP cohorts and I applied the quivariate predictor separately to risk classes determined by IPI scores<sup>(Figure III.2.5.8)</sup>.

As expected, a general trend can be observed that higher IPI risk classes show lower average survival and hence have more patients in high risk intervals of the molecular predictor. As not all patients have IPI annotations and due to splitting into four risk classes, remaining numbers are not very high. This is probably the major reason, why some neighboring curves do not show significant survival differences. However, survival trends are consistent with predicted molecular risks within all IPI risk classes and thus it may be anticipated that even neighboring curves will gain significance for future larger sample counts.

Already now, significant differences in patient outcome between top and bottom risk intervals can be molecularly predicted within each IPI risk class. This demonstrates that macroscopic clinical observables underlying the IPI score can no longer serve as surrogates for molecular prediction, as soon as different therapies for different molecular subtypes are clinically available.

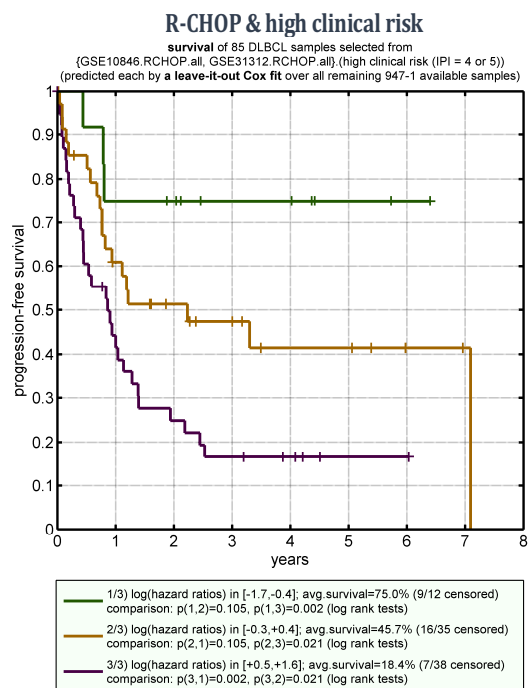
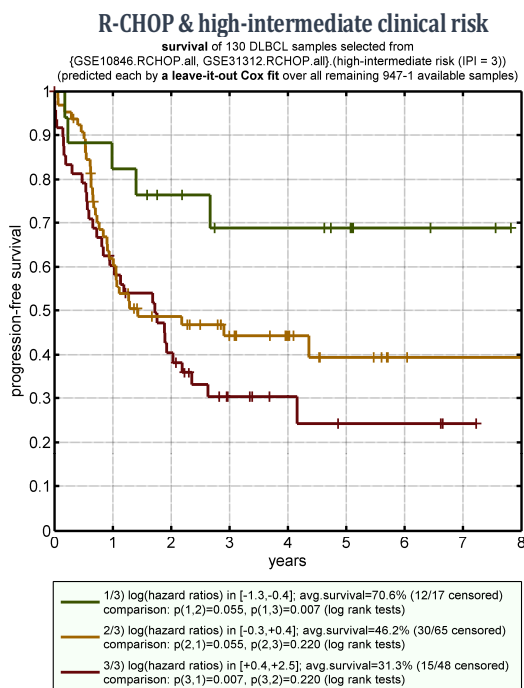
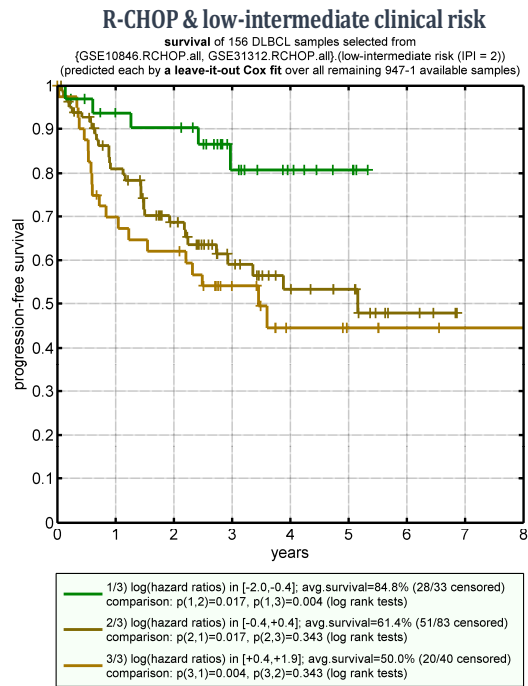
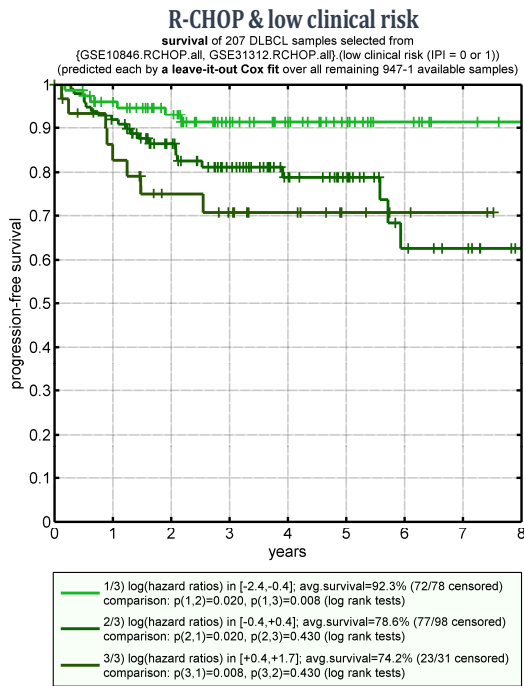


Figure III.2.5.8) Quinvariate model, predicted survival differences within IPI risk classes

163 samples from GSE10846.R-CHOP and 415 samples from GSE31312.R-CHOP having all clinical data for IPI score computation were partitioned into: a) low clinical risk (IPI=0 or 1), low-intermediate clinical risk (IPI=2), high-intermediate clinical risk (IPI=3) and a high clinical risk (IPI=4 or 5). The quinvariate predictor was applied to each sample subset. Chosen split points to present survival spreads in three risk intervals equal  $\pm \log(\text{hazard ratios of } 150\%)$ .

## III.3 Biostatistical Evaluation

---

*To decode and to model cellular pathways that might explain why top genes of discovered effects are highly correlated to each other, a first step is to biologically examine all validated GEP effects that significantly differentiate patients with Diffuse Large B-Cell Lymphoma.*

*This evaluation usually consists of two collaborative parts that may be iterated: genetic interpretation and experimental validation by experts in molecular biology as well as biostatistical evaluation of detected effects and of subsequent experiments based on available computable genomic and clinical knowledge.*

*This section presents the initial iteration of the latter. These analyses might help to identify disease-specific effects that are promising targets for further experimental validation in molecular biology.*

*Selected effects are evaluated in detail based on several clinical and genomic analyses; similar analyses are available at [D=Interpretation](#) for all 135 detected effects (15304 Excel® tables and 36588 editable EPS plots). Additionally, statistical overview tables are provided that describe and link most plots and sub tables to quickly find files of biological interest.*

### III.3.1 Analyses and Statistical Tests

---

Biostatistical analysis methods have already been introduced<sup>(cf. 1.3)</sup>. This section briefly describes their application to discovered and validated GEP effects.

#### III.3.1.1 Association with clinical knowledge

---

Basic and nominal clinical annotations like patient gender can be associated with effects via usual contingency tables: Samples are sorted based on their effect eigensignal strengths  $|u_{v,c}^s|$  (Table III.1.5) and then partitioned by cutting at zero. The association between resulting binary partitions and categories of nominal clinical covariates is quantified by  $\chi^2$  tests.

Besides this binary partitioning, effect eigensignal strengths  $|u_{v,c}^s|$  may alternatively be cut at  $\pm \frac{1}{2}$  standard deviations of all samples  $\langle e_{c,j}^s | u_{v,c}^s \rangle$  in a cohort  $C$ . This cuts samples into three groups for each effect and is useful to reveal significant associations that only exist e.g. for upper effect strengths. For each validated GEP effect  $v$ , associations of all available clinical covariates are computed for both partitioning alternatives.

Besides for nominal covariates, the same sample partitions are used to associate effects with patient outcome. More precisely, Kaplan-Meier survival estimates are computed for each sample partition and log rank tests are used to compare these estimates<sup>(1.3.3)</sup>. These survival analyses are univariate (i.e. effect-centric) and are useful for validation of single GEP effects by independently observed survival. They cannot reveal survival dependencies based on *two or more* effects; to analyze these dependencies, multivariate survival analyses have been performed<sup>(III.2)</sup>.

All associations of individual GEP effects with available clinical knowledge about samples are provided in graphical and tabular form at [D=Interpretation\clinical\v\(effect index\)\\(clinical cohort name\)](#).

To interpret effects biologically and their DLBCL-specificity in particular, existing knowledge about their top genes and their function should to be systematically collected.


In the easiest case, the same set of genes has already been discovered in other fields of biology or medicine and is clearly DLBCL-unspecific, for example differential expressions between different blood cells.

Usually, effects are only roughly associated with existing genomic knowledge and it is not trivial to see a biological connection or an underlying pathway that links all significant associations. Hence, often further biological experiments asking specific questions about individual genes in model cell lines are required to reverse-engineer underlying pathways step by step.

Notably, even for effects that do not have any significant association with survival, these genomic analyses might still help to identify parts of the molecular pathogenesis of DLBCL, as observed patient outcome is based on either the CHOP or R-CHOP therapy. For example, effects might represent potential attack vectors that are not utilized by current therapies. On the other hand, they could also be DLBCL-unspecific effects like the gender induced effect.

Gene set enrichment analyses and other basic signature statistics are utilized to reveal related sets of genes from a combined database comprised of 13584 known gene signatures that have been discovered and studied in a wide range of fields in human biology and were imported from various sources<sup>(cf.1.3.1)</sup>. Genes are ranked by the same maximally biologically informative gene scores that were selected for cross-cohort validation<sup>(III.1.3.2)</sup>. Hence, top genes in this gene order show both a strong differential signal between patients and are highly correlated to other top genes of the same effect.

Gene ontology overrepresentation analyses<sup>(I.3.2)</sup> are utilized to reveal terms in cellular components, molecular functions and biological processes that are significantly related to top genes of respective effects. For each effect, overrepresentation is tested for several gene signatures of top-correlated or of top-anti-correlated effect genes (genes are ranked as above and then cut by their relative correlation in 10% steps).

These genomic analyses have also been computed for all 135 validated GEP effects and are available in effect subfolders at  `D=Interpretation\genomic`.

An example of another complementary genomic association analysis is the correlation of top effect genes to copy number differences measured by array comparative genomic hybridization (aCGH). This analysis is not part of this work, but might reveal amplifications or losses *on DNA level* at or around genomic loci of effect top genes. Ideally, this could identify oncogenes respectively tumor suppressor genes that are the key regulators of validated GEP effects.

### III.3.2 Effects Identified by Sample Annotations

First, the quad-discovered gender effect is presented that also serves as independent control of the whole analysis pipeline.

Next, effects are interrogated for their association with standard DLBCL subtypes. One particular unsupervisedly discovered effect is identified as the rediscovered cell of origin induced effect<sup>[4]</sup>, because its sample axes show 95% agreement with public classifications into ABC DLBCL and GCB DLBCL for all samples. Several other genetically distinct effects are also significantly related to DLBCL subtypes and have markedly different predictive capabilities, demonstrating that a binary classification can only provide a very rough summary of survival trends over gene expressions.

#### III.3.2.1 $\nu = 2$ : Gender effect and annotation mistakes

##### Effect overview

The strongest effect in terms of differential expression between patients was observed for genes that significantly differentiate males and females<sup>(Table III.3.2.1)</sup>. Naturally, this effect exists in all four detection cohorts. It is not DLBCL-specific. Detection ranks were #3<sup>(GSE10846.CHOP)</sup>, #4<sup>(GSE10846.R-CHOP)</sup>, #6<sup>(GSE4475.nonMBL)</sup> and #6<sup>(GSE31312.R-CHOP)</sup>. The average correlation of these detected effect across cohorts is the second-highest with  $\mathcal{R}(3,4,6,6) = 0.88$ <sup>(cf. III.1.2.2)</sup>. This effect and its validation is also an *independent control* of detection, validation, annotation and interpretation pipelines for real world data.

As expected, exemplary detections from two cohorts show<sup>(Figure III.3.2.1)</sup> that the effect signal has a *binary nature*. Clearly, patients are correctly predicted and sorted by gender<sup>(pink/blue)</sup>. Naturally, the effect is carried by genes from  $X$  and  $Y$  chromosomes, for example  $XIST$  and  $EIF1AY$ <sup>(cf. Table III.3.2.1.a)</sup>. Only men can have an expression for  $Y$  chromosome genes and consequently women have negative  $\log_2(\text{ratio})$ s for these genes<sup>(blue)</sup>. (All expression ratios are computed relative to cohort-average gene expressions.)

This is a prime example for a reflection of well-known *chromosomal features* on gene expression level. Due to very strong foldings relative to the noise level, correlations of top genes with this effect are very high. Typical gene expression effects that are based on regulation networks rather than on chromosomal features typically cannot reach correlations  $|r| > 0.9$  due to noise. Typical correlations in these cases are only  $|r| \approx 0.5$ , but associated  $p$  values are still approximately  $10^{-20}$  or less.

Despite being directly related to the chromosomal level, it is interesting that only few of in total 1822 measured probesets for the two gender chromosomes are correlated to the gender effect. One conceivable explanation could be that other measured probesets are simply not expressed in measured samples, but this is not the case here. (Probesets with  $|r_2^g| \geq 0.4$  have an

| Top genes in $\nu=2$ |             |            |             |             |             |  |
|----------------------|-------------|------------|-------------|-------------|-------------|--|
| Probeset             | Gene Symbol | HG19 locus | $a_2^{g,c}$ | $r_2^{g,c}$ | $p_2^{g,c}$ | $\left  \frac{a_2^{g,c}}{r_2^{g,c}} \right $ |
| 204409_s_at          | EIF1AY      | Yq11.223   | 2.19        | 0.94        | 1.2E-116    | 2.05   |
| 205000_at            | DDX3Y       | Yq11.21    | 2.11        | 0.96        | 3.2E-138    | 2.02   |
| 201909_at            | RPS4Y1      | Yp11.31    | 2.03        | 0.94        | 3.6E-121    | 1.91   |
| 206700_s_at          | KDM5D       | Yq11.222   | 1.62        | 0.93        | 5.9E-116    | 1.51   |
| 232618_at            | TXLNG2P     | Yq11.222   | 1.43        | 0.87        | 4.1E-90     | 1.24   |
| 236694_at            | TXLNG2P     | Yq11.222   | 1.39        | 0.87        | 3.7E-93     | 1.21   |
| 228492_at            | USP9Y       | Yq11.21    | 1.38        | 0.87        | 6.3E-93     | 1.20   |
| 204410_at            | EIF1AY      | Yq11.223   | 1.31        | 0.86        | 1.4E-76     | 1.13   |
| 214131_at            | TXLNG2P     | Yq11.222   | 1.20        | 0.83        | 1.7E-65     | 0.99   |
| 223646_s_at          | TXLNG2P     | Yq11.222   | 1.14        | 0.84        | 5.0E-79     | 0.96   |
| 223645_s_at          | TXLNG2P     | Yq11.222   | 1.03        | 0.84        | 3.0E-79     | 0.86   |
| 205001_s_at          | DDX3Y       | Yq11.21    | 0.93        | 0.88        | 1.9E-83     | 0.82   |
| 230760_at            | ZFY         | Yp11.31    | 0.99        | 0.80        | 1.7E-65     | 0.79   |
| 206624_at            | USP9Y       | Yq11.21    | 0.80        | 0.76        | 1.1E-49     | 0.61   |
| 211149_at            | UTY         | Yq11.221   | 0.68        | 0.82        | 3.1E-62     | 0.55   |
| 214983_at            | TTY15       | Yq11.21    | 0.65        | 0.79        | 5.7E-57     | 0.51   |
| 244482_at            |             | Yq11.223   | 0.61        | 0.69        | 2.6E-42     | 0.42   |
| 243712_at            | XIST        | Xq13.2     | -0.94       | -0.69       | 3.3E-42     | -0.65  |
| 231592_at            | TSIX        | Xq13.2     | -1.24       | -0.86       | 4.9E-88     | -1.07  |
| 235446_at            |             | Xq13.2     | -1.36       | -0.81       | 1.1E-69     | -1.11  |
| 224589_at            | XIST        | Xq13.2     | -2.17       | -0.95       | 9.7E-145    | -2.05  |
| 227671_at            | XIST        | Xq13.2     | -2.34       | -0.95       | 9.5E-145    | -2.21  |
| 221728_x_at          | XIST        | Xq13.2     | -2.51       | -0.92       | 6.7E-104    | -2.31  |
| 214218_s_at          | XIST        | Xq13.2     | -2.58       | -0.92       | 7.5E-105    | -2.37  |
| 224590_at            | XIST        | Xq13.2     | -2.70       | -0.97       | 1.0E-175    | -2.61  |
| 224588_at            | XIST        | Xq13.2     | -3.10       | -0.98       | 1.4E-198    | -3.03  |

Table III.3.2.1.a) Top genes in validated effect  $\nu=2$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx<sup>TM</sup>[97] v33)  
 $a_2^{g,c}$  Components of the consensus gene axis of effect  $\nu=2$  (cf. Table III.1.5); filtered  $|a_2^{g,c}| \geq 0.4$ .  
 $r_2^{g,c}$  Consensus gene correlations; filtered  $|r_2^{g,c}| \geq 0.4$ .  
 $p_2^{g,c}$   $p$  values for the correlations (cf. II.5.2.1)

average  $\log_2(\text{intensity})$  over all samples of 3.89 and show a clearly differential signal (Figure III.3.2.1). For probesets that are correlated less, 916/1796 show a higher average  $\log_2(\text{intensity})$ ; 99 of them even have average  $\log_2(\text{intensities})$  of  $>7$ . This indicates that they are expressed, but not in a gender-modulated way.)

### ■ Potential gender annotation mistakes

Some patients show significant (as per  $p$  value for correlations to the effect's consensus gene axis) and clearly defined gender on gene expression level that is *different* from their annotated gender (Table III.3.2.1.b and pink/blue classification in Figure III.3.2.1). Annotations for these samples should be reexamined. (Patient IDs together with their correlations to the effect's gene axis and associated  $p$  values are available in definition tables for consensus effect  $\nu = 2$  in cohort subfolders of [C=Consensus Effects](#).)

### ■ Survival and clinical associations

Besides gender, no other clinical covariate was consistently and significantly associated with this effect. (All clinical correlations and contingency tables are available in *c) clinical correlations\** tables in subfolders of [D=Interpretation\clinical\002](#).)

Interestingly, females were borderline significantly associated with favorable outcome in one cohort ( $p_{\text{GSE10846.R-CHOP}} = 0.049$ , log rank test for patients partitioned at zero effect eigensignal). A study<sup>[101]</sup> showed a similar survival bias. However, this trend could not be validated in any of the three other cohorts ( $p_{\text{GSE10846.CHOP}} = 0.68$ ,  $p_{\text{GSE4475.DLBCL}} = 0.82$ ,  $p_{\text{GSE31312}} = 0.32$ ). All relevant and validated DLBCL survival factors based on gene expressions have been systematically analyzed<sup>(III.2)</sup> and the gender effect was not significant. Possibly, this weak association is a shadow of age related biases in this cohort, potentially caused by general lower life expectancy for males.

### ■ Genomic associations

Surprisingly, in the combined signatures database comprising 13586 published gene signatures from various sources<sup>(I.3.1)</sup>, only two signatures are related to this gender effect and only remotely so. The first is the positional gene set [chryq11](#) (cf. [online interpretation card](#)) with 204 defined and 29 measured genes, an average  $\log_2(\text{ratio})$  of  $-0.88$  and a rather low enrichment score of 0.5 ( $p_{\text{GSEA}} < 0.002$  based on 514 permutations,  $FDR = 0.2\%$ ). The second signature on the other regulation side is [Disteche, escaped from X inactivation](#) (cf. [online interpretation card](#)) with 13 defined and measured genes, an average  $\log_2(\text{ratio})$  of  $+0.28$  and an enrichment score of  $-0.85$  ( $p_{\text{GSEA}} < 0.002$  based on 489 permutations,  $FDR = 0.2\%$ ). For comparison, the average  $\log_2(\text{ratio})$  of top genes of  $\nu = 2$  is  $-3.89$  and  $+4.34$  respectively, i.e. much stronger. (The complete signature analyses for the gender effect with all statistics is available at [D=Interpretation\genomic\002\SA](#).)

### ■ Inference

Given its perfect suitability as control effect, I expected more gender-specific gene expression signatures in public signature databases. It might be a useful addition, especially as only a specific subset of expressed probesets from  $X$  and  $Y$  chromosomes are actually strongly correlated to this gender effect.

Besides its use in method validation contexts, it might be utilized for quality control of gender annotations when publishing large studies with gene expression measurements.

| GEP effect strength                          | # annotated as males                     | # annotated as females |
|--|--|------------------------|
| GSE10846.CHOP (163/181 patients annotated)   |  |                        |
| < 0  | 2  | 72                     |
| ≥ 0  | 88                                       | 1                      |
| <i>p</i> value                               | 9.8 · 10 <sup>-35</sup> ( $\chi^2$ test) |                        |
| GSE10846.R-CHOP (220/233 patients annotated) |  |                        |
| < 0  | 2  | 91                     |
| ≥ 0  | 126                                      | 1                      |
| <i>p</i> value                               | 4.0 · 10 <sup>-47</sup> ( $\chi^2$ test) |                        |
| GSE4475 (DLBCL only, 76/166 patients)        |  |                        |
| < 0  | 2  | 35                     |
| ≥ 0  | 39                                       | 0                      |
| <i>p</i> value                               | 1.3 · 10 <sup>-16</sup> ( $\chi^2$ test) |                        |
| GSE31312 (470/498 patients annotated)        |  |                        |
| < 0  | 9  | 185                    |
| ≥ 0  | 262                                      | 14                     |
| <i>p</i> value                               | 1.0 · 10 <sup>-84</sup> ( $\chi^2$ test) |                        |

Table III.3.2.1.b) Gender effect, contingency with clinical annotations

Patients of each cohort were partitioned at zero eigensignal strength (Table III.L.5) for GEP effect  $\nu = 2$ .



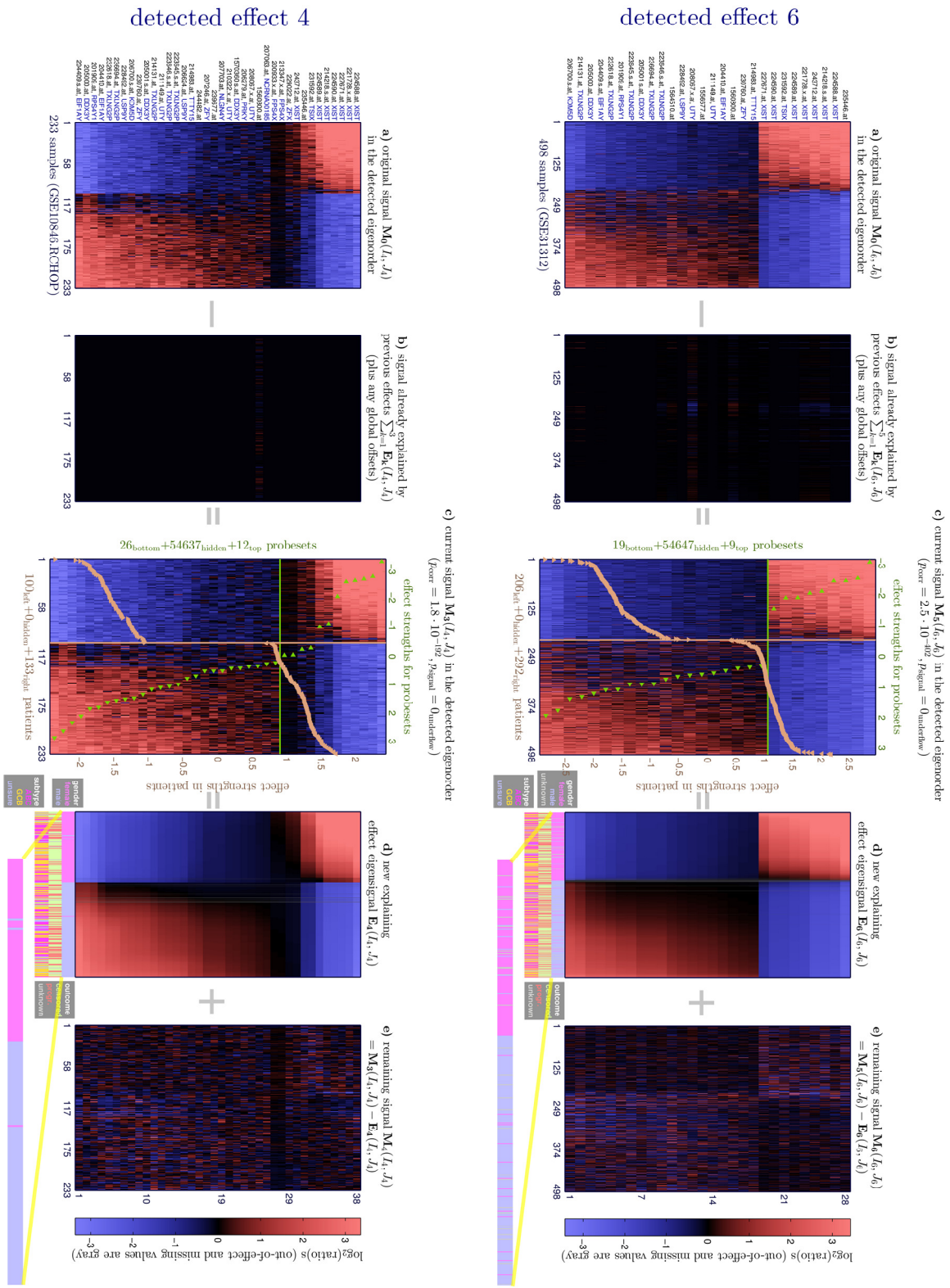


Figure III.3.2.1) Detected gender effects

Gender effects detected in GSE10846.R-CHOP (233 patients) and GSE31312 (498 patients). Both are part of the quad-discovered consensus gender effect. Sample effect strengths (orange curves in center panels) are clearly either positive or negative with a jump in-between, as can be expected for a naturally binary effect. Enlarged gender classifications show some mismatches between GEP based genders and annotations. Due to the obvious signal of the gender effect on GEP level, these are presumably annotation mistakes.

### III.3.2.2 $\nu = 129$ : Cell of origin induced effect (standard DLBCL subtypes)

#### Identification of the cell of origin induced effect

The COO induced effect distinguishes ABC DLBCL from GCB DLBCL (cf. [4] and Figure 1.1.2.1). To test whether it has been unsupervisedly rediscovered, contingency tables have been computed for previously identified patient subtypes and for each of the 135 validated GEP effects.

Many validated GEP effects can arrange patients with a significant bias to subtype. All effects are depicted (Figure III.3.2.2.a) that agree with  $\geq 70\%$  to previous subtype classifications when partitioning samples at zero effect eigensignal (III.3.1.1). In principle, all these GEP effects might help to comprehend DLBCL subtype biology. However, the smaller the agreement, the more likely it is that the effect represents a pathway that is not restricted to only one subtype.

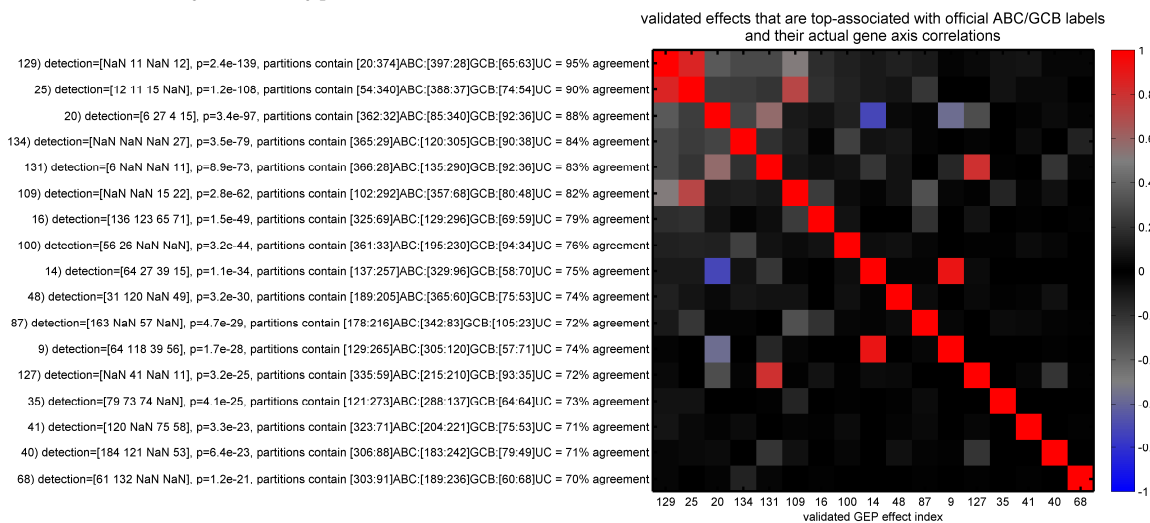


Figure III.3.2.2.a Validated GEP effects that are top-associated with public sample classifications as ABC DLBCL or GCB DLBCL

Validated GEP effects with  $\geq 70\%$  agreement to public classifications as ABC DLBCL or GCB DLBCL from all four cohorts when partitioning samples by cutting at zero eigensignal of the respective consensus effect. Contingency tables for both resulting effect partitions and for public ABC/UC/GCB annotations are provided in text form; their  $p$  values are based on  $\chi^2$  tests. The matrix shows whether effects are genetically associated with each other (weighted correlations of their consensus gene axes; cf. III.1.2.1).

With 95% agreement over all four cohorts ( $p = 2 \cdot 10^{-139}$ ), GEP effect  $\nu = 129$  is clearly at rank #1, thereby identifying it as *the rediscovered cell of origin induced effect*. It is followed by  $\nu = 25$  (90% agreement,  $p = 10^{-108}$ ) that has a highly correlated consensus gene scores to the first effect with  $r_{(129;25)}^c = 0.85$  (cf. Eqn. III.1.3.2.b) and thus can be considered biologically similar (for one cohort, both consensus effects are even based on the identical discovered effect).

At rank #3, a quad-discovered effect follows ( $\nu = 20$ , 88% agreement,  $p = 3 \cdot 10^{-97}$ ). This effect is interesting, because it is based on rather different genes ( $r_{(129;20)}^c = -0.36$ ) and can only predict relatively weak survival differences (cf. III.3.4.1). This demonstrates that a binary classification into subtypes can only provide a relatively rough summary of genetic survival trends compared to gradual effect strengths, as no patient ordering *within* classes can be prescribed in this way.

Effect  $\nu = 134$  at rank #4 can still arrange patients into ABC DLBCL and GCB DLBCL with 84% agreement, but like  $\nu = 20$  it is based on rather different genes ( $r_{(129;134)}^c = -0.29$ ). Because of its excellent survival prediction capability,  $\nu = 134$  is evaluated in detail (cf. III.3.3.1).

Effect  $\nu = 131$  follows at rank #5. It is strongly correlated to effect  $\nu = 127$  ( $r_{(131;127)}^c = 0.80$ ). Both are not genetically correlated to  $\nu = 129$  ( $r_{(129;127)}^c = -0.04$  and  $r_{(129;131)}^c = -0.29$ ), but these effect play an important role in multivariate survival analysis (III.2.2.3) and are, therefore, also presented in detail (in III.3.3.2).

## Effect overview

Effect  $v = 129$  has been unsupervisedly detected in two cohorts (at rank #11 in GSE10846.R-CHOP and at rank #12 in GSE31312.R-CHOP). Application of this effect sorts patients from GCB DLBCL to ABC DLBCL, as is depicted exemplary for the largest cohort (Figure III.3.2.2.b, page 179). Similar plots and definition tables are available for all four cohorts in  $\mathcal{C}=\text{Consensus Effects}\backslash(\text{cohort subfolder})$ ; they supervisedly validate the effect's existence in each cohort.

Other than for the gender effect (III.3.2.1), *effect strengths for patients are gradual* here. Patients in the center do not have sufficient correlation to the effect's gene axis and hence their eigensignal is gray (II.4.2.1). Patients with effect strengths near zero therefore should be considered neither ABC DLBCL nor GCB DLBCL. (Previous subtype classifications implicitly assumed that every patient is either ABC DLBCL or GCB DLBCL by using a two-class Bayes predictor model<sup>[83]</sup> and hence could underestimate the ratio of "unclassified" patients with respect to the COO induced effect; cf. III.2.1.9.)

The effect is two-sided, i.e. it is comprised both of correlated "ABC>GCB" and anti-correlated "GCB>ABC" genes. All correlated genes with strong folding between patients as per gene scores (III.1.3.2)  $|a_{129}^g| \cdot r_{129}^g$  are listed (Table III.3.2.2.a) for cutoffs  $|a_{129}^g| \geq 0.4$  and  $|r_{129}^g| \geq 0.4$ ; the complete list is available in  $\mathcal{DLBCL Master Table 2015, gene orders.xlsx}$ .

| Correlated genes in $v=129$ (ABC>GCB DLBCL) |             |            |                 |                 |                 |                                       | Anti-correlated genes in $v=129$ (GCB>ABC DLBCL) |              |            |                 |                 |                 |                                       |
|---|-------------|------------|-----------------|-----------------|-----------------|---------------------------------------|--|--------------|------------|-----------------|-----------------|-----------------|---------------------------------------|
| Probeset                                    | Gene Symbol | HG19 locus | $a_{129}^{g,c}$ | $r_{129}^{g,c}$ | $p_{129}^{g,c}$ | $ a_{129}^{g,c}  \cdot r_{129}^{g,c}$ | Probeset   | Gene Symbol  | HG19 locus | $a_{129}^{g,c}$ | $r_{129}^{g,c}$ | $p_{129}^{g,c}$ | $ a_{129}^{g,c}  \cdot r_{129}^{g,c}$ |
| 44790_s_at                                  | KIAA0226L   | 13q14.13   | 1.39            | 0.63            | 3.1E-27         | 0.88                                  | 1554413_s_at                                     | SNX29P2      | 16p11.2    | -1.06           | -0.53           | 3.1E-18         | -0.56                                 |
| 219471_at                                   | KIAA0226L   | 13q14.13   | 1.15            | 0.61            | 3.7E-25         | 0.70                                  | 219753_at  | STAG3        | 7q22.1     | -1.02           | -0.52           | 3.9E-17         | -0.52                                 |
| 224838_at                                   | FOXP1       | 3p13       | 0.92            | 0.69            | 9.4E-34         | 0.63                                  | 204249_s_at                                      | LMO2         | 11p13      | -0.96           | -0.54           | 7.8E-19         | -0.52                                 |
| 224837_at                                   | FOXP1       | 3p13       | 0.79            | 0.66            | 3.7E-30         | 0.52                                  | 236981_at  | C17orf99     | 17q25.3    | -1.12           | -0.45           | 3.6E-13         | -0.51                                 |
| 223287_s_at                                 | FOXP1       | 3p13       | 0.79            | 0.61            | 1.1E-24         | 0.48                                  | 206310_at  | SPINK2       | 4q12       | -1.05           | -0.45           | 3.6E-13         | -0.48                                 |
| 235444_at                                   | FOXP1       | 3p13       | 0.74            | 0.59            | 2.0E-23         | 0.44                                  | 242334_at  | NLRP4        | 19q13.43   | -1.14           | -0.41           | 6.5E-11         | -0.47                                 |
| 212827_at                                   | IGHM        | 14q32.33   | 1.05            | 0.40            | 1.8E-10         | 0.42                                  | 226281_at  | DNER         | 2q36.3     | -1.13           | -0.41           | 9.1E-11         | -0.46                                 |
| 244413_at                                   | CLECL1      | 12p13.31   | 0.80            | 0.47            | 5.6E-14         | 0.37                                  | 211597_s_at                                      | HOPX         | 4q12       | -1.00           | -0.45           | 7.6E-13         | -0.45                                 |
| 229844_at                                   | FOXP1       | 3p13       | 0.62            | 0.60            | 2.0E-24         | 0.37                                  | 223159_s_at                                      | NEK6         | 9q33.3     | -0.74           | -0.59           | 6.4E-23         | -0.43                                 |
| 229114_at                                   | GAB1        | 4q31.21    | 0.80            | 0.46            | 9.7E-14         | 0.37                                  | 207599_at  | MMP20        | 11q22.2    | -0.99           | -0.42           | 1.8E-11         | -0.42                                 |
| 1558996_at                                  | FOXP1       | 3p13       | 0.68            | 0.54            | 4.8E-19         | 0.37                                  | 223158_s_at                                      | NEK6         | 9q33.3     | -0.68           | -0.61           | 4.0E-25         | -0.41                                 |
| 227198_at                                   | AFF3        | 2q11.2     | 0.82            | 0.45            | 6.3E-13         | 0.37                                  | 227703_s_at                                      | SYTL4        | Xq22.1     | -0.88           | -0.46           | 9.5E-14         | -0.41                                 |
| 203753_at                                   | TCF4        | 18q21.2    | 0.66            | 0.55            | 1.7E-19         | 0.36                                  | 202119_s_at                                      | CPNE3        | 8q21.3     | -0.74           | -0.54           | 7.2E-19         | -0.40                                 |
| 213891_s_at                                 | TCF4        | 18q21.2    | 0.66            | 0.53            | 3.3E-18         | 0.35                                  | 206181_at  | SLAMF1       | 1q23.3     | -0.75           | -0.53           | 1.7E-18         | -0.40                                 |
| 212386_at                                   | TCF4        | 18q21.2    | 0.68            | 0.50            | 2.6E-16         | 0.34                                  | 231049_at  | LMO2         | 11p13      | -0.77           | -0.51           | 5.1E-17         | -0.40                                 |
| 222762_x_at                                 | LIMD1       | 3p21.31    | 0.61            | 0.56            | 1.7E-20         | 0.34                                  | 213906_at  | MYBL1        | 8q13.1     | -0.94           | -0.42           | 2.7E-11         | -0.39                                 |
| 225331_at                                   | CCDC50      | 3q28       | 0.64            | 0.52            | 4.0E-17         | 0.33                                  | 229041_s_at                                      | LOC100505746 | 21q22.3    | -0.83           | -0.47           | 2.0E-14         | -0.39                                 |
| 220230_s_at                                 | CYB5R2      | 11p15.4    | 0.74            | 0.44            | 3.5E-12         | 0.32                                  | 204604_at  | CDK14        | 7q21.13    | -0.76           | -0.50           | 8.9E-16         | -0.38                                 |
| 212382_at                                   | TCF4        | 18q21.2    | 0.60            | 0.52            | 1.3E-17         | 0.31                                  | 231455_at  | LINC00487    | 2p25.2     | -0.86           | -0.44           | 2.7E-12         | -0.38                                 |
| 1553369_at                                  | FAM129C     | 19p13.11   | 0.66            | 0.47            | 2.1E-14         | 0.31                                  | 244467_at  | SHISA8       | 22q13.2    | -0.91           | -0.41           | 6.0E-11         | -0.37                                 |
| 203313_s_at                                 | TGIF1       | 18p11.31   | 0.61            | 0.49            | 1.8E-15         | 0.30                                  | 242794_at  | MAML3        | 4q31.1     | -0.74           | -0.49           | 1.5E-15         | -0.37                                 |
| 204562_at                                   | IRF4        | 6p25.3     | 0.66            | 0.45            | 5.3E-13         | 0.30                                  | 231455_s_at                                      | CPNE3        | 8q21.3     | -0.68           | -0.51           | 9.0E-17         | -0.35                                 |
| 212387_at                                   | TCF4        | 18q21.2    | 0.55            | 0.53            | 2.1E-18         | 0.29                                  | 239697_x_at                                      | C3orf67      | 3p14.2     | -0.78           | -0.44           | 3.7E-12         | -0.34                                 |
| 222146_s_at                                 | TCF4        | 18q21.2    | 0.58            | 0.51            | 1.7E-16         | 0.29                                  | 244367_at  |              | 11p13      | -0.66           | -0.50           | 4.1E-16         | -0.33                                 |
| 212385_at                                   | TCF4        | 18q21.2    | 0.61            | 0.48            | 6.9E-15         | 0.29                                  | 200644_at  | MARCKSL1     | 1p35.1     | -0.62           | -0.51           | 5.3E-17         | -0.32                                 |
| 232739_at                                   | SPIB        | 19q13.33   | 0.60            | 0.48            | 4.5E-15         | 0.29                                  | 218640_s_at                                      | PLEKHF2      | 8q22.1     | -0.60           | -0.53           | 4.8E-18         | -0.32                                 |
| 1557049_at                                  | BTBD19      | 1p34.1     | 0.65            | 0.44            | 1.5E-12         | 0.29                                  | 219874_at  | SLC12A8      | 3q21.2     | -0.68           | -0.46           | 2.8E-13         | -0.31                                 |
| 228837_at                                   | TCF4        | 18q21.2    | 0.55            | 0.50            | 9.3E-16         | 0.27                                  | 234284_at  | GNG8         | 19q13.32   | -0.75           | -0.41           | 1.0E-10         | -0.31                                 |
| 230983_at                                   | FAM129C     | 19p13.11   | 0.65            | 0.42            | 2.7E-11         | 0.27                                  | 223624_at  | ZFAND4       | 10q11.22   | -0.62           | -0.50           | 6.6E-16         | -0.31                                 |
| 1565034_s_at                                | AFF3        | 11q23.3    | 0.65            | 0.41            | 1.4E-10         | 0.26                                  | 230509_at  | SNX22        | 15q22.31   | -0.61           | -0.50           | 3.9E-16         | -0.31                                 |
| 204269_at                                   | PIM2        | Xp11.23    | 0.52            | 0.48            | 7.6E-15         | 0.25                                  | 218862_at  | ASB13        | 10p15.1    | -0.63           | -0.49           | 3.2E-15         | -0.30                                 |
| 226304_at                                   | HSPB6       | 19q13.12   | 0.59            | 0.42            | 1.7E-11         | 0.25                                  | 243185_at  |              | 10p14      | -0.66           | -0.45           | 5.3E-13         | -0.30                                 |
| 235056_at                                   | ETV6        | 12p13.2    | 0.54            | 0.46            | 3.0E-13         | 0.24                                  | 212314_at  | SEL1L3       | 4p15.2     | -0.57           | -0.52           | 1.0E-17         | -0.30                                 |
| 205222_at                                   | EHHADH      | 3q27.2     | 0.59            | 0.41            | 8.5E-11         | 0.24                                  | 205570_at  | PIP4K2A      | 10p12.2    | -0.59           | -0.50           | 2.5E-16         | -0.30                                 |
| 212345_s_at                                 | CREB3L2     | 7q33       | 0.50            | 0.47            | 1.9E-14         | 0.24                                  | 1553499_s_at                                     | SERPINA9     | 14q32.13   | -0.66           | -0.44           | 1.3E-12         | -0.29                                 |
| 239973_at                                   |             | 7p15.3     | 0.52            | 0.45            | 6.7E-13         | 0.23                                  | 211502_s_at                                      | CDK14        | 7q21.13    | -0.62           | -0.45           | 4.0E-13         | -0.28                                 |
| 208690_s_at                                 | PDLIM1      | 10q23.33   | 0.48            | 0.49            | 2.6E-15         | 0.23                                  | 235213_at  | ITPKB        | 1q42.12    | -0.57           | -0.48           | 7.7E-15         | -0.28                                 |
| 204083_s_at                                 | TPM2        | 9p13.3     | 0.53            | 0.44            | 3.3E-12         | 0.23                                  | 224102_at  | P2RY12       | 3q25.1     | -0.64           | -0.42           | 1.8E-11         | -0.27                                 |
| 218792_s_at                                 | BSPRY       | 9q32       | 0.55            | 0.42            | 3.8E-11         | 0.23                                  | 235353_at  | SEL1L3       | 4p15.2     | -0.57           | -0.47           | 2.9E-14         | -0.27                                 |
| 233483_at                                   | TBC1D27     | 17p11.2    | 0.51            | 0.45            | 7.5E-13         | 0.23                                  | 200965_s_at                                      | ABLIM1       | 10q25.3    | -0.64           | -0.42           | 3.3E-11         | -0.27                                 |
| 218699_at                                   | RAB7L1      | 1q32.1     | 0.50            | 0.46            | 2.1E-13         | 0.23                                  | 225637_at  | DEFB8        | 16q24.3    | -0.66           | -0.41           | 1.2E-10         | -0.27                                 |
| 235051_at                                   | CCDC50      | 3q28       | 0.49            | 0.46            | 1.4E-13         | 0.23                                  | 212975_at  | DENND3       | 8q24.3     | -0.52           | -0.50           | 2.2E-16         | -0.26                                 |
| 200953_s_at                                 | CCND2       | 12p13.32   | 0.55            | 0.41            | 1.1E-10         | 0.23                                  | 225622_at  | PAG1         | 8q21.13    | -0.56           | -0.47           | 2.7E-14         | -0.26                                 |
| 203988_s_at                                 | FUT8        | 14q23.3    | 0.55            | 0.41            | 5.7E-11         | 0.23                                  | 222699_s_at                                      | PLEKHF2      | 8q22.1     | -0.55           | -0.48           | 8.2E-15         | -0.26                                 |
| 244845_at                                   |             | 3p13       | 0.50            | 0.44            | 3.3E-12         | 0.22                                  | 227684_at  | S1PR2        | 19p13.2    | -0.53           | -0.50           | 8.5E-16         | -0.26                                 |

|             |           |          |      |      |         |      |              |              |          |       |       |         |       |
|-------------|-----------|----------|------|------|---------|------|--------------|--------------|----------|-------|-------|---------|-------|
| 213262_at   | SACS      | 13q12.12 | 0.47 | 0.46 | 9.2E-14 | 0.22 | 225626_at    | PAG1         | 8q21.13  | -0.56 | -0.46 | 8.9E-14 | -0.26 |
| 223218_s_at | NFKB1Z    | 3q12.3   | 0.53 | 0.41 | 5.3E-11 | 0.22 | 212311_at    | SEL1L3       | 4p15.2   | -0.57 | -0.46 | 2.6E-13 | -0.26 |
| 225436_at   | FAM108C1  | 15q25.1  | 0.49 | 0.45 | 7.5E-13 | 0.22 | 1563621_at   |              | 7p14.2   | -0.51 | -0.50 | 4.5E-16 | -0.26 |
| 201160_s_at | CSDA      | 12p13.2  | 0.50 | 0.43 | 6.8E-12 | 0.22 | 221039_s_at  | ASAP1        | 8q24.21  | -0.52 | -0.49 | 3.6E-15 | -0.25 |
| 207641_at   | TNFRSF13B | 17p11.2  | 0.47 | 0.46 | 9.8E-14 | 0.22 | 217991_x_at  | SSBP3        | 1p32.3   | -0.49 | -0.51 | 1.2E-16 | -0.25 |
| 226818_at   | MPEG1     | 11q12.1  | 0.48 | 0.44 | 1.9E-12 | 0.21 | 1554575_a_at | BPNT1        | 1q41     | -0.50 | -0.49 | 1.7E-15 | -0.25 |
| 233955_x_at | CXCC5     | 5q31.2   | 0.49 | 0.43 | 5.1E-12 | 0.21 | 224790_at    | ASAP1        | 8q24.21  | -0.51 | -0.48 | 6.6E-15 | -0.24 |
| 223422_s_at | ARHGAP24  | 4q21.23  | 0.49 | 0.43 | 8.4E-12 | 0.21 | 1568817_at   |              | 21q22.3  | -0.60 | -0.41 | 1.3E-10 | -0.24 |
| 201811_x_at | SH3BP5    | 3p25.1   | 0.44 | 0.47 | 7.6E-14 | 0.21 | 230278_at    |              | 1q42.3   | -0.58 | -0.41 | 6.7E-11 | -0.24 |
| 204642_at   | S1PR1     | 1p21.2   | 0.44 | 0.46 | 9.2E-14 | 0.20 | 204137_at    | GPR137B      | 1q42.3   | -0.57 | -0.42 | 2.8E-11 | -0.24 |
| 228693_at   | CCDC50    | 3q28     | 0.48 | 0.42 | 3.9E-11 | 0.20 | 241942_at    | PXDNL        | 8q11.22  | -0.57 | -0.41 | 6.0E-11 | -0.23 |
| 203761_at   | SLA       | 8q24.22  | 0.44 | 0.45 | 4.4E-13 | 0.20 | 235841_at    |              | 4q31.1   | -0.53 | -0.44 | 3.2E-12 | -0.23 |
| 201810_s_at | SH3BP5    | 3p25.1   | 0.46 | 0.43 | 4.6E-12 | 0.20 | 235632_at    |              | 9q33.3   | -0.46 | -0.50 | 4.4E-16 | -0.23 |
| 212654_at   | TPM2      | 9p13.3   | 0.43 | 0.46 | 2.3E-13 | 0.20 | 206348_s_at  | PKD3         | Xp22.11  | -0.51 | -0.45 | 5.7E-13 | -0.23 |
| 244480_at   |           | 18q21.2  | 0.44 | 0.44 | 1.6E-12 | 0.20 | 224796_at    | ASAP1        | 8q24.21  | -0.48 | -0.47 | 6.2E-14 | -0.22 |
| 205965_at   | BATF      | 14q24.3  | 0.45 | 0.43 | 5.4E-12 | 0.19 | 210829_s_at  | SSBP2        | 5q14.1   | -0.54 | -0.41 | 7.2E-11 | -0.22 |
| 218700_s_at | RAB7L1    | 1q32.1   | 0.46 | 0.42 | 2.3E-11 | 0.19 | 201209_at    | HDAC1        | 1p35.1   | -0.47 | -0.47 | 3.4E-14 | -0.22 |
| 236831_at   | CCDC50    | 3q28     | 0.45 | 0.42 | 4.2E-11 | 0.19 | 204891_s_at  | LCK          | 1p35.1   | -0.50 | -0.44 | 1.9E-12 | -0.22 |
| 239231_at   |           | 19p13.11 | 0.43 | 0.44 | 3.1E-12 | 0.19 | 210461_s_at  | ABLIM1       | 10q25.3  | -0.55 | -0.40 | 1.9E-10 | -0.22 |
| 1561167_at  |           | 12p13.2  | 0.45 | 0.41 | 5.2E-11 | 0.19 | 225214_at    | LOC100129034 | 9q33.3   | -0.43 | -0.51 | 7.3E-17 | -0.22 |
| 207237_at   | KCNA3     | 1p13.3   | 0.45 | 0.41 | 1.0E-10 | 0.18 | 238353_at    | RASL11A      | 13q12.2  | -0.45 | -0.49 | 4.0E-15 | -0.22 |
| 212097_at   | CAV1      | 7q31.2   | 0.41 | 0.44 | 1.6E-12 | 0.18 | 229713_at    | PIP4K2A      | 10p12.2  | -0.47 | -0.46 | 8.6E-14 | -0.22 |
| 203068_at   | KLHL21    | 1p36.31  | 0.41 | 0.43 | 5.3E-12 | 0.18 | 208456_s_at  | RRAS2        | 11p15.2  | -0.51 | -0.42 | 1.6E-11 | -0.22 |
| 203143_s_at | KIAA0040  | 1q25.1   | 0.42 | 0.43 | 7.4E-12 | 0.18 | 229040_at    | LOC100505746 | 21q22.3  | -0.48 | -0.45 | 4.6E-13 | -0.21 |
| 205861_at   | SPIB      | 19q13.33 | 0.41 | 0.44 | 3.3E-12 | 0.18 | 212829_at    | PIP4K2A      | 10p12.2  | -0.46 | -0.47 | 5.7E-14 | -0.21 |
| 207655_s_at | BLNK      | 10q24.1  | 0.42 | 0.41 | 1.1E-10 | 0.17 | 203723_at    | ITPKB        | 1q42.12  | -0.48 | -0.45 | 9.6E-13 | -0.21 |
| 209939_x_at | CFLAR     | 2q33.1   | 0.42 | 0.40 | 2.3E-10 | 0.17 | 1555626_a_at | SLAMF1       | 1q23.3   | -0.43 | -0.49 | 1.4E-15 | -0.21 |
| 243878_at   |           | 3p13     | 0.41 | 0.40 | 1.5E-10 | 0.17 | 203537_at    | PRPSAP2      | 17p11.2  | -0.43 | -0.49 | 2.3E-15 | -0.21 |
| 200599_s_at | HSP90B1   | 12q23.3  | 0.40 | 0.42 | 4.1E-11 | 0.17 | 228360_at    | LYPD6B       | 2q23.1   | -0.52 | -0.41 | 1.1E-10 | -0.21 |
|             |           |          |      |      |         |      | 212590_at    | RRAS2        | 11p15.2  | -0.48 | -0.44 | 3.2E-12 | -0.21 |
|             |           |          |      |      |         |      | 227354_at    | PAG1         | 8q21.13  | -0.48 | -0.43 | 5.5E-12 | -0.21 |
|             |           |          |      |      |         |      | 205922_at    | VNN2         | 6q23.2   | -0.48 | -0.43 | 5.2E-12 | -0.21 |
|             |           |          |      |      |         |      | 212646_at    | RFTN1        | 3p25.1   | -0.50 | -0.42 | 3.7E-11 | -0.21 |
|             |           |          |      |      |         |      | 222942_s_at  | TIAM2        | 6q25.2   | -0.46 | -0.44 | 1.2E-12 | -0.20 |
|             |           |          |      |      |         |      | 232103_at    | BPNT1        | 1q41     | -0.47 | -0.43 | 9.1E-12 | -0.20 |
|             |           |          |      |      |         |      | 1569481_s_at | SNX22        | 15q22.31 | -0.45 | -0.44 | 1.3E-12 | -0.20 |
|             |           |          |      |      |         |      | 221781_s_at  | DNAJC10      | 2q32.1   | -0.43 | -0.46 | 1.6E-13 | -0.20 |
|             |           |          |      |      |         |      | 236533_at    | ASAP1        | 8q24.21  | -0.44 | -0.44 | 1.7E-12 | -0.20 |
|             |           |          |      |      |         |      | 203521_s_at  | ZNF318       | 6p21.1   | -0.46 | -0.43 | 1.0E-11 | -0.19 |
|             |           |          |      |      |         |      | 201425_at    | ALDH2        | 12q24.12 | -0.47 | -0.41 | 7.2E-11 | -0.19 |
|             |           |          |      |      |         |      | 224791_at    | ASAP1        | 8q24.21  | -0.43 | -0.44 | 1.8E-12 | -0.19 |
|             |           |          |      |      |         |      | 204890_s_at  | LCK          | 1p35.1   | -0.43 | -0.44 | 2.4E-12 | -0.19 |
|             |           |          |      |      |         |      | 215886_x_at  | USP12        | 13q12.13 | -0.46 | -0.41 | 8.6E-11 | -0.19 |
|             |           |          |      |      |         |      | 212974_at    | DENND3       | 8q24.3   | -0.40 | -0.46 | 1.8E-13 | -0.19 |
|             |           |          |      |      |         |      | 241155_at    |              | 10p12.2  | -0.42 | -0.44 | 3.6E-12 | -0.18 |
|             |           |          |      |      |         |      | 242650_at    |              | 10q11.22 | -0.41 | -0.45 | 1.1E-12 | -0.18 |
|             |           |          |      |      |         |      | 235242_at    |              | 2p16.1   | -0.43 | -0.42 | 1.8E-11 | -0.18 |
|             |           |          |      |      |         |      | 201201_at    | CSTB         | 21q22.3  | -0.40 | -0.45 | 6.2E-13 | -0.18 |
|             |           |          |      |      |         |      | 212589_at    | RRAS2        | 11p15.2  | -0.40 | -0.44 | 1.4E-12 | -0.18 |
|             |           |          |      |      |         |      | 220694_at    | ASAP1-IT1    | 8q24.21  | -0.40 | -0.42 | 3.2E-11 | -0.17 |
|             |           |          |      |      |         |      | 1563513_at   | SYTL4        | Xq22.1   | -0.40 | -0.40 | 2.2E-10 | -0.16 |
|             |           |          |      |      |         |      | 221496_s_at  | TOB2         | 22q13.2  | -0.40 | -0.40 | 1.8E-10 | -0.16 |

Table III.3.2.2.a) Top genes in validated effect  $v=129$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx<sup>TM</sup>[97] v33)  
 $a_{129}^{g,c}$  Components of the consensus gene axis of effect  $v=129$  (cf. Table III.1.5); filtered  $|a_{129}^{g,c}| \geq 0.4$ .  
 $r_{129}^{g,c}, p_{129}^{g,c}$  Consensus gene correlations of  $v=129$ ; filtered  $|r_{129}^{g,c}| \geq 0.4$  and  $p$  values for the correlations (cf. II.5.2.1).

### Role in survival analysis

Consistent with known differences in average survival of ABC DLBCL and GCB DLBCL, effect  $v=129$  has been found to *predict the second-most significant survival differences* between all DLBCL patients with  $p = 1.1 \cdot 10^{-12}$  (Figure III.2.5.1.a) (and if restricted to R-CHOP treated patients still with  $p = 1.8 \cdot 10^{-9}$  (Figure III.2.2.1.a)). It has also been selected and analyzed as primary variable for a bivariate Cox proportional hazard survival model (III.2.3) that can predict strong survival differences both for R-CHOP and CHOP-treated patients (Figure III.2.3.2).

### Clinical associations

Besides significant association with DLBCL subtype in all four clinical cohorts, the effect is also significantly associated with IPI scores (both R-CHOP treated cohorts have IPI annotations and both are significantly associated:  $p_{GSE10846.R-CHOP} = 6.2 \cdot 10^{-3}$  and  $p_{GSE31312} = 7.8 \cdot 10^{-5}$ ). This is not untypical for effects with strong predictive performance and indicates that some molecularly explainable survival differences can already be predicted by macroscopic clinical variables underlying the IPI score.



## Genomic associations

Gene ontology analysis did not reveal any specific and significantly overrepresented terms; some terms like sequence-specific DNA binding (cf. [GO:0043565](https://doi.org/10.1093/bioinformatics/btu078) [78]) are significant, but not very specific (8/62 top genes belong to this term comprised of 465/20370 measured genes;  $p = 8.0 \cdot 10^{-5}$  via hypergeometric test). (All gene ontology analyses for this effect are available at [D=Interpretation\genomic\v129\GOA](#).)

Signature analyses for  $v = 129$  were able to confirm rediscovery of the COO induced effect, as several previously published DLBCL subtype signatures are significantly and strongly enriched for its top genes (Table III.3.2.2).

| Signatures            |  |           |            | GSEA             |        |      | Basic Statistics      |                          |                 |              |
|-----------------------|--|-----------|------------|------------------|--------|------|-----------------------|--------------------------|-----------------|--------------|
| Signatures DB         | Signature Name   | # defined | # measured | Enrichment score | $p$    | FDR  | Mean $\log_2$ (ratio) | $p$ (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| StaudtSigDB_dNov2012  | ABC_gt_GCB_LC  | 15        | 15         | 0.946            | 0.0019 | 0.2% | -0.887                | 5.9E-07                  | 100.0%          | 0.0%         |
| StaudtSigDB_dNov2012  | ABC_gt_GCB_PMBL_MCLBL_U133AB                               | 53        | 52         | 0.840            | 0.0019 | 0.2% | -0.700                | 1.3E-17                  | 100.0%          | 0.0%         |
| StaudtSigDB_dNov2012  | ABC_gt_GCB_Affy  | 20        | 20         | 0.871            | 0.0020 | 0.2% | -0.645                | 2.3E-06                  | 90.1%           | 9.9%         |
| StaudtSigDB_dNov2012  | ABCgtGCB_U133AB  | 286       | 281        | 0.701            | 0.0019 | 0.2% | -0.527                | 1.5E-70                  | 98.9%           | 1.1%         |
| GeneSigDB_v4_Sept2011 | Lymphoma_Poulson05_48genes                                 | 48        | 47         | 0.727            | 0.0019 | 0.2% | -0.501                | 1.9E-10                  | 94.2%           | 5.8%         |
| MolSigDBv4_0_dMay2014 | Reactome_purine_ribonucleoside_monop_Hosphate_biosynthesis | 11        | 11         | 0.673            | 0.0020 | 0.2% | -0.090                | 0.0031                   | 97.6%           | 2.4%         |
| MolSigDBv4_0_dMay2014 | Spindle_organization_and_biogenesis                        | 11        | 11         | 0.744            | 0.0020 | 0.2% | -0.040                | 0.3830                   | 58.5%           | 41.5%        |
| MolSigDBv4_0_dMay2014 | Tsai_dnajb4_targets_up                                     | 13        | 13         | -0.766           | 0.0022 | 0.2% | 0.073                 | 0.3299                   | 39.5%           | 60.5%        |
| MolSigDBv4_0_dMay2014 | Biocarta_tcytotoxicpathway                                 | 14        | 12         | -0.783           | 0.0020 | 0.2% | 0.059                 | 0.2917                   | 32.2%           | 67.8%        |
| MolSigDBv4_0_dMay2014 | Biocarta_theelper_pathway                                  | 14        | 12         | -0.764           | 0.0020 | 0.2% | 0.078                 | 0.1434                   | 22.7%           | 77.3%        |
| MolSigDBv4_0_dMay2014 | Module_293   | 12        | 11         | -0.795           | 0.0020 | 0.2% | 0.064                 | 0.0548                   | 21.5%           | 78.5%        |
| MolSigDBv4_0_dMay2014 | Module_143   | 14        | 13         | -0.774           | 0.0021 | 0.2% | 0.107                 | 0.0810                   | 21.3%           | 78.7%        |
| MolSigDBv4_0_dMay2014 | Reactome_translocation_of_zap_70_to_immunologicalsynapse   | 14        | 11         | -0.682           | 0.0020 | 0.2% | 0.131                 | 0.2075                   | 24.4%           | 75.6%        |
| HGNCSigDB_dMay2014    | Histocompatibility_complex                                 | 44        | 21         | -0.791           | 0.0019 | 0.2% | 0.144                 | 0.0039                   | 7.1%            | 92.9%        |
| MolSigDBv4_0_dMay2014 | Biocarta_blymphocytepathway                                | 11        | 11         | -0.771           | 0.0021 | 0.2% | 0.190                 | 0.0889                   | 20.0%           | 80.0%        |
| MolSigDBv4_0_dMay2014 | Reactome_phosphorylation_of_cd3_and_tcr_zeta_chains        | 16        | 13         | -0.693           | 0.0021 | 0.2% | 0.227                 | 0.0342                   | 7.3%            | 92.7%        |
| StaudtSigDB_dNov2012  | GC_B_cell_Blow_DLBLhigh                                    | 49        | 48         | -0.690           | 0.0019 | 0.2% | 0.521                 | 1.1E-06                  | 11.8%           | 88.2%        |
| GeneSigDB_v4_Sept2011 | Lymphoma_Chin09_65genes                                    | 47        | 47         | -0.734           | 0.0019 | 0.2% | 0.655                 | 1.1E-09                  | 3.6%            | 96.4%        |
| StaudtSigDB_dNov2012  | GCB_gt_ABC_U133plus  | 307       | 298        | -0.752           | 0.0021 | 0.2% | 0.598                 | 1.4E-75                  | 0.4%            | 99.6%        |
| GeneSigDB_v4_Sept2011 | Lymphoma_Tome05_151genes                                   | 46        | 44         | -0.823           | 0.0019 | 0.2% | 0.816                 | 1.5E-12                  | 4.4%            | 95.6%        |
| StaudtSigDB_dNov2012  | Germinal_center_Bcell_DLBL                                 | 59        | 55         | -0.792           | 0.0021 | 0.2% | 0.780                 | 2.6E-15                  | 0.0%            | 100.0%       |
| StaudtSigDB_dNov2012  | Germinal_center_B_cell_DLBL-survival_predictor             | 37        | 34         | -0.867           | 0.0020 | 0.2% | 0.945                 | 2.0E-11                  | 0.4%            | 99.6%        |
| StaudtSigDB_dNov2012  | GCB_gt_ABC_LC  | 11        | 10         | -0.980           | 0.0020 | 0.2% | 1.395                 | 0.0001                   | 0.0%            | 100.0%       |

Table III.3.2.2.b) Top-enriched signatures by  $v = 129$

Signatures with  $|\text{enrichment score}| \geq 0.67$  and at least 10 measured members are listed for genes ranked by GEP effect  $v = 129$ . All GSEA  $p$  values are based on permutation tests; hence, they are lower-bounded by  $1/(\#\text{permutations})$  and true  $p$  values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of  $t$ -tests against zero regulation as weights).

Enrichment plots (Figure III.3.2.2.c) for the two largest top signatures [ABCgtGCB\\_U133AB](#) (online interpretation card) and [GCB\\_gt\\_ABC\\_U133plus](#) (online interpretation card) visualize their significance and again confirm rediscovery of the standard COO induced effect.

## The rediscovery of the COO induced effect by correlations provides a filtered view on subtype-specific genes

Gene selection criteria for these DLBCL signatures show that they are based on (a)  $t$ -tests between sample subsets by previously predicted subtypes, (b) deselection of genes correlated with  $r > 0.2$  to a proliferation signature and (c) in case of [GCB\\_gt\\_ABC\\_U133plus](#) also deselection of genes correlated with  $r > 0.2$  to a lymph node signature.

Using previously predicted subtypes as basis for gene signature definition can only test every gene individually, but does not examine correlations between genes. Given the low information base (every subtype annotation essentially carries only one bit of information per patient as there are only two subtypes), criterion (a) probably selects many genes that are biologically *unspecific* with respect to DLBCL subtype

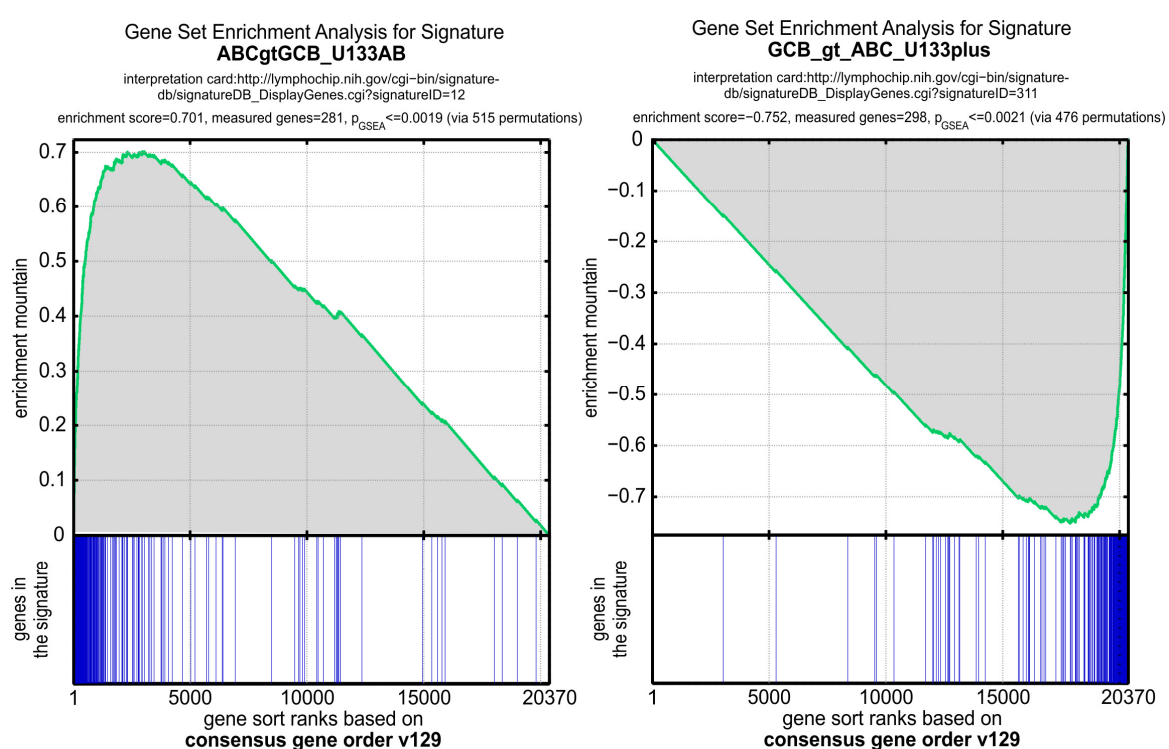


Figure III.3.2.2.c) Significant enrichment of known ABC-versus-GCB DLBCL signatures by effect  $\nu = 129$

biology. Deselecting genes with respect to other signatures (i.e. criteria (b) and (c)) has probably been performed in order to filter some of these unspecific genes out again. However, correlation between remaining genes is still ignored and thus many genetically distinct effects are included in and mixed by resulting gene signatures. Such signatures were also utilized by previous subtype classifiers<sup>[83]</sup>.

In contrast, signal dissection forms effects based on correlation maximization<sup>(cf. II.3.2)</sup>, i.e. top genes in every effect are as highly correlated to each other as permitted by the signal. Hence, dissection into effects with maximal inter-gene correlations can help to dissect pathways that were only visible in overlapped and mixed form in previous gene signatures.

Interestingly, most other discovered GEP effects that are significantly related to DLBCL subtypes have rather uncorrelated gene axes to effect  $\nu = 129$ , i.e. their top genes are rather different from each other<sup>(cf. Figure III.3.2.2.a)</sup>. This indicates that the unsupervisedly rediscovered GEP effect  $\nu = 129$  does not only match previous subtypes form four independent cohorts best, but can also be considered a *genetically filtered* redefinition of above ABC-versus-GCB-DLBCL gene signatures and presumably of corresponding subtype biology.

Still, the known genetic heterogeneity of DLBCL remains even for this filtered view in form of a rough signal<sup>(see Figure III.3.2.2.b)</sup>; thus correlations of top genes to the effect (and thus to each other) are only moderate. Other discovered effects with strong survival impact<sup>(III.3.3)</sup> like  $\nu = 134$  are smaller and possess some top genes of higher correlation. Assuming that higher correlations between genes indicate tighter biological relations, these effects could be biologically even more specific than this filtered redefinition of the COO induced effect in form of  $\nu = 129$ .

### ■ Preliminary top genes analysis

For probesets with  $|r_{129}^g| \geq 0.4$ , 151 unique genes from 22 chromosomes participate in the effect, suggesting that it reflects a functional genomic network, rather than just reflecting local aberrations on DNA level. For  $|r_{129}^g| \geq 0.5$ , 19 unique genes from 12 chromosomes remain.

Most top genes of  $v = 129$  are contained in the two above-mentioned ABC-versus-GCB-DLBCL gene signatures. Several have already been investigated. To illuminate the filtered redefinition of the COO induced effect by  $v = 129$ , an individual review of its top-correlated genes in context of existing research is indicated. To start, the most specific genes for  $v = 129$  are briefly presented here.

FOXP1 is the best-correlated gene with  $r_{129,FOXP1}^g = 0.69$ ; higher correlations do not exist due to the relatively rough signal of this effect. This gene has already been identified for DLBCL survival predictor models before<sup>[29]</sup> and a recent study<sup>[102]</sup> disclosed its molecular function as oncogene in lymphomas relying on NF- $\kappa$ B activity. It directly represses transcription of seven pro-apoptotic genes and its aberrant expression can complement (constitutive) NF- $\kappa$ B activity, which in total may contribute to lymphomagenesis<sup>[102]</sup>.

KIAA0226L is the second-best-correlated gene with  $r_{129,KIAA0226L}^g = 0.63$ . No literature that directly relates this gene to DLBCL has been found, but a study in molecular oncology<sup>[103]</sup> observed silencing of KIAA0226L (aka C13orf18) through hypermethylation in cervical cancer. Its re-expression via artificial gene-specific transcription factors significantly inhibited cell growth and/or induced apoptosis. However, this cannot explain its role in DLBCL, because KIAA0226L is expressed higher in ABC DLBCL, i.e. for patients associated with adverse outcome.

On the anti-correlated side, for example LMO2 with  $r_{129,LMO2}^g = -0.54$  is known as a powerful prognostic indicator in DLBCL<sup>[104]</sup>. It is specifically upregulated in the germinal center and a study on LMO2's interactome<sup>[104]</sup> revealed that it increases transcriptional activity of NFATc1. An immunohistochemical study<sup>[105]</sup> on LMO2 confirmed its exclusive expression in GCB DLBCL (20/20) and negativity in ABC DLBCL (0/15). This gene is also expressed in Hodgkin lymphoma cases (23/23) and in Burkitt's lymphoma (9/10). Regarding healthy tissues, it was exclusively found in the germinal center, but not in mantle, marginal and T cell zones.

### ■ Inference

Several genes of this rediscovered COO induced effect have already been biologically investigated with respect to their contribution to DLBCL lymphomagenesis. The functional interplay of identified top-correlated genes and the role of KIAA0226L in particular may be interesting for further investigation. However, genes in GEP effect  $v = 134$  presented next<sup>(III.3.3.1)</sup> might be more interesting, as  $v = 134$  predicts survival most consistently and is comprised of fewer and higher correlated top genes (for the same correlation cutoff).



Multivariate survival analysis<sup>(III.2)</sup> has revealed that several discovered GEP effects and combinations thereof are significantly associated with observed patient outcome. The COO induced effect<sup>(cf. III.3.2.2)</sup> has been shown to be only one among several<sup>(cf. Figure III.2.5.1.a)</sup>. For interpretation of constructed survival predictors, relevant GEP effects are presented and biostatistically evaluated here.

Some effects could be identified as unsupervisedly rediscovered versions of genomic entities that are known from previous DLBCL studies. Some effects are novel. Their biostatistical evaluations can provide a basis for their biological interpretation.



The discussion of each effect starts with a summary of its role in survival analysis. Then its definition and an overview of its signal are provided, both in graphical and tabular form. Subsequently significant genomic associations or clinical associations are presented. If indicated, this is followed by a preliminary discussion of specific top correlated genes. An effect-specific outlook may infer analytically promising experimental investigations that might eventually lead to advances in modeling the molecular pathogenesis of DLBCL.

## III.3.3.1

 $\nu = 134$ : Primary survival effect in DLBCL■ *Role in survival analysis*

Already during detection and before any systematic survival analyses, one GEP effect emerged by showing an obvious sorting of patients by outcome. A systematic comparison<sup>(Figure III.2.5.1.a)</sup> showed that this effect can explain survival more consistently (with  $p = 4.5 \cdot 10^{-17}$ ) compared to the COO induced effect<sup>(cf. III.3.2.2)</sup> ( $p = 1.1 \cdot 10^{-12}$ ), despite consisting of only 69 unique correlated top genes, whereas the COO induced effect has 151 unique top genes for an identical correlation cutoff ( $|r| \geq 0.4$ ). This might indicate that it is biologically more specific to DLBCL. Another indication for that is the existence of a hierarchical survival effect that strongly affects only one risk partition of  $\nu = 134$ , whereas the other partition is unaffected<sup>(III.2.2.3)</sup>. Using the COO induced effect as primary predictor variable did not disclose such hierarchical dependencies<sup>(cf. III.2.3.1 and III.2.3.3)</sup>. Additionally,  $\nu = 134$  also ranks first as primary predictor variable when training only with samples from R-CHOP treated patients<sup>(cf. III.2.2)</sup>. Because of these properties, it is evaluated here in detail.

■ *Effect overview*

Effect  $\nu = 134$  was originally detected and dissected in the GEP signal of cohort GSE31312 at rank #27. It has been supervisedly validated on GEP level in all three other cohorts<sup>(e.g. Figure III.3.3.1.b)</sup>. (See 134\* files in  *C=Consensus Effects\{cohort subfolders\}*.) Additionally, it validates against survival data in all four cohorts (see  *D=Interpretation\clinical\134\{cohort subfolders\}*).

Its application to the largest patient cohort illustrates<sup>(Figure III.3.3.1.a)</sup> that this effect is comprised of approximately 100 correlated and only few anti-correlated probesets (with relative correlation  $\geq 0.5$ ). Survival as indicated by censored/progression information<sup>(green/orange)</sup> shows that *higher expression of correlated genes is associated with better patient outcome*. On the upregulated side, GCB DLBCL patients are overrepresented, as could be expected due to higher average survival of GCB DLBCL. Comparison of subtype information<sup>(yellow/pink)</sup> shows that the sample order by  $\nu = 134$  is partially correlated, but not identical to the COO induced sample order<sup>(cf. Figure III.3.2.2.b)</sup>. More importantly, only a weak correlation between this effect's consensus gene scores and the COO consensus gene scores exists ( $r_{(134;129)}^C = -0.29$ <sup>(cf. Eqn. III.1.3.2.b)</sup>), demonstrating that these effects are based on different top genes.



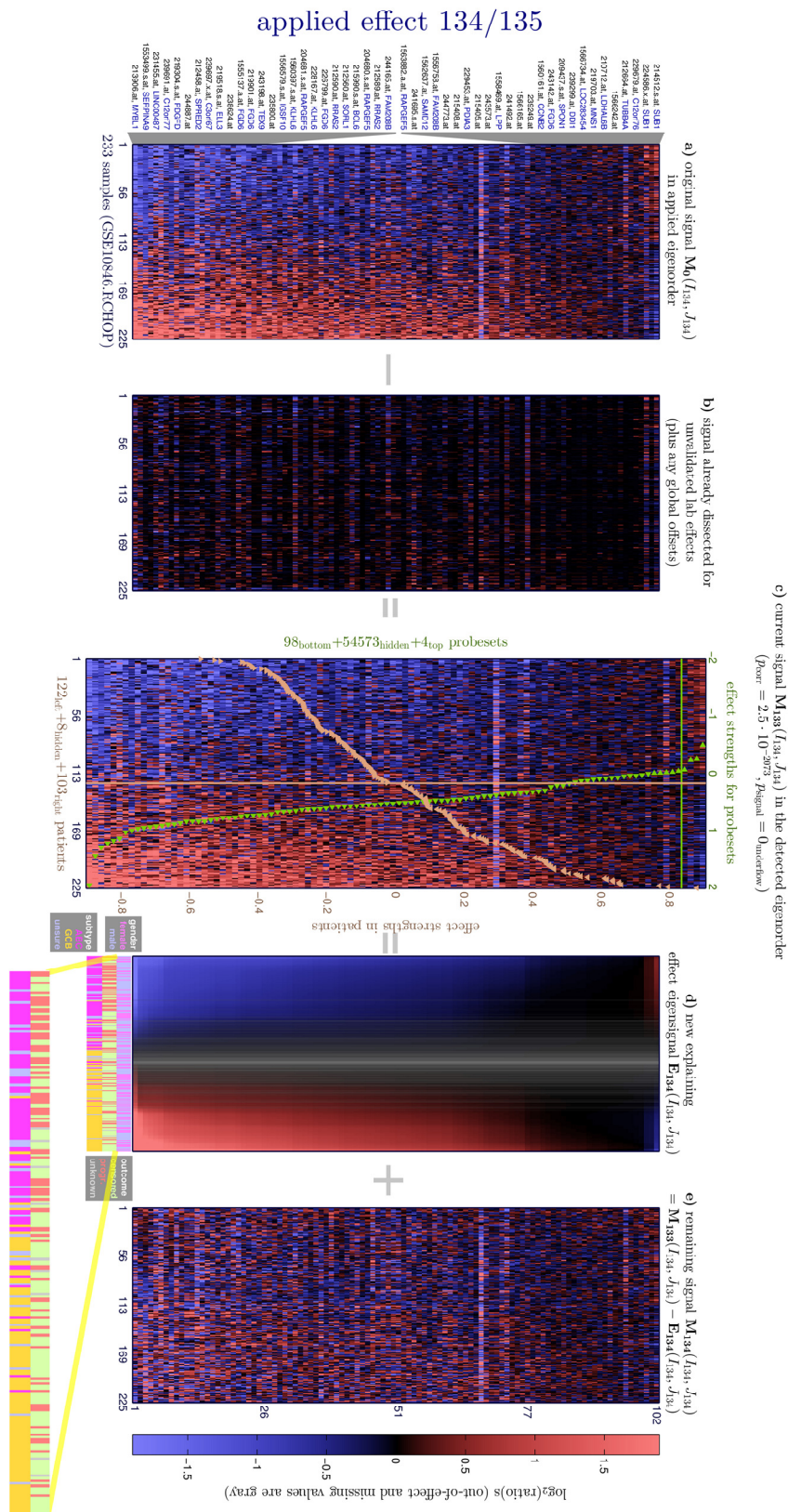


Figure III.3.3.1.b) Validated effect  $v=134$ , applied to GSE10846.R-CHOP

Primary survival effect by multi-cohort survival analysis<sup>(II.2.5)</sup> applied to GSE10846.R-CHOP (233 patients). Enlarged binary classifications show censored/progression follow-up information (green/orange) as well as public subtypes (yellow/pink).

(The genomic consensus effect is applied to the cleaned signal without lab effects<sup>(cf III.1.4.2)</sup>. Samples and probesets are ordered by their effect strengths in this cohort<sup>(cf Table III.1.5)</sup>. Additionally, probesets are filtered by demanding a relative correlation stronger than 0.5. The effect's bimonotonic eigensignal<sup>[panel d)</sup> is grayed for samples having insufficient or insignificant correlation to this effect<sup>(II.4.2.1)</sup>.)

Similar effect plots and definition tables are available for all four cohorts in  $C=Consensus\ Effects\ (cohort\ subfolder)$ ; they supervisedly validate the effect's existence in each cohort. All probesets with a relative correlation  $\geq 0.5$  to the consensus sample axis are depicted; they are ordered by their differential signal between samples. While this signal ordering is necessary for bimonotonic regression and effect dissection, highly correlated genes with relatively weak differential signal may also be biologically important to understand underlying pathogenic pathways<sup>(III.1.3.2)</sup>. For this reason, cohort-independent genomic analyses are based on both differential expression *and* correlation strength<sup>(cf. Table III.1.5)</sup>.

### Clinical associations

As the COO induced effect before, this effect is significantly associated with IPI scores, albeit slightly less so (both R-CHOP treated cohorts have IPI annotations and both are significantly associated:  $p_{GSE10846.R-CHOP} = 3.8 \cdot 10^{-3}$  and  $p_{GSE31312} = 0.03$ ). Again, this is not untypical for effects with strong predictive performance and indicates that some molecularly explainable survival differences can already be predicted by macroscopic clinical variables underlying the IPI score.

### Genomic associations

Using the same gene scores for ranking, gene set enrichment analyses revealed only six significantly enriched signatures (filtering by  $|\text{enrichment score}| \geq 0.67$  as before). As could be expected by rank #4 for the association with subtypes<sup>(cf. Figure III.3.2.2.a)</sup>, again several known ABC-versus-GCB-DLBCL gene signatures are significantly enriched<sup>(Table III.3.3.1.a)</sup>. However, one should be cautious not to overinterpret significant enrichment as it is not sufficient for high correlation of signature genes to an effect<sup>(cf. III.3.4.1)</sup>.

| Signatures            |   |           |            | GSEA             |          |      | Basic Statistics              |                               |                 |              |
|-----------------------|---|-----------|------------|------------------|----------|------|-------------------------------|-------------------------------|-----------------|--------------|
| Signatures DB         | Signature Name                                  | # defined | # measured | Enrichment score | <i>p</i> | FDR  | Mean log <sub>2</sub> (ratio) | <i>p</i> (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| StaudtSigDB_dNov2012  | GCB_gt_ABC_LC                                   | 11        | 10         | 0.849            | 0.0020   | 0.2% | -1.346                        | 0.0009                        | 100.0%          | 0.0%         |
| StaudtSigDB_dNov2012  | Germinal_center_B_cell_DLBCL-survival_predictor | 37        | 34         | 0.765            | 0.0021   | 0.2% | -0.987                        | 4.9E-11                       | 98.3%           | 1.7%         |
| StaudtSigDB_dNov2012  | Germinal_center_Bcell_DLBCL                     | 59        | 55         | 0.741            | 0.0020   | 0.2% | -0.905                        | 1.4E-15                       | 97.0%           | 3.0%         |
| GeneSigDB_v4_Sept2011 | Lymphoma_Tome05_151genes                        | 46        | 44         | 0.736            | 0.0019   | 0.2% | -0.957                        | 2.9E-13                       | 95.6%           | 4.4%         |
| StaudtSigDB_dNov2012  | ABC_gt_GCB_LC                                   | 15        | 15         | -0.678           | 0.0019   | 0.2% | 0.439                         | 5.6E-05                       | 0.0%            | 100.0%       |
| MolSigDBv4_0_dMay2014 | Xu_response_to_tretinoin-and_nsc682994_dn       | 15        | 15         | -0.688           | 0.0020   | 0.2% | 0.053                         | 0.2324                        | 38.7%           | 61.3%        |

Table III.3.3.1.a) Top-enriched signatures by  $v = 134$

Signatures with  $|\text{enrichment score}| \geq 0.67$  and at least 10 measured members are listed for genes ranked by GEP effect  $v = 134$ .

All GSEA *p* values are based on permutation tests; hence, they are lower-bounded by  $1/(\#\text{permutations})$  and true *p* values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of *t*-tests against zero regulation as weights).

Signatures with a positive enrichment score are more relevant for interpretation, as the effect consists predominantly of top correlated and only of few and relatively weakly anti-correlated genes. None of the four top enriched signatures does contain all top genes of  $v = 134$ ; however, they contain genes that are only weakly correlated to the effect. This implies that  $v = 134$  contains a subset of known subtype-specific genes that are highly correlated to each other, but so far scattered over various different gene signatures or embedded in larger and less specific signatures. To further elucidate biological implications of this, an individual review of top genes of  $v = 134$  is indicated.

Like for the COO induced effect<sup>(III.3.2.2)</sup>, gene ontology analysis did not reveal any specific and significantly overrepresented terms; some terms like the nuclear envelope<sup>(cf. GO:0005635)</sup> are significant, but not specific (4/44 top genes belong to this term, but it is comprised of 109/20370 measured genes;  $p = 8.9 \cdot 10^{-5}$  via hypergeometric test).

### ■ Preliminary top genes analysis

For probesets with  $r_{134}^g \geq 0.4$ , 69 unique genes from 19 chromosomes participate in the effect, suggesting that this effect reflects a functional genomic network, rather than just reflecting local aberrations on DNA level. For  $r_{134}^g \geq 0.5$ , 19 unique genes from 7 chromosomes remain.

The most specific effect gene is FGD6 with  $r_{134,FGD6}^g = 0.84$  (for its top correlated probeset). Its association with the effect is verified by five independently measured and highly correlated probesets. FGD6 is located in 12q22 and is a protein coding gene with validated sequence as per RefSeq<sup>[106]</sup> status (information via <http://www.ncbi.nlm.nih.gov/gene>, [107], April 2015). Literature screening did not reveal any direct association with DLBCL. A retrospective medical study<sup>[108]</sup> about toxic side effects of several groups of chemotherapeutic agents revealed that a single-nucleotide polymorphism within FGD6 (adenine instead of guanine at base pair 95490248) can cause severe neutropenia/leucopenia for patients of various cancers when administered with Paclitaxel and carboplatin agents ( $p = 2.46 \cdot 10^{-7}$ ). This might indicate a possible functional role of FGD6 in lymphocytes and maybe for B cells.

KLHL6 follows at correlation rank #2 ( $r_{134,KLHL6}^g = 0.75$  for its top probeset); its association with this effect is confirmed by four independently measured probesets. It is located in 3q27.1 and is also a protein coding gene with validated RefSeq status. A related molecular immunology study<sup>[109]</sup> compared gene expressions of ex-vivo B cells from sheep that undergo hypermutation during antigen-independent development with in-vitro B cells having the same process arrested. (Normally, antigen-independent hypermutation builds the primary antibody reservoir.) The study revealed that KLHL6 might be involved in the germinal center B-cell differentiation pathway. Furthermore, KLHL6 was exclusively expressed in lymphoid tissues compared to several other human tissues. Like BCL6 (that is also among top genes of  $v = 134$ ), the KLHL6 protein contains a domain that is known for transcriptional repression activity. Combining several other results like the relevance of receptor cross-linking in the germinal center differentiation program, it was hypothesized that KLHL6 may be involved in re-modeling actin micro-filaments during germinal center differentiation.

LPP follows at correlation rank #3 ( $r_{134,LPP}^g = 0.74$  for its top probeset; supported by four other correlated probesets). It is a protein coding gene with reviewed RefSeq status and is located in 3q28. A related high resolution genome-wide association study<sup>[110]</sup> revealed a susceptibility locus in the intergenic region between BCL6 and LPP on 3q27 (of length ~400kB) for non-Hodgkin lymphoma in Chinese population. This locus is associated with increased risk, especially in DLBCL ( $p = 1.14 \cdot 10^{-11}$ ), but not in non-B cell lymphomas.

| Top genes in $v=134$ |             |            |                 |                 |                 |  |
|----------------------|-------------|------------|-----------------|-----------------|-----------------|--|
| Probeset             | Gene Symbol | HG19 locus | $a_{134}^{g,c}$ | $r_{134}^{g,c}$ | $p_{134}^{g,c}$ | $\left  \frac{a_{134}^{g,c}}{r_{134}^{g,c}} \right $ |
| 1555137_a_at         | FGD6        | 12q22      | 1.18            | 0.84            | 1.3E-82         | 0.99   |
| 228167_at            | KLHL6       | 3q27.1     | 1.05            | 0.75            | 1.8E-55         | 0.79   |
| 243142_at            | FGD6        | 12q22      | 0.94            | 0.83            | 7.6E-79         | 0.79   |
| 241879_at            | LPP         | 3q28       | 1.03            | 0.72            | 1.2E-48         | 0.74   |
| 202822_at            | LPP         | 3q28       | 1.00            | 0.74            | 5.4E-53         | 0.74   |
| 1555275_a_at         | KLHL6       | 3q27.1     | 1.06            | 0.69            | 3.7E-44         | 0.73   |
| 226799_at            | FGD6        | 12q22      | 0.89            | 0.82            | 7.8E-73         | 0.72   |
| 235000_at            | LPP         | 3q28       | 0.92            | 0.72            | 7.2E-50         | 0.67   |
| 1555136_at           | FGD6        | 12q22      | 0.96            | 0.65            | 6.8E-37         | 0.62   |
| 224811_at            | LPP         | 3q28       | 0.92            | 0.67            | 5.2E-40         | 0.61   |
| 1556579_s_at         | IGSF10      | 3q25.1     | 1.02            | 0.58            | 3.5E-28         | 0.59   |
| 240866_at            |             | 3q28       | 0.82            | 0.72            | 2.3E-48         | 0.59   |
| 219901_at            | FGD6        | 12q22      | 0.71            | 0.75            | 5.8E-56         | 0.53   |
| 241695_s_at          |             | 3q27.1     | 0.72            | 0.73            | 3.7E-51         | 0.52   |
| 1560397_s_at         | KLHL6       | 3q27.1     | 0.84            | 0.57            | 3.4E-27         | 0.48   |
| 1560396_at           | KLHL6       | 3q27.1     | 0.92            | 0.50            | 1.8E-20         | 0.46   |
| 219304_s_at          | PDGFD       | 11q22.3    | 0.74            | 0.59            | 2.4E-29         | 0.44   |
| 239697_x_at          | C3orf67     | 3p14.2     | 0.84            | 0.52            | 6.2E-22         | 0.44   |
| 243573_at            |             | 3q28       | 0.61            | 0.70            | 8.0E-46         | 0.43   |
| 231455_at            | LINC00487   | 2p25.2     | 0.92            | 0.45            | 1.8E-16         | 0.42   |
| 1569344_a_at         |             | 7p21.1     | 0.85            | 0.47            | 3.9E-18         | 0.40   |
| 1562637_at           | SAMD12      | 8q24.12    | 0.69            | 0.55            | 7.3E-25         | 0.38   |
| 212458_at            | SPRED2      | 2p14       | 0.64            | 0.54            | 7.5E-24         | 0.35   |
| 1558469_at           | LPP         | 3q27.3     | 0.60            | 0.56            | 3.0E-26         | 0.34   |
| 213906_at            | MYBL1       | 8q13.1     | 0.67            | 0.50            | 1.7E-20         | 0.33   |
| 218331_s_at          | FAM208B     | 10p15.1    | 0.71            | 0.47            | 1.2E-17         | 0.33   |
| 218862_at            | ASB13       | 10p15.1    | 0.69            | 0.47            | 5.4E-18         | 0.32   |
| 240144_at            | DNASE1      | 16p13.3    | 0.63            | 0.50            | 6.4E-20         | 0.31   |
| 238181_at            |             | 1q31.2     | 0.67            | 0.47            | 1.8E-17         | 0.31   |
| 235521_at            | HOXA3       | 7p15.2     | 0.66            | 0.46            | 4.7E-17         | 0.30   |
| 217966_s_at          | FAM129A     | 1q25.3     | 0.72            | 0.42            | 5.4E-14         | 0.30   |
| 204530_s_at          | TOX         | 8q12.1     | 0.65            | 0.45            | 1.3E-16         | 0.30   |
| 244165_at            | FAM208B     | 10p15.1    | 0.62            | 0.47            | 4.0E-18         | 0.29   |
| 243040_at            |             | 8q24.12    | 0.53            | 0.54            | 7.3E-24         | 0.28   |
| 244887_at            |             | 1q31.2     | 0.65            | 0.43            | 3.9E-15         | 0.28   |
| 235800_at            |             | 10q25.3    | 0.63            | 0.44            | 1.3E-15         | 0.28   |
| 243198_at            | TEX9        | 15q21.3    | 0.55            | 0.49            | 8.2E-20         | 0.27   |
| 212560_at            | SORL1       | 11q24.1    | 0.52            | 0.52            | 1.8E-22         | 0.27   |
| 225997_at            | MOB1B       | 4q13.3     | 0.57            | 0.48            | 1.7E-18         | 0.27   |
| 227354_at            | PAG1        | 8q21.13    | 0.53            | 0.51            | 2.1E-21         | 0.27   |
| 204680_s_at          | RAPGEF5     | 7p15.3     | 0.52            | 0.51            | 1.3E-21         | 0.27   |
| 210712_at            | LDHAL6B     | 15q22.2    | 0.48            | 0.56            | 4.7E-26         | 0.27   |
| 1560180_at           |             | 2p23.1     | 0.62            | 0.43            | 5.7E-15         | 0.27   |
| 215408_at            |             | 15q22.2    | 0.61            | 0.43            | 6.2E-15         | 0.26   |
| 231442_at            | ZBP2        | 17q12      | 0.55            | 0.48            | 1.2E-18         | 0.26   |
| 1556755_s_at         | LOC286149   | 8q22.1     | 0.54            | 0.47            | 4.5E-18         | 0.26   |

Other top probesets with  $r_{134,241695\_s\_at}^g = 0.73$ ,  $r_{134,240866\_at}^g = 0.72$  and  $r_{134,243573\_at}^g = 0.70$  are located in genomic vicinity of either KLHL6 or LPP. Presumably they measure expressions of the same genes, because they are genomically located *directly* before or after them (with in part overlapping probeset sequence intervals). But they do not have a gene annotation so far. In any case, they seem to be biologically related, because sequences underlying 243573\_at([transcript AA648962](#)) and 241695\_s\_at([transcript AA648986](#)) were both defined with human tonsillar cells that were *enriched for germinal center B cells* by flow sorting([provided by L.M. Staudt, National Cancer Institute, 1997](#)).

The locus 3q27 is known for typical translocations in B-cell lymphoma, but these translocations alone could not predict significant survival differences in DLBCL<sup>[111]</sup>. However, as this study was only based on 14/93 DLBCL patients having 3q27 translocations, the study size might be too small to reach significance on survival level. Anyhow, the low 15% incidence of these 3q27 translocations cannot explain the consistent GEP signal of effect  $\nu = 134$ , as it exists in most samples.

While both BCL6 (3q27.3) and LPP (3q28) are top genes of  $\nu = 134$ , the other top gene KLHL6 in genomic vicinity (3q27.1) is located approximately 4.3MB before them. This indicates that a functional relationship may be needed to establish the correlation to the reported 400kB short susceptibility locus between BCL6 and LPP. To further examine the possibility of a reflection of a chromosomal feature on GEP level, 147 additional probesets have been analyzed that are located between KLHL6 in 3q27.1 and LPP in 3q28. Expressions of these 147 probesets are not correlated to the effect, but absolute expression levels (measured by their average  $\log_2(\text{intensities})$  over all samples) are higher for 35/147 probesets than average  $\log_2(\text{intensities})$  for KLHL6 and LPP (see [DLBCL Master Table 2015, gene orders.xlsx](#)). If a chromosomal feature without any connection to or modulation by a functional genomic network was reflected by this GEP effect, these 35 expressed probesets between KLHL6 and LPP should also be correlated to this effect, but this is not the case.

MYBL1 (8q13.1) is depicted at rank #1 in the effect plot([Figure III.3.3.1.a](#)), but more due to its strongly differential expression, rather than by high correlation to other genes in the effect ( $r_{134,MYBL1}^g = 0.50$ ; this is comparable to MYBL1's moderate correlation to the COO induced effect, i.e.  $r_{129,MYBL1}^g = -0.42$ ). No direct relation of this gene to DLBCL has been found in the literature. However, for diffuse pediatric low-grade gliomas (PLGG), MYBL1 is known as partially duplicated transcription factor based on gains of its 8q13.1 locus<sup>[112]</sup>. (These aberrations result in truncated MYBL1 transcripts. A correspondingly transformed cell line formed tumors in nude mice, whereas the same cell line having full-length MYBL1 wild type constructs could not form any tumors.)

|              |              |          |       |       |         |       |
|--------------|--------------|----------|-------|-------|---------|-------|
| 202821_s_at  | LPP          | 3q28     | 0.48  | 0.53  | 2.7E-23 | 0.26  |
| 203284_s_at  | HS2ST1       | 1p22.3   | 0.58  | 0.44  | 7.8E-16 | 0.26  |
| 215405_at    |              | 15q22.2  | 0.52  | 0.49  | 1.3E-19 | 0.25  |
| 219703_at    | MNS1         | 15q21.3  | 0.52  | 0.49  | 1.6E-19 | 0.25  |
| 1566165_at   |              | 2q31.1   | 0.54  | 0.47  | 8.7E-18 | 0.25  |
| 235171_at    | LOC100505501 | 8q12.1   | 0.51  | 0.49  | 1.1E-19 | 0.25  |
| 203769_s_at  | STS          | Xp22.31  | 0.51  | 0.48  | 5.3E-19 | 0.25  |
| 228464_at    | MIR3685      | 12q22    | 0.44  | 0.55  | 1.0E-24 | 0.24  |
| 229588_at    | DNAJC10      | 2q32.1   | 0.52  | 0.46  | 3.9E-17 | 0.24  |
| 231181_at    |              | 8q23.2   | 0.58  | 0.41  | 1.7E-13 | 0.24  |
| 1556758_at   | FAM208B      | 10p15.1  | 0.49  | 0.48  | 1.8E-18 | 0.23  |
| 244185_at    |              | 12q22    | 0.47  | 0.48  | 1.1E-18 | 0.23  |
| 232471_at    |              | 15q22.2  | 0.49  | 0.47  | 1.5E-17 | 0.23  |
| 1568751_at   | RGS13        | 1q31.2   | 0.55  | 0.41  | 3.0E-13 | 0.22  |
| 239691_at    | C12orf77     | 12p12.1  | 0.46  | 0.48  | 7.8E-19 | 0.22  |
| 220168_at    | CASC1        | 12p12.1  | 0.49  | 0.45  | 5.6E-16 | 0.22  |
| 239249_at    |              | 3q27.3   | 0.41  | 0.51  | 1.2E-21 | 0.21  |
| 213156_at    |              | 3q13.31  | 0.49  | 0.42  | 1.9E-14 | 0.21  |
| 215990_s_at  | BCL6         | 3q27.3   | 0.41  | 0.50  | 1.3E-20 | 0.21  |
| 241492_at    |              | 5q31.3   | 0.45  | 0.45  | 2.0E-16 | 0.20  |
| 232170_at    | SORL1        | 11q24.1  | 0.42  | 0.48  | 1.3E-18 | 0.20  |
| 240777_at    | SYNE2        | 14q23.2  | 0.46  | 0.44  | 2.0E-15 | 0.20  |
| 1554168_a_at | SH3KBP1      | Xp22.12  | 0.45  | 0.44  | 7.2E-16 | 0.20  |
| 203140_at    | BCL6         | 3q27.3   | 0.42  | 0.47  | 3.4E-18 | 0.20  |
| 232125_at    |              | 3q13.31  | 0.47  | 0.41  | 7.4E-14 | 0.20  |
| 219551_at    | EAF2         | 3q13.33  | 0.48  | 0.40  | 5.3E-13 | 0.19  |
| 214276_at    | KLF12        | 13q22.1  | 0.42  | 0.47  | 1.1E-17 | 0.19  |
| 1566242_at   |              | 7q22.1   | 0.41  | 0.46  | 3.9E-17 | 0.19  |
| 212640_at    | PTPLB        | 3q21.1   | 0.41  | 0.45  | 1.4E-16 | 0.18  |
| 225626_at    | PAG1         | 8q21.13  | 0.46  | 0.40  | 4.8E-13 | 0.18  |
| 227713_at    | KATNAL1      | 13q12.3  | 0.45  | 0.41  | 2.7E-13 | 0.18  |
| 1554122_a_at | HSD17B12     | 11p11.2  | 0.44  | 0.41  | 1.4E-13 | 0.18  |
| 239516_at    |              | 1q41     | 0.44  | 0.41  | 2.7E-13 | 0.18  |
| 209967_s_at  | CREM         | 10p11.21 | 0.42  | 0.41  | 7.5E-14 | 0.17  |
| 224586_x_at  | SUB1         | 5p13.3   | -0.44 | -0.49 | 2.9E-19 | -0.21 |
| 1566734_at   | LOC283454    | 12q24.22 | -0.54 | -0.44 | 8.2E-16 | -0.24 |
| 212664_at    | TUBB4A       | 19p13.3  | -0.68 | -0.49 | 7.2E-20 | -0.34 |

Table III.3.3.1.b) Top genes in validated effect  $\nu=134$

|                 |  |
|-----------------|--|
| (probesets)     | from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx <sup>TM</sup> [97] v33) |
| $a_{134}^{g,c}$ | Components of the consensus gene axis of effect $\nu=134$ (cf. Table III.1.5); filtered $\geq 0.4$ . |
| $r_{134}^{g,c}$ | Consensus gene correlations of $\nu=134$ ; filtered $\geq 0.4$ .                                     |
| $p_{134}^{g,c}$ | $p$ values for the correlations (cf. II.5.2.1)   |

For comparison, the probesets discussed for the three top genes are correlated only with  $r_{129,FGD6}^g = -0.22$ ,  $r_{129,KLHL6}^g = -0.27$  and  $r_{129,LPP}^g = -0.25$  to the COO induced effect  $\nu = 129$ , underlining that these two GEP effects describe genetically distinct or at the most partially correlated biology.

### ■ Inference

As KLHL6 proteins contain a domain that is known for transcriptional repression activity and might be involved in the germinal center B-cell differentiation pathway<sup>[109]</sup>, it could potentially be a tumor suppressor gene whose expression is required to switch off the hypermutation program in a subset of DLBCL cells. This would be consistent with significantly adverse patient outcome for lower KLHL6 expression. As they are linked by high GEP correlation, the same biological function might be associated with FGD6 and other top-correlated genes of  $\nu = 134$ .

These hypotheses about potential tumor suppressor genes could possibly be investigated by overexpression experiments in DLBCL cell lines that show low expression of effect  $\nu = 134$ . To identify these cell lines, they could be screened for their protein levels of KLHL6 and FGD6. Ideally, the proliferation of cells with low or nonexistent levels of these proteins can be stopped by corresponding overexpression experiments.

### III.3.3.2 $\nu \in \{127, 131\}$ : A hierarchical survival effect prevailing in GCB DLBCL

#### ■ Role in survival analysis

Effect  $\nu = 127$  is the best secondary predictor variable with  $p = 5.6 \cdot 10^{-8}$  on top of the primary effect  $\nu = 134$  for the bivariate model trained with all samples from R-CHOP treated patients<sup>(III.2.2.1)</sup>. It predicts hierarchical survival dependencies<sup>(III.2.2.3)</sup> that exclusively exist in the lower risk partition of the primary effect<sup>(Figure III.2.2.3.a)</sup>. It can also significantly predict survival differences between GCB DLBCL patients, but not between ABC DLBCL patients. More precisely<sup>(cf. Figure III.2.2.3.b)</sup>, 89/327 R-CHOP treated GCB DLBCL patients in the high risk interval of  $\nu = 127$  have an average survival of only 55.1% while 218/327 patients in the baseline risk interval have an average survival of 78.4% ( $p = 5.6 \cdot 10^{-6}$ , log rank test). From the remaining 20/327 GCB DLBCL patients in the low risk interval of effect  $\nu = 127$  17/20 survived (average survival of 85%, still significant with  $p = 0.018$  relative to the high risk interval, despite the low sample number).

Effect  $\nu = 131$  qualified as best secondary variable on top of effect  $\nu = 134$  with  $p = 1.6 \cdot 10^{-9}$  for the quivariate predictor model trained with all samples<sup>(III.2.5.1)</sup>. Here,  $\nu = 127$  followed on rank #2 with  $p = 2.0 \cdot 10^{-9}$ . Due to the high correlation of their consensus gene scores ( $r_{(127;131)}^c = 0.80$ <sup>(cf. Eqn. III.1.3.2.b)</sup>) both effects share many top probesets (for one cohort, they are even based on the identical detected effect). Additionally, their sample eigensignal strengths<sup>(cf. Table III.1.5)</sup> from all four cohorts are correlated with 0.90, i.e. sample arrangements by either GEP effect are highly similar. Hence, it suffices to evaluate  $\nu = 127$  in detail here; results should be transferrable to  $\nu = 131$ .

To provide a direct validation of GEP effect  $\nu = 127$  on patient survival level (rather than as secondary variable within larger predictor models), I fitted additional univariate Cox models that are only based on effect  $\nu = 127$  to each of the four GCB DLBCL sub cohorts separately. Independent and highly significant validations succeeded based on both R-CHOP treated GCB DLBCL sub cohorts:  $p_{GSE10846.R-CHOP.GCB} = 1.0 \cdot 10^{-5}$  and  $p_{GSE31312.R-CHOP.GCB} = 5.4 \cdot 10^{-7}$ . However, these analyses revealed another striking difference, this time between R-CHOP and CHOP therapy, as for CHOP-treated GCB DLBCL samples  $p_{GSE10846.CHOP.GCB} = 0.44$  and  $p_{GSE4475.nonMBL.GCB} = 0.77$  only. The same difference (significant in R-CHOP.GCB, but not in CHOP.GCB) exists for  $\nu = 131$ .

## applied effect 127/135

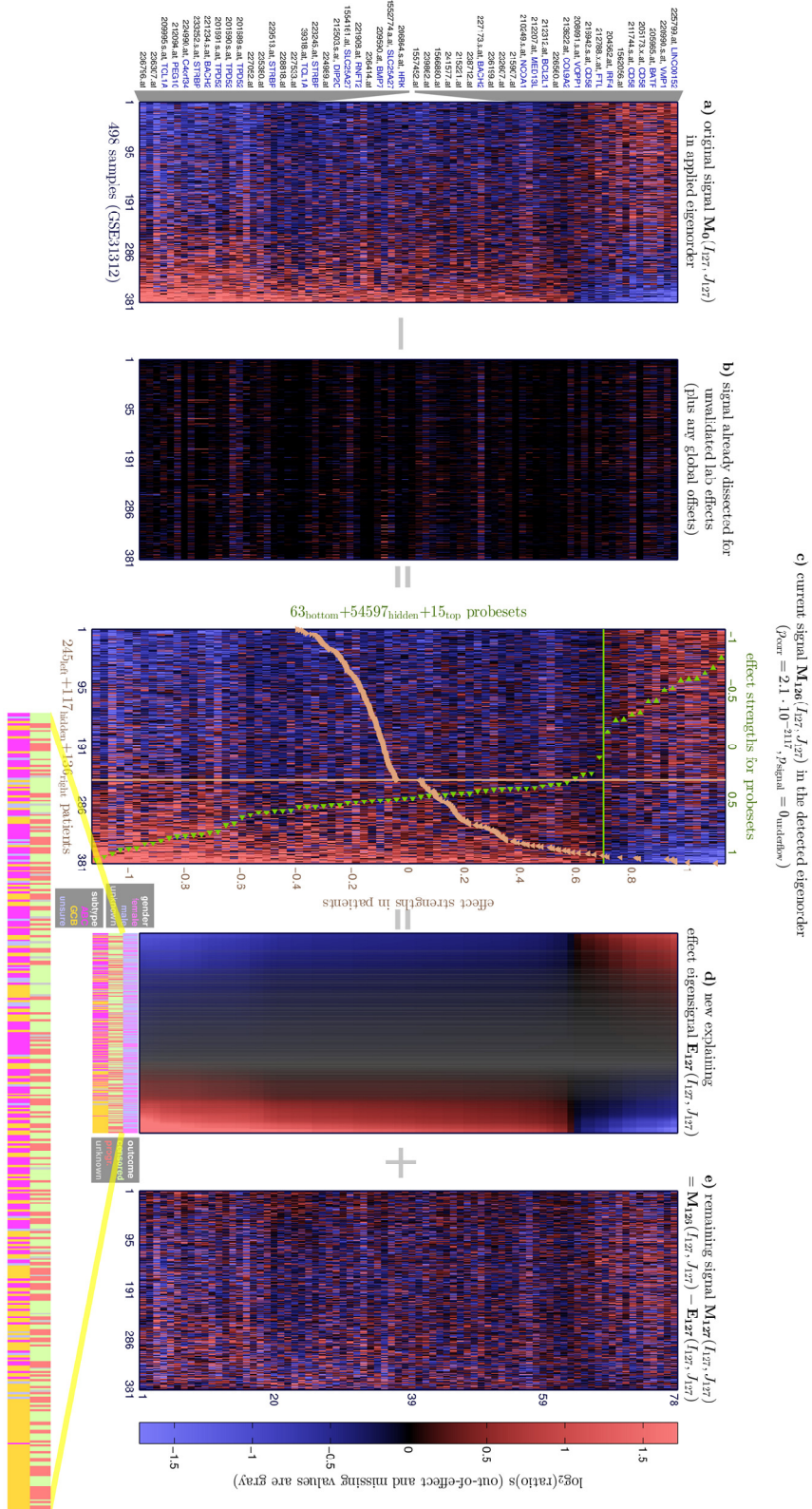


Figure III.3.3.2) Validated effect  $v=127$ , applied to GSE31312


Secondary survival effect by R-CHOP based survival analysis<sup>[III.2.2]</sup> applied to GSE31312.R-CHOP (498 patients); it is genetically highly correlated to effect  $v=131$ , the secondary survival effect for survival analyses based on all samples<sup>[III.2.5]</sup>. Enlarged binary classifications show censored/progression follow-up information (green/orange) as well as public subtypes (yellow/pink). At the right/lower tail, more GCB DLBCL patients with deaths or progressions were observed than could be expected by average GCB DLBCL survival.

(The genomic consensus effect is applied to the cleaned signal without lab effects<sup>[cf. III.1.4.2]</sup>. Samples and probesets are ordered by their effect strengths in this cohort<sup>[cf. Table III.1.5]</sup>. Additionally, probesets are filtered by demanding a relative correlation stronger than 0.5. The effect's bimonotonic eigensignal<sup>[panel d]</sup> is grayed for samples having insufficient or insignificant correlation to this effect<sup>[II.4.2.1]</sup>.)



When sorting all patients by  $v = 127$ , subtypes are separated (cf. Figure III.3.2.2.a) with 72% agreement ( $p = 3.2 \cdot 10^{-25}$ ) and for  $v = 131$  even with 83% agreement ( $p = 8.9 \cdot 10^{-73}$ ). This trend together with the adverse outcome of ABC DLBCL compared to GCB DLBCL *overlaps and hides the opposite survival trend within GCB DLBCL*. Hence,  $v = 127$  cannot predict a strong survival trend when applied to all DLBCL samples. This demonstrates that a multivariate survival analysis was necessary to reveal these dependencies and also shows that significant association of an effect to subtypes cannot provide a biologically complete characterization of underlying genes. With the background knowledge of these two opposite survival trends, they can now even be noticed visually: While the subtype trend is clearly visible, more GCB DLBCL patients had a progression or died in the upper/right interval than could be expected by average GCB DLBCL survival (Figure III.3.3.2).

### Effect overview

Effect  $v = 127$  was unsupervisedly detected at rank #41 in GSE10846.R-CHOP and at rank #11 in GSE31312. It has been supervisedly validated on GEP level in all four cohorts (see 127\* files at  C=Consensus Effects (cohort subfolders)).

Top probesets of effect  $v = 127$  are depicted (cf. Figure III.3.3.2) for a relative correlation  $\geq 0.5$ . It has a two-sided gradual gene expression eigensignal with approximately 65 correlated and 15 anti-correlated top probesets.

Higher expressions of correlated genes ( $r_{127}^g > 0$ ) are associated with adverse outcome. Consequently, lower expressions of anti-correlated genes ( $r_{127}^g < 0$ ) are also associated with adverse outcome. As summarized above, this association exists within the lower risk partition of effect  $v = 134$  and within GCB DLBCL, but neither in the high risk partition of  $v = 134$  nor in ABC DLBCL.

### Genomic associations

Gene ontology analysis did not reveal any significantly overexpressed and specific terms. On gene set level, 4/6 top enriched signatures are related to differential expression of Burkitt's lymphoma (BL) relative to DLBCL:

| Signatures            |   |           |            | GSEA             |        |      | Basic Statistics              |                          |                 |              |
|-----------------------|---|-----------|------------|------------------|--------|------|-------------------------------|--------------------------|-----------------|--------------|
| Signatures DB         | Signature Name  | # defined | # measured | Enrichment score | $p$    | FDR  | Mean log <sub>2</sub> (ratio) | $p$ (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| GeneSigDB_v4_Sept2011 | Lymphoma_Hummel06_24genes   | 14        | 14         | -0.950           | 0.0020 | 0.2% | 0.314                         | 0.0018                   | 8.7%            | 91.3%        |
| MolSigDBv4_0_dMay2014 | Hummel_Burkittslymphoma_dn  | 15        | 15         | -0.883           | 0.0021 | 0.2% | 0.316                         | 0.0008                   | 8.1%            | 91.9%        |
| StaudtSigDB_dNov2012  | ABC_gt_GCB_LC   | 15        | 15         | -0.755           | 0.0019 | 0.2% | 0.464                         | 8.3E-06                  | 5.4%            | 94.6%        |
| HGNCSigDB_dMay2014    | Protein tyrosine phosphatases / Class I Cys-based PTPs: MAP kinase phosphatases | 11        | 11         | -0.685           | 0.0019 | 0.2% | 0.180                         | 0.0502                   | 11.5%           | 88.5%        |
| MolSigDBv4_0_dMay2014 | Hummel_Burkittslymphoma_up  | 43        | 43         | 0.725            | 0.0021 | 0.2% | -0.311                        | 9.6E-06                  | 91.5%           | 8.5%         |
| StaudtSigDB_dNov2012  | GC_B_cell_BLhigh_DLBCLlow   | 36        | 35         | 0.734            | 0.0021 | 0.2% | -0.550                        | 2.0E-08                  | 98.4%           | 1.6%         |

Table III.3.3.2.a) Top-enriched signatures by  $v = 127$

Signatures with  $|\text{enrichment score}| \geq 0.67$  and at least 10 measured members are listed for genes ranked by GEP effect  $v = 127$ . All GSEA  $p$  values are based on permutation tests; hence, they are lower-bounded by  $1/(\#\text{permutations})$  and true  $p$  values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of  $t$ -tests against zero regulation as weights).

GCB DLBCL patients with adverse outcome have downregulated Hummel\_Burkittslymphoma\_dn (online interpretation card, [95]) and upregulated Hummel\_Burkittslymphoma\_up (online interpretation card, [95]). Superficially, this might suggest that these GCB DLBCL patients are misclassified BL rather than DLBCL cases. However, molecularly similarly determined BL patients showed a much higher 5-years survival of approximately 80% [95.figure 3] that makes this conclusion questionable. Furthermore, even LymphomaHummel06\_24genes (online interpretation card, [95]) with enrichment score  $-0.95$  does not contain any of the GEP effect's top genes (as listed in Table III.3.3.2.b), i.e. all signature genes are correlated with  $|r_{127}^g| \leq 0.4$ . As before, enrichment results are not specific enough and top correlated genes of this effect need to be analyzed individually to further elucidate its biological meaning.

### ■ Preliminary top genes analysis

For probesets with  $|r_{127}^g| \geq 0.4$ , 71 unique genes from 19 chromosomes participate in the effect, suggesting that it reflects a functional genomic network, rather than just reflecting local aberrations on DNA level. For  $|r_{127}^g| \geq 0.5$ , 14 unique genes from 14 chromosomes remain.

On the co-regulated side, several probesets with high correlations to the effect are available. The list is led by BACH2 ( $r_{127, \text{BACH2}}^g = 0.86$ , 6q15). Three unannotated probesets that measure transcribed sequences in direct genomic vicinity to BACH2 (with no other measured probeset in-between) are also highly correlated. Because their annotated sequence intervals are overlapped by BACH2-annotated probesets, they potentially still measure BACH2 RNA.

A recent study<sup>[113]</sup> performed parallel to this work has already tested immunohistochemical BACH2 expression for its prognostic value in DLBCL. (Such prognostic factors are important in order to optimize therapeutic strategies, as molecular analyses are not readily available in clinical practice.) Indeed, for a cohort size of 76 DLBCL patients, the BACH2-low group ( $n = 36$ ) showed an average overall survival of 91% that was significantly higher ( $p = 0.026$ ) than for the BACH2-high group ( $n = 40$ ) with 72% average overall survival. Progression free survival showed the same trend ( $p = 0.068$ ). However, there was no significant difference between GCB DLBCL and non-GCB DLBCL classes. Non-GCB DLBCL overall survival was even higher on average than GCB DLBCL overall survival. This fact and the small study size made it impossible to recognize that BACH2 belongs to a *hierarchical* survival effect.

Interestingly, the study also summarizes inconsistent findings from previous studies with respect to the survival impact of BACH2 expression in DLBCL. These inconsistencies are probably caused by and may be explained by this hierarchical survival impact of  $v = 127$ . (Its anti-aligned survival trend relative to the overlapping global subtype trend for the same genes might cause confusing and misleading results for univariate analyses.)

| Top genes in $v=127$ |             |            |                 |                 |                 |                   |
|----------------------|-------------|------------|-----------------|-----------------|-----------------|-------------------|
| Probeset             | Gene Symbol | HG19 locus | $a_{127}^{g,c}$ | $r_{127}^{g,c}$ | $P_{127}^{g,c}$ | $ r_{127}^{g,c} $ |
| 221234_s_at          | BACH2       | 6q15       | 1.46            | 0.86            | 1.3E-62         | 1.25              |
| 236796_at            |             | 6q15       | 1.40            | 0.86            | 2.5E-64         | 1.21              |
| 236307_at            |             | 6q15       | 1.36            | 0.82            | 1.6E-53         | 1.12              |
| 209995_s_at          | TCL1A       | 14q32.13   | 1.48            | 0.51            | 9.5E-16         | 0.76              |
| 212094_at            | PEG10       | 7q21.3     | 1.46            | 0.51            | 3.6E-15         | 0.74              |
| 39318_at             | TCL1A       | 14q32.13   | 1.44            | 0.51            | 3.4E-15         | 0.73              |
| 224990_at            | C4orf34     | 4p14       | 1.14            | 0.57            | 6.6E-20         | 0.65              |
| 229513_at            | STRBP       | 9q33.3     | 0.93            | 0.68            | 9.8E-31         | 0.64              |
| 1556451_at           |             | 6q15       | 0.80            | 0.75            | 2.3E-40         | 0.61              |
| 201691_s_at          | TPD52       | 8q21.13    | 1.01            | 0.59            | 1.6E-21         | 0.60              |
| 227052_at            |             | 4p14       | 1.04            | 0.55            | 3.3E-18         | 0.57              |
| 227173_s_at          | BACH2       | 6q15       | 0.77            | 0.73            | 2.8E-36         | 0.56              |
| 235380_at            |             | 10q11.21   | 0.92            | 0.53            | 1.8E-16         | 0.48              |
| 238919_at            |             | 13q21.32   | 1.13            | 0.41            | 3.5E-10         | 0.47              |
| 212092_at            | PEG10       | 7q21.3     | 0.93            | 0.50            | 1.2E-14         | 0.46              |
| 224989_at            |             | 4p14       | 0.94            | 0.49            | 5.7E-14         | 0.46              |
| 206864_s_at          | HRK         | 12q24.22   | 0.95            | 0.47            | 2.8E-13         | 0.45              |
| 223245_at            | STRBP       | 9q33.3     | 0.75            | 0.60            | 2.1E-22         | 0.45              |
| 233252_s_at          | STRBP       | 9q33.3     | 0.73            | 0.58            | 2.0E-20         | 0.42              |
| 1566734_at           | LOC283454   | 12q24.22   | 0.99            | 0.42            | 1.9E-10         | 0.42              |
| 221908_at            | RNF2        | 12q24.22   | 0.78            | 0.53            | 7.3E-17         | 0.41              |
| 212503_s_at          | DIP2C       | 10p15.3    | 0.91            | 0.44            | 1.5E-11         | 0.40              |
| 1554161_at           | SLC25A27    | 6p12.3     | 0.76            | 0.53            | 1.6E-16         | 0.40              |
| 206896_s_at          | GNG7        | 19p13.3    | 0.81            | 0.49            | 3.3E-14         | 0.40              |
| 201690_s_at          | TPD52       | 8q21.13    | 0.74            | 0.53            | 2.5E-17         | 0.39              |
| 208651_x_at          | CD24        | Yq11.222   | 0.90            | 0.43            | 5.9E-11         | 0.39              |
| 228818_at            |             | 8q21.13    | 0.75            | 0.52            | 6.2E-16         | 0.39              |
| 266_s_at             | CD24        | Yq11.222   | 0.94            | 0.40            | 9.3E-10         | 0.38              |
| 216379_x_at          | CD24        | 6q21       | 0.89            | 0.41            | 5.5E-10         | 0.36              |
| 236414_at            |             | 8q21.13    | 0.61            | 0.57            | 6.1E-20         | 0.35              |
| 201689_s_at          | TPD52       | 8q21.13    | 0.69            | 0.50            | 7.1E-15         | 0.35              |
| 209771_x_at          | CD24        | Yq11.222   | 0.85            | 0.40            | 1.3E-09         | 0.34              |
| 227798_at            | SMAD1       | 4q31.21    | 0.85            | 0.40            | 1.4E-09         | 0.34              |
| 209590_at            | BMP7        | 20q13.31   | 0.73            | 0.45            | 2.9E-12         | 0.33              |
| 227407_at            | TAPT1       | 4p15.32    | 0.65            | 0.50            | 3.9E-15         | 0.33              |
| 238712_at            |             | 3p13       | 0.65            | 0.49            | 2.1E-14         | 0.32              |
| 203434_s_at          | MME         | 3q25.2     | 0.76            | 0.40            | 1.2E-09         | 0.31              |
| 204165_at            | WASF1       | 6q21       | 0.66            | 0.46            | 1.1E-12         | 0.30              |
| 223246_s_at          | STRBP       | 9q33.3     | 0.57            | 0.53            | 5.6E-17         | 0.30              |
| 232286_at            |             | 2q11.2     | 0.74            | 0.41            | 7.6E-10         | 0.30              |
| 222336_at            | C4orf34     | 4p14       | 0.63            | 0.48            | 1.8E-13         | 0.30              |
| 227533_at            |             | 1q25.2     | 0.59            | 0.51            | 2.3E-15         | 0.30              |
| 225978_at            | RIMKLB      | 12p13.31   | 0.70            | 0.42            | 1.1E-10         | 0.30              |
| 223522_at            | MIR600HG    | 9q33.3     | 0.52            | 0.57            | 1.0E-19         | 0.30              |
| 1552774_a_at         | SLC25A27    | 6p12.3     | 0.54            | 0.54            | 2.5E-17         | 0.29              |
| 1566880_at           |             | 2q11.2     | 0.59            | 0.49            | 4.2E-14         | 0.29              |
| 226164_x_at          | RIMKLB      | 12p13.31   | 0.58            | 0.49            | 3.4E-14         | 0.29              |
| 229670_at            |             | 1q24.2     | 0.63            | 0.45            | 9.7E-12         | 0.28              |
| 241577_at            |             | 2q11.2     | 0.63            | 0.44            | 1.9E-11         | 0.28              |
| 239884_at            | CADPS       | 3p14.2     | 0.64            | 0.43            | 4.6E-11         | 0.28              |
| 225999_at            | RIMKLB      | 12p13.31   | 0.61            | 0.45            | 7.4E-12         | 0.27              |
| 225421_at            | PM20D2      | 6q15       | 0.55            | 0.49            | 2.9E-14         | 0.27              |
| 1557814_a_at         |             | 5q14.1     | 0.51            | 0.50            | 4.1E-15         | 0.26              |
| 236655_at            | TPD52       | 8q21.13    | 0.52            | 0.50            | 1.0E-14         | 0.26              |
| 239862_at            |             | 8q21.13    | 0.52            | 0.48            | 1.2E-13         | 0.25              |
| 242090_x_at          |             | NA         | 0.57            | 0.43            | 6.3E-11         | 0.25              |
| 229552_at            | LOC283454   | 12q24.22   | 0.58            | 0.42            | 1.9E-10         | 0.24              |
| 208754_s_at          | NAP1L1      | 12q21.2    | 0.50            | 0.48            | 1.0E-13         | 0.24              |
| 242681_at            |             | 1p36.22    | 0.56            | 0.43            | 5.9E-11         | 0.24              |
| 231817_at            | USP53       | 4q26       | 0.57            | 0.42            | 2.8E-10         | 0.24              |
| 233251_at            | STRBP       | 9q33.3     | 0.46            | 0.51            | 2.5E-15         | 0.23              |
| 202478_at            | TRIB2       | 2p24.3     | 0.53            | 0.43            | 3.8E-11         | 0.23              |
| 229344_x_at          | RIMKLB      | 12p13.31   | 0.48            | 0.47            | 3.3E-13         | 0.23              |
| 215221_at            |             | 3p13       | 0.44            | 0.51            | 1.9E-15         | 0.23              |
| 225763_at            | RCS1        | 1q24.2     | 0.54            | 0.40            | 1.0E-09         | 0.22              |
| 219655_at            | C7orf10     | 7p14.1     | 0.51            | 0.43            | 9.4E-11         | 0.22              |
| 235310_at            | GCE2        | 3q13.2     | 0.53            | 0.41            | 6.7E-10         | 0.22              |
| 218988_at            | SLC35E3     | 12q15      | 0.52            | 0.41            | 3.9E-10         | 0.21              |
| 228991_at            | CDK13       | 7p14.1     | 0.44            | 0.48            | 1.2E-13         | 0.21              |
| 228081_at            | CCNG2       | 4q21.1     | 0.46            | 0.45            | 3.5E-12         | 0.21              |
| 227369_at            | SERBP1      | 1p31.3     | 0.51            | 0.41            | 5.7E-10         | 0.21              |
| 238484_s_at          | SSBP2       | 5q14.1     | 0.47            | 0.44            | 2.6E-11         | 0.21              |
| 241933_at            | QRS1        | 6q21       | 0.45            | 0.44            | 1.0E-11         | 0.20              |
| 201688_s_at          | TPD52       | 8q21.13    | 0.50            | 0.40            | 9.5E-10         | 0.20              |
| 238483_at            | SSBP2       | 5q14.1     | 0.46            | 0.43            | 5.2E-11         | 0.20              |

The second gene in terms of correlation is STRBP ( $r_{127,STRBP}^g = 0.68, 9q33.3$ ). Regarding STRBP, no B cell or lymphomagenesis related literature could be found.

Other co-regulated probesets for TCL1A ( $r_{127,TCL1A}^g = 0.51, 14q32.13$ ), PEG10 ( $r_{127,PEG10}^g = 0.51, 7q21.3$ ) and C4orf34 ( $r_{127,C4orf34}^g = 0.57, 4p14$ ) show strong differential expressions, but not as high correlation to the effect.

The anti-regulated side is led by LINC00152 ( $r_{127,LINC00152}^g = -0.58, 2p11.2$ ); this is a long intergenic *non-protein-coding* RNA. No direct link to DLBCL could be located, but LINC00152 was found to be significantly expressed in gastric cancer<sup>[114]</sup> compared to normal adjacent tissue (fold change = 1.93,  $p = 6.9 \cdot 10^{-9}$ ). It has subsequently been suggested as biomarker<sup>[115]</sup> for gastric cancer, because its expression levels were significantly increased compared to mucosa from healthy controls ( $p = 0.004$ ) as well as when comparing gastric juice between gastric cancer patients and normal controls ( $p = 0.002$ ).

Interestingly, another study<sup>[116]</sup> showed that LINC00152 is being significantly and strongly upregulated in HeLa cells in response to chemical stressors, especially by Cisplatin (fold change of 32). Hence, it was suggested as surrogate indicator of general or specific cell stress.

Other anti-regulated genes include BATF ( $r_{127,BATF}^g = -0.53, 14q24.3$ ) and IRF4 ( $r_{127,IRF4}^g = -0.49, 6p25.3$ ). These genes are related to T cell lymphomas and *inhibiting* IRF4 (and MYC) caused toxicity in ALCL cell lines<sup>[76]</sup>. While these properties might be transferrable to ABC DLBCL cells (given their relative overexpression of IRF4 and BATF), for the GCB DLBCL subgroup in question *low* expressions of these genes are associated with significantly adverse patient outcome.

### Inference

To elucidate the biological pathway underlying  $\nu = 127$  and the cause of its hierarchical survival impact relative to effect  $\nu = 134$ , further biological experiments might potentially be helpful. Analytically, BACH2 seems to be a promising oncogene candidate for the lower risk partition of  $\nu = 134$ , i.e. predominantly for GCB DLBCL patients. To test this hypothesis, GCB DLBCL cell lines with high BACH2 protein expression could be selected for BACH2 knockdown experiments.

Additionally, it might be biologically illuminating to investigate the cause why the survival impact of  $\nu = 127$  is highly significant for GCB DLBCL patients following R-CHOP therapy, but not for the CHOP therapy. The 5-years survival in the higher risk partition of  $\nu = 127$  for R-CHOP-treated GCB DLBCL patients equals approximately 55%. This is comparable to GCB DLBCL 5-years survival following CHOP therapy. Hence, one conceivable hypothesis pending further validation might be that GCB DLBCL patients in this higher risk partition of  $\nu = 127$  cannot profit from Rituximab.

|             |           |              |       |       |         |       |
|-------------|-----------|--------------|-------|-------|---------|-------|
| 236199_at   |           | 10q11.21     | 0.43  | 0.46  | 2.5E-12 | 0.20  |
| 218949_s_at | QRSL1     | 6q21         | 0.48  | 0.40  | 1.0E-09 | 0.20  |
| 244185_at   |           | 12q22        | 0.44  | 0.44  | 2.7E-11 | 0.19  |
| 1557452_at  |           | 5q14.1       | 0.42  | 0.44  | 1.8E-11 | 0.19  |
| 223624_at   | ZFAND4    | 10q11.22     | 0.44  | 0.42  | 2.6E-10 | 0.18  |
| 230624_at   | SLC25A27  | 6p12.3       | 0.45  | 0.40  | 1.1E-09 | 0.18  |
| 237187_at   |           | 12q24.22     | 0.44  | 0.41  | 8.6E-10 | 0.18  |
| 214042_s_at | RPL22     | 3q26.2       | 0.43  | 0.42  | 2.2E-10 | 0.18  |
| 201678_s_at | C3orf37   | 3q21.3       | 0.44  | 0.40  | 1.3E-09 | 0.18  |
| 240176_at   |           | 2p11.2, 2q13 | -0.41 | -0.42 | 2.1E-10 | -0.17 |
| 221658_s_at | IL21R     | 16p12.1      | -0.43 | -0.41 | 4.1E-10 | -0.18 |
| 213622_at   | COL9A2    | 1p34.2       | -0.42 | -0.43 | 3.6E-11 | -0.18 |
| 244612_at   | NA        |              | -0.45 | -0.41 | 3.6E-10 | -0.19 |
| 202644_s_at | TNFAIP3   | 6q23.3       | -0.48 | -0.40 | 8.9E-10 | -0.19 |
| 235668_at   | PRDM1     | 6q21         | -0.46 | -0.43 | 8.0E-11 | -0.20 |
| 219424_at   | EBI3      | 19p13.3      | -0.47 | -0.42 | 1.3E-10 | -0.20 |
| 238567_at   | SGPP2     | 2q36.1       | -0.53 | -0.43 | 9.0E-11 | -0.23 |
| 216942_s_at | CD58      | 1p13.1       | -0.58 | -0.44 | 2.4E-11 | -0.25 |
| 1562056_at  |           | 2p11.2, 2q13 | -0.54 | -0.51 | 1.4E-15 | -0.28 |
| 220990_s_at | VMP1      | 17q23.1      | -0.57 | -0.50 | 8.1E-15 | -0.28 |
| 226560_at   |           | 2q36.1       | -0.64 | -0.47 | 3.7E-13 | -0.30 |
| 211744_s_at | CD58      | 1p13.1       | -0.71 | -0.47 | 4.2E-13 | -0.33 |
| 205173_x_at | CD58      | 1p13.1       | -0.72 | -0.46 | 1.1E-12 | -0.33 |
| 205965_at   | BATF      | 14q24.3      | -0.63 | -0.53 | 4.2E-17 | -0.33 |
| 204562_at   | IRF4      | 6p25.3       | -0.71 | -0.49 | 4.4E-14 | -0.35 |
| 225799_at   | LINC00152 | 2p11.2       | -0.69 | -0.58 | 1.3E-20 | -0.40 |

Table III.3.3.2.b) Top genes in validated effect  $\nu = 127$

|                 |  |
|-----------------|--|
| (probesets)     | from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx <sup>TM</sup> (97) v33)                   |
| $a_{127}^{g,c}$ | Components of the consensus gene axis of effect $\nu = 127$ (cf. Table III.1.5); filtered $ a_{127}^{g,c}  \geq 0.4$ . |
| $r_{127}^{g,c}$ | Consensus gene correlations; filtered $ r_{127}^{g,c}  \geq 0.4$ .   |
| $p_{127}^{g,c}$ | $p$ values for the correlations (cf. II.5.2.1)   |



### Inference

Due to its high genomic specificity and the lack of strongly correlated genes from other chromosomes, this GEP effect may reflect e.g. a DNA level aberration; copy number measurements<sup>(e.g. from GSE11318, [11])</sup> for its locus could clarify this.

Due to its prognostic value on top of the COO induced effect and its anti-aligned survival effect, analyzing *KIAA1217* experimentally might also be interesting. To this end, DLBCL cell lines could first be screened for low *KIAA1217* protein levels relative to DLBCL average. Ideally, proliferation of a subset of these cell lines could be reduced by *KIAA1217* overexpression experiments.

However, this effect cannot explain any strong survival dependencies on top of the primary survival effect  $\nu = 134$ ; it only ranks #37 with  $p = 0.006$  (cf. Figure III.2.5.1.b). Therefore, experiments for this gene seem to be of second priority from an analytical point of view.

| Top genes in $\nu = 105$ |             |            |                 |                 |                 |  |
|--------------------------|-------------|------------|-----------------|-----------------|-----------------|--|
| Probeset                 | Gene Symbol | HG19 locus | $a_{105}^{g,c}$ | $r_{105}^{g,c}$ | $p_{105}^{g,c}$ | $\left  \frac{a_{105}^{g,c}}{r_{105}^{g,c}} \right $ |
| 231807_at                | KIAA1217    | 10p12.1    | 1.26            | 0.92            | 5.9E-68         | 1.16   |
| 1554438_at               | KIAA1217    | 10p12.1    | 1.01            | 0.87            | 5.7E-51         | 0.88   |
| 232762_at                | KIAA1217    | 10p12.1    | 0.79            | 0.86            | 3.7E-49         | 0.69   |
| 244147_at                |             | 10p12.1    | 0.58            | 0.78            | 6.8E-34         | 0.45   |
| 214912_at                |             | 10p12.2    | 0.60            | 0.74            | 1.9E-29         | 0.45   |
| 242846_at                |             | 10p12.2    | 0.45            | 0.69            | 8.4E-24         | 0.31   |
| 1562966_at               | KIAA1217    | 10p12.1    | 0.50            | 0.61            | 1.7E-17         | 0.30   |
| 235333_at                | B4GALT6     | 18q12.1    | 0.51            | 0.37            | 1.1E-06         | 0.19   |
| 203562_at                | FEZ1        | 11q24.2    | 0.59            | 0.31            | 5.8E-05         | 0.18   |
| 241163_at                |             | 3q26.31    | 0.50            | 0.31            | 5.7E-05         | 0.16   |
| 229070_at                | ADTRP       | 6p24.1     | 0.48            | 0.31            | 6.1E-05         | 0.15   |
| 224374_s_at              | EMILIN2     | 18p11.32   | 0.38            | 0.35            | 5.2E-06         | 0.13   |
| 225202_at                | RHOBTB3     | 5q15       | 0.40            | 0.32            | 3.8E-05         | 0.13   |
| 232352_at                | ISL2        | 15q24.3    | 0.37            | 0.34            | 8.4E-06         | 0.13   |
| 226099_at                | ELL2        | 5q15       | 0.37            | 0.32            | 3.0E-05         | 0.12   |
| 225662_at                | ZAK         | 2q31.1     | 0.36            | 0.32            | 3.9E-05         | 0.12   |
| 204083_s_at              | TPM2        | 9p13.3     | 0.37            | 0.30            | 1.1E-04         | 0.11   |
| 206034_at                | SERPINB8    | 18q22.1    | 0.31            | 0.35            | 4.1E-06         | 0.11   |
| 202950_at                | CRYZ        | 1p31.1     | 0.36            | 0.30            | 1.0E-04         | 0.11   |
| 206490_at                | DLGAP1      | 18p11.31   | 0.34            | 0.30            | 9.8E-05         | 0.10   |
| 233002_at                | PPP4R4      | 14q32.12   | 0.33            | 0.30            | 1.0E-04         | 0.10   |
| 213060_s_at              | CHI3L2      | 1p13.3     | -0.42           | -0.31           | 7.1E-05         | -0.13  |

Table III.3.3.a) Top genes in validated effect  $\nu = 105$

|                 |  |
|-----------------|--|
| (probesets)     | from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx™ <sup>[97]</sup> v33)                     |
| $a_{105}^{g,c}$ | Components of the consensus gene axis of effect $\nu = 105$ (cf. Table III.1.5); filtered $ a_{105}^{g,c}  \geq 0.3$ . |
| $r_{105}^{g,c}$ | Consensus gene correlations; filtered $ r_{105}^{g,c}  \geq 0.3$ .   |
| $p_{105}^{g,c}$ | $p$ values for the correlations (cf. II.5.2.1)   |

### III.3.3.4 $\nu = 5$ : A stromal effect (CHOP based model, 1<sup>st</sup> predictor variable)

#### Role in survival analysis

Sorting patients by the eigensignal of effect  $\nu = 5$  shows the most consistent survival prediction ( $p = 2.2 \cdot 10^{-8}$ ) of all discovered GEP effects for *CHOP-treated* patients (Figure III.2.4.1.a). But for *R-CHOP treated* patients it only ranks #14<sup>th</sup> ( $p = 8.4 \cdot 10^{-4}$ ).

#### Effect overview

The effect has been unsupervisedly quad-discovered effect with detection ranks #1 (GSE10846.CHOP), #1 (GSE10846.R-CHOP), #1 (GSE4475.nonMBL) and #2 (GSE31312.R-CHOP). It is one of the largest discovered GEP effects in terms of number of correlated and differentially expressed genes. It is *the* largest effect that is associated with a significant survival impact.

Higher expressions of this effect correspond to lower risk and more favorable outcome (cf. Table III.2.4.2).

The effect has a one-sided eigensignal (Figure III.3.3.4.d, page 199) that has more than 650 probesets correlated stronger than  $|r_5^g| \geq 0.5$ . To unveil its biological meaning, I focus this evaluation on its top genes only (Table III.3.3.4.a).

| Top genes in $\nu = 5$ |             |            |             |             |             |  |
|------------------------|-------------|------------|-------------|-------------|-------------|--|
| Probeset               | Gene Symbol | HG19 locus | $a_5^{g,c}$ | $r_5^{g,c}$ | $p_5^{g,c}$ | $\left  \frac{a_5^{g,c}}{r_5^{g,c}} \right $ |
| 202404_s_at            | COL1A2      | 7q21.3     | 1.88        | 0.96        | 7.5E-125    | 1.81   |
| 201852_x_at            | COL3A1      | 2q32.2     | 1.87        | 0.96        | 6.5E-120    | 1.78   |
| 1555778_a_at           | POSTN       | 13q13.3    | 2.02        | 0.88        | 4.2E-84     | 1.78   |
| 211161_s_at            | COL3A1      | NA         | 1.80        | 0.95        | 7.0E-117    | 1.72   |
| 202310_s_at            | COL1A1      | 17q21.33   | 1.80        | 0.94        | 7.9E-104    | 1.69   |
| 202311_s_at            | COL1A1      | 17q21.33   | 1.85        | 0.91        | 4.3E-88     | 1.69   |
| 215076_s_at            | COL3A1      | 2q32.2     | 1.72        | 0.95        | 4.9E-114    | 1.63   |
| 210809_s_at            | POSTN       | 13q13.3    | 1.85        | 0.87        | 1.4E-71     | 1.61   |
| 212464_s_at            | FN1         | 2q35       | 1.66        | 0.93        | 3.1E-96     | 1.53   |
| 227140_at              | INHBA       | 7p14.1     | 1.72        | 0.87        | 5.9E-81     | 1.50   |
| 221730_at              | COL5A2      | 2q32.2     | 1.58        | 0.93        | 1.2E-101    | 1.48   |
| 202403_s_at            | COL1A2      | 7q21.3     | 1.55        | 0.95        | 9.3E-118    | 1.48   |
| 221729_at              | COL5A2      | 2q32.2     | 1.56        | 0.94        | 1.8E-105    | 1.47   |
| 212489_at              | COL5A1      | 9q34.3     | 1.58        | 0.91        | 2.6E-88     | 1.44   |
| 211719_x_at            | FN1         | 2q35       | 1.54        | 0.93        | 4.1E-97     | 1.43   |
| 212354_at              | SULF1       | 8q13.2     | 1.54        | 0.90        | 8.4E-85     | 1.39   |
| 216442_x_at            | FN1         | 2q35       | 1.50        | 0.93        | 4.6E-99     | 1.39   |
| 210495_x_at            | FN1         | 2q35       | 1.49        | 0.93        | 8.5E-97     | 1.38   |
| 225664_at              | COL12A1     | 6q13       | 1.63        | 0.84        | 1.8E-69     | 1.37   |
| 225681_at              | CTHRC1      | 8q22.3     | 1.56        | 0.86        | 2.3E-76     | 1.35   |
| 209335_at              | DCN         | 12q21.33   | 1.53        | 0.87        | 3.8E-70     | 1.33   |
| 221731_x_at            | VCAN        | 5q14.3     | 1.47        | 0.90        | 2.0E-83     | 1.33   |
| 203083_at              | THBS2       | 6q27       | 1.43        | 0.92        | 7.3E-92     | 1.32   |
| 201744_s_at            | LUM         | 12q21.33   | 1.47        | 0.89        | 4.1E-77     | 1.31   |
| 211896_s_at            | DCN         | 12q21.33   | 1.49        | 0.87        | 2.8E-70     | 1.29   |
| 212353_at              | SULF1       | 8q13.2     | 1.51        | 0.85        | 1.6E-64     | 1.29   |
| 209596_at              | MXRA5       | Xp22.33    | 1.45        | 0.88        | 1.4E-73     | 1.28   |
| 204620_s_at            | VCAN        | 5q14.2     | 1.38        | 0.89        | 4.7E-79     | 1.23   |
| 202620_s_at            | PLOD2       | 3q24       | 1.39        | 0.87        | 3.0E-72     | 1.22   |
| 201893_x_at            | DCN         | 12q21.33   | 1.36        | 0.89        | 4.0E-77     | 1.21   |
| 201069_at              | MMP2        | 16q12.2    | 1.40        | 0.86        | 3.6E-68     | 1.20   |
| 215646_s_at            | VCAN        | 5q14.3     | 1.39        | 0.85        | 1.0E-63     | 1.18   |
| 212488_at              | COL5A1      | 9q34.3     | 1.33        | 0.89        | 2.7E-76     | 1.18   |
| 211813_x_at            | DCN         | 12q21.33   | 1.33        | 0.88        | 1.8E-72     | 1.17   |
| 203325_s_at            | COL5A1      | 9q34.3     | 1.29        | 0.88        | 2.8E-75     | 1.14   |

### Identification by gene ontology

For so far presented survival effects  $v \in \{129, 134, 127, 131, 105\}$  no highly significant gene ontology terms existed. For  $v = 5$  however, several GO terms are significantly overrepresented for cellular components (cf. Figure III.3.3.4.a), molecular functions and biological processes (cf. Figure III.3.3.4.b). The top 30-40 genes of this effect already suffice for significance. Based on these results,  $v = 5$  can be interpreted as a measurement of extracellular matrix structural constituents (cf. GO:0005201) with  $p < 2.2 \cdot 10^{-16}$  and of related terms for the extracellular matrix.

Small GO terms might serve as starting points for a biologically more specific interpretation. For example, the molecular function of platelet-derived growth factor binding (cf. GO:0048407) is significant (5 of the top 34 genes belong to this term, while the term contains only 11 genes in total from all measured 20370 genes; this results in  $p = 4.4 \cdot 10^{-12}$  via hypergeometric test).

|              |         |          |      |      |         |      |
|--------------|---------|----------|------|------|---------|------|
| 201438_at    | COL6A3  | 2q37.3   | 1.23 | 0.92 | 1.6E-90 | 1.13 |
| 1556499_s_at | COL1A1  | 17q21.33 | 1.27 | 0.88 | 7.8E-84 | 1.12 |
| 224694_at    | ANTXR1  | 2p13.3   | 1.32 | 0.83 | 1.2E-65 | 1.09 |
| 204619_s_at  | VCAN    | 5q14.2   | 1.22 | 0.86 | 8.3E-69 | 1.06 |
| 226777_at    | ADAM12  | 10q26.2  | 1.30 | 0.80 | 6.9E-59 | 1.04 |
| 229218_at    | COL1A2  | 7q21.3   | 1.18 | 0.86 | 9.7E-76 | 1.02 |
| 203131_at    | PDGFRA  | 4q12     | 1.21 | 0.83 | 5.1E-58 | 1.00 |
| 213905_x_at  | BGN     | Xq28     | 1.13 | 0.87 | 5.3E-71 | 0.99 |
| 203477_at    | COL15A1 | 9q22.33  | 1.18 | 0.81 | 2.3E-53 | 0.95 |
| 227399_at    | VGLL3   | 3p12.1   | 1.19 | 0.80 | 8.8E-59 | 0.95 |
| 207173_x_at  | CDH11   | 16q21    | 1.12 | 0.83 | 1.1E-57 | 0.93 |
| 202766_s_at  | FBN1    | 15q21.1  | 1.08 | 0.85 | 4.9E-64 | 0.92 |
| 202237_at    | NNMT    | 11q23.2  | 1.09 | 0.84 | 3.9E-60 | 0.91 |
| 212344_at    | SULF1   | 8q13.2   | 1.05 | 0.85 | 7.4E-64 | 0.90 |
| 225242_s_at  | CCDC80  | 3q13.2   | 1.10 | 0.81 | 7.6E-61 | 0.89 |
| 201261_x_at  | BGN     | Xq28     | 1.03 | 0.86 | 3.9E-67 | 0.89 |
| 211571_s_at  | VCAN    | 5q14.2   | 1.06 | 0.83 | 8.0E-59 | 0.88 |
| 201505_at    | LAMB1   | 7q31.1   | 1.08 | 0.81 | 4.9E-54 | 0.88 |
| 232458_at    | COL3A1  | 2q32.2   | 1.07 | 0.81 | 2.2E-60 | 0.87 |
| 210986_s_at  | TPM1    | 15q22.2  | 1.07 | 0.81 | 5.8E-53 | 0.86 |
| 209955_s_at  | FAP     | 2q24.2   | 0.99 | 0.85 | 3.1E-65 | 0.85 |
| 212667_at    | SPARC   | 5q33.1   | 1.01 | 0.82 | 2.2E-56 | 0.83 |
| 208782_at    | FSTL1   | 3q13.33  | 0.94 | 0.86 | 1.4E-68 | 0.81 |
| 202202_s_at  | LAMA4   | 6q21     | 0.95 | 0.84 | 2.3E-62 | 0.80 |
| 227628_at    | GPX8    | 5q11.2   | 0.95 | 0.82 | 7.4E-64 | 0.78 |
| 228141_at    | GPX8    | 5q11.2   | 0.95 | 0.82 | 2.0E-62 | 0.77 |
| 200665_s_at  | SPARC   | 5q33.1   | 0.92 | 0.83 | 9.8E-59 | 0.77 |
| 204517_at    | PPIC    | 5q23.2   | 0.93 | 0.81 | 1.6E-53 | 0.75 |
| 202619_s_at  | PLOD2   | 3q24     | 0.91 | 0.82 | 6.4E-55 | 0.74 |
| 210139_s_at  | PMP22   | 17p12    | 0.87 | 0.83 | 1.2E-58 | 0.73 |
| 202351_at    | ITGAV   | 2q32.1   | 0.87 | 0.81 | 8.4E-55 | 0.71 |
| 211651_s_at  | LAMB1   | 7q31.1   | 0.81 | 0.80 | 1.0E-51 | 0.65 |

Table III.3.3.4.a) Top genes in validated effect  $v=5$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx™<sup>[97]</sup> v33)  
 $a_s^{g,c}$  Components of the consensus gene axis of effect  $v=5$  (cf. Table III.1.5); filtered  $|a_s^{g,c}| \geq 0.8$ .  
 $r_s^{g,c}$  Consensus gene correlations; filtered  $|r_s^{g,c}| \geq 0.8$ .  
 $p_s^{g,c}$   $p$  values for the correlations (cf. II.5.2.1)

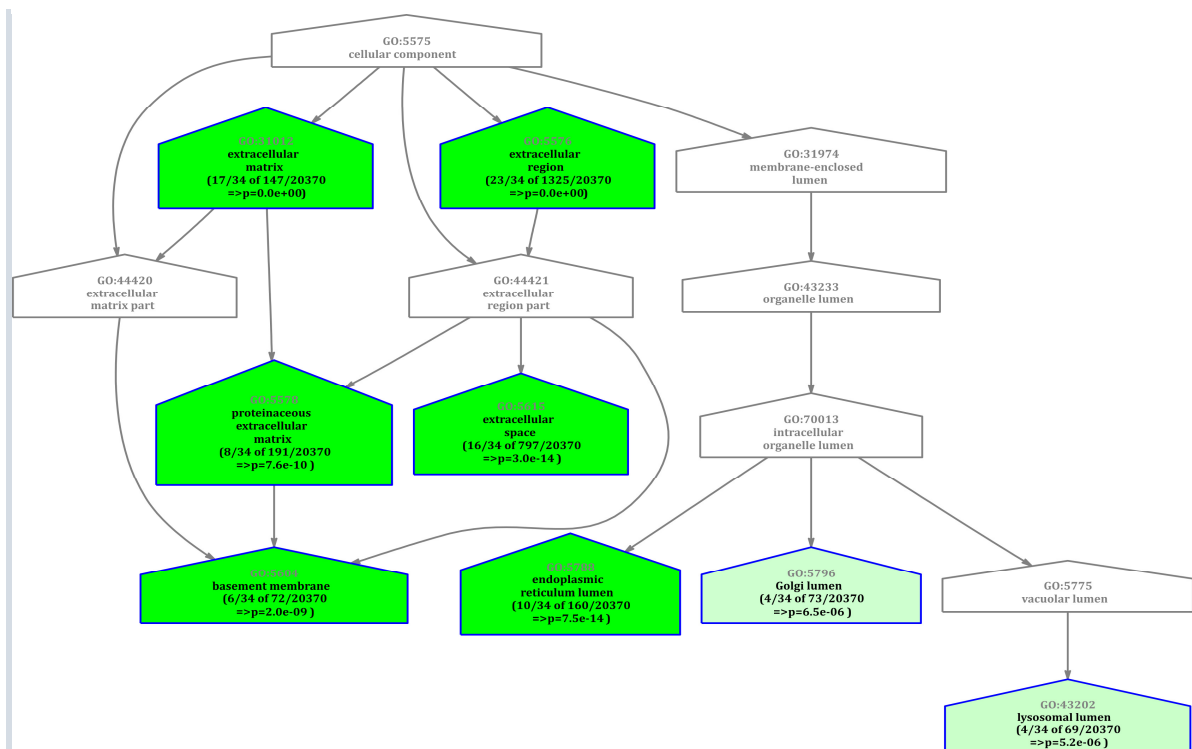


Figure III.3.3.4.a) Gene ontology overrepresentation analyses of cellular components in effect  $v=5$

All  $p$  values are based on hypergeometric tests. A  $p$  value of zero indicates an underflow, i.e.  $p < 2.2 \cdot 10^{-16}$ , which is the numeric resolution limit for differences to one; the true  $p$  value is never exactly zero.

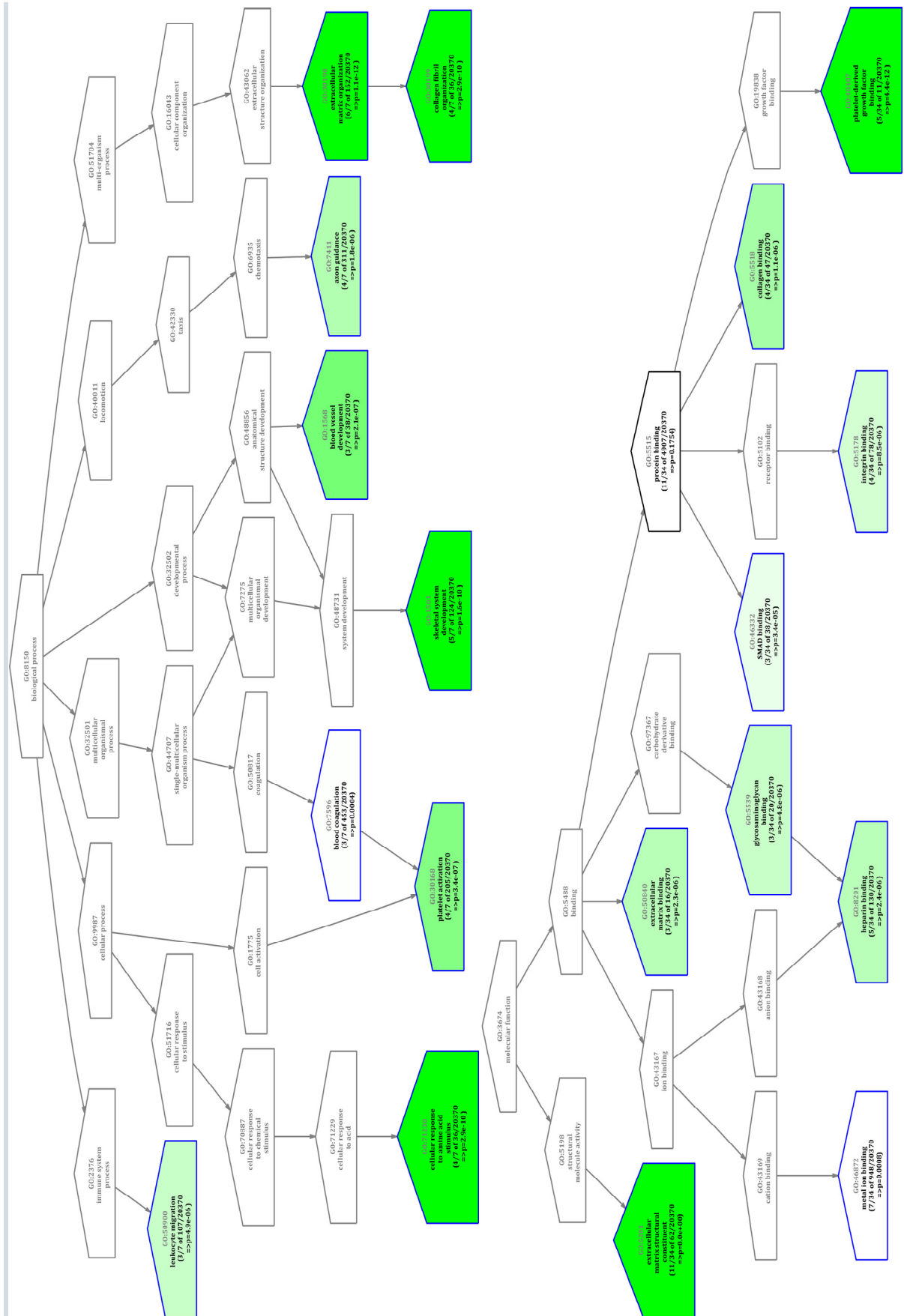


Figure III.3.3.4.b) Gene ontology overrepresentation analyses of biological processes and molecular functions in effect  $v=5$

All  $p$  values are based on hypergeometric tests. A  $p$  value of zero indicates an underflow, i.e.  $p < 2.2 \cdot 10^{-16}$ , which is the numeric resolution limit for differences to one; the true  $p$  value is never exactly zero.

Another example is the biological process of leukocyte migration<sup>(cf. GO:0050900)</sup> (with COL1A1, COL1A2 and FN1, three of the top 7 genes belong to this term, while 107 genes from all measured 20370 genes are directly annotated with this term; this results in with  $p = 4.9 \cdot 10^{-6}$  via hypergeometric test). (All GO results including associated gene IDs are available in graphical and tabular form in [D=Interpretation\genomic\v005\GOA.](#))

### Signature analyses

Many known signatures are significantly enriched for top genes of  $v = 5$ . Only those with enrichments stronger than 0.8 are listed<sup>(Table III.3.3.4.b)</sup>. The full signature analyses table is available at [D=Interpretation\genomic\v005\SA.](#)

| Signatures            |   |           |            | GSEA             |        |      | Basic Statistics              |                          |                 |              |
|-----------------------|---|-----------|------------|------------------|--------|------|-------------------------------|--------------------------|-----------------|--------------|
| Signatures DB         | Signature Name                                      | # defined | # measured | Enrichment score | $p$    | FDR  | Mean log <sub>2</sub> (ratio) | $p$ (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| MolSigDBv4_0_dMay2014 | Ffarmer_breast_cancer_cluster_5                     | 19        | 19         | 0.982            | 0.0019 | 0.2% | -1.697                        | 4.2E-11                  | 100.0%          | 0.0%         |
| MolSigDBv4_0_dMay2014 | Gnf2_cdh11  | 25        | 25         | 0.937            | 0.0019 | 0.2% | -1.562                        | 1.2E-11                  | 100.0%          | 0.0%         |
| StaudtSigDB_dNov2012  | Normal_mesenchymal-1_Node1643                       | 77        | 77         | 0.936            | 0.0021 | 0.2% | -1.368                        | 2.3E-34                  | 100.0%          | 0.0%         |
| MolSigDBv4_0_dMay2014 | Anastassiou_cancer_mesenchymal-transition_signature | 64        | 64         | 0.921            | 0.002  | 0.2% | -1.654                        | 8.7E-29                  | 99.7%           | 0.3%         |
| StaudtSigDB_dNov2012  | Lymph_node_LymphDx                                  | 56        | 56         | 0.902            | 0.0021 | 0.2% | -1.183                        | 1.5E-19                  | 98.2%           | 1.8%         |
| StaudtSigDB_dNov2012  | Lymph_node_High_vs_low                              | 651       | 645        | 0.871            | 0.002  | 0.2% | -0.872                        | 4.9E-188                 | 99.5%           | 0.5%         |
| MolSigDBv4_0_dMay2014 | GNF2_PTX3   | 36        | 36         | 0.860            | 0.002  | 0.2% | -1.346                        | 8.5E-13                  | 100.0%          | 0.0%         |
| StaudtSigDB_dNov2012  | <b>Stromal-1_DLBCL_survival_predictor</b>           | 263       | 248        | 0.855            | 0.0021 | 0.2% | -0.997                        | 6.9E-76                  | 98.4%           | 1.6%         |
| GeneSigDB_v4_Sept2011 | Lymphoma_Piccaluga07_64genes                        | 60        | 59         | 0.839            | 0.002  | 0.2% | -1.026                        | 1.4E-18                  | 98.3%           | 1.7%         |
| GeneSigDB_v4_Sept2011 | Breast_Miller07_19genes                             | 17        | 17         | 0.837            | 0.002  | 0.2% | -1.267                        | 7.2E-06                  | 98.5%           | 1.5%         |
| StaudtSigDB_dNov2012  | Lymph_node_U133plus                                 | 217       | 215        | 0.828            | 0.0022 | 0.2% | -0.843                        | 5.1E-60                  | 98.8%           | 1.2%         |
| GeneSigDB_v4_Sept2011 | StemCell_Menicarin09_13genes                        | 12        | 12         | 0.813            | 0.002  | 0.2% | -1.084                        | 1.2E-03                  | 95.2%           | 4.8%         |
| GeneSigDB_v4_Sept2011 | Prostate_Chambers09_40genes                         | 28        | 28         | 0.805            | 0.0019 | 0.2% | -1.195                        | 1.8E-08                  | 100.0%          | 0.0%         |
| GeneSigDB_v4_Sept2011 | Lymphoma_VanLoo09_11genes                           | 11        | 11         | -0.810           | 0.0019 | 0.2% | 0.265                         | 5.7E-03                  | 6.3%            | 93.7%        |
| MolSigDBv4_0_dMay2014 | RRNA_metabolic_process                              | 16        | 16         | -0.817           | 0.0019 | 0.2% | 0.252                         | 3.8E-07                  | 0.0%            | 100.0%       |
| MolSigDBv4_0_dMay2014 | Ribosome_biogenesis_and_assembly                    | 18        | 18         | -0.880           | 0.0019 | 0.2% | 0.246                         | 5.7E-08                  | 0.0%            | 100.0%       |
| MolSigDBv4_0_dMay2014 | RRNA_processing                                     | 15        | 15         | -0.880           | 0.002  | 0.2% | 0.265                         | 2.1E-07                  | 0.0%            | 100.0%       |

Table III.3.3.4.b) Top-enriched signatures by  $v = 5$

Signatures with  $|\text{enrichment score}| \geq 0.8$  and at least 10 measured members are listed for genes ranked by GEP effect  $v = 5$ . All GSEA  $p$  values are based on permutation tests; hence, they are lower-bounded by  $1/(\#\text{permutations})$  and true  $p$  values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of  $t$ -tests against zero regulation as weights).

Given the size of effect  $v = 5$  and the high correlation of top genes to it (and thus between each other), it is interesting that several significantly and highly enriched signatures are relatively small. As any definition that includes only some but not all highly correlated genes would be biased, I would have expected larger signatures. (Maybe these signatures were defined in a constraint signal context or were intersected with biologically motivated tertiary signatures.)

### Role as "stromal-1" signature in a previous CHOP-based DLBCL survival predictor

One of the larger listed top signatures is the stromal-1 signature<sup>(online interpretation card, Figure III.3.3.4.c)</sup> that has already been identified for a previous DLBCL survival predictor<sup>[5]</sup> and has also been associated with the extracellular matrix. The previous predictor was trained with samples from CHOP-treated patients only. Hence, it is consistent that it included effect  $v = 5$ , as it shows the most consistent survival prediction for CHOP-treated patients<sup>(Figure III.2.4.1.a)</sup>.

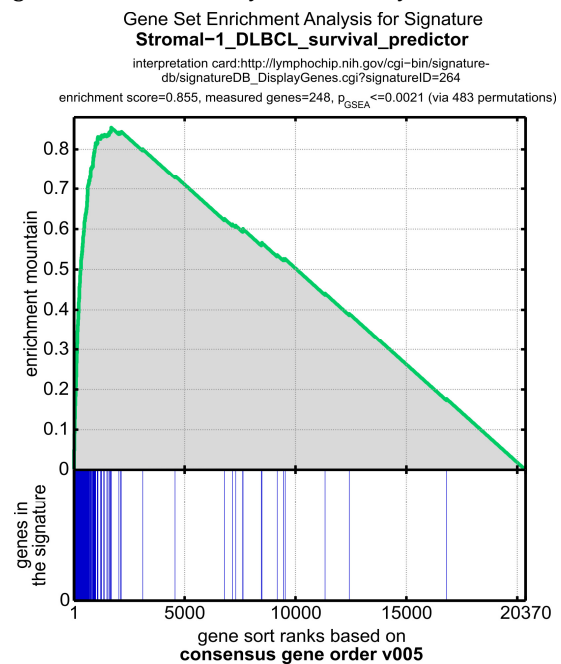


Figure III.3.3.4.c) Significant enrichment of the stromal-1 signature<sup>[5]</sup>





### ■ Top genes overview

For probesets with  $|r_5^g| \geq 0.4$ , 783 unique genes from 24 chromosomes participate in the effect. This clearly indicates that it does not just reflect genomically local aberrations on DNA level, as could be expected by its identified biological function. Even for  $|r_5^g| \geq 0.5$ , 417 unique genes from 23 chromosomes remain.

As the biological function of this effect has already been identified by gene ontology terms, an individual analysis of its top-correlated genes has not been performed.

### ■ Inference

Gene ontology terms and gene signatures with significant enrichment only concern small subsets of the effect's correlated top genes. Hence, they may not be able to biologically describe this large effect completely. Maybe a common yet unidentified biological cause exists that regulates *all* these genes coordinately, resulting in observed correlations. It might be biologically interesting to search for such a common cause.

As previously quantified, the predictive value of this stromal effect is limited in the Rituximab era. If this prognostic difference between CHOP and R-CHOP therapies validates biologically, it might be interesting to find the cause, why Rituximab seems to predominantly help patients with lower expressions of  $\nu = 5$ .

Already in the CHOP era, this effect seems to have measured an effect of the tumor microenvironment. Additionally, it did not qualify as one of the best five survival effects when training with all samples. (Survival dependencies explained by  $\nu = 5$  can already be explained by the selected primary and secondary predictor variables  $\nu \in \{134, 131\}$  (cf. III.2.5.1).) Hence and analytically,  $\nu = 5$  seems to be of second priority towards disclosing DLBCL's molecular pathogenesis.

## III.3.3.5 $\nu = 44$ : Another stromal effect with a hierarchical survival dependency and revisiting a previous DLBCL survival predictor


---

### ■ Role in survival analysis

Effect  $\nu = 44$  facilitates a significant additional survival prediction ( $p = 2.5 \cdot 10^{-7}$ ) on top of effect  $\nu = 5$  for CHOP treated DLBCL patients (Figure III.2.4.1.b). Like  $\nu = 127$  relative to  $\nu = 134$ , this survival trend is *anti-aligned* (Table III.2.4.2) to the primary CHOP predictor variable  $\nu = 5$ .

The survival dependency predicted by this effect is hierarchical. More precisely, it can predict significant survival differences within the higher risk partition of  $\nu = 5$ , but not within the lower risk partition of this stromal effect. The same hierarchical effect is still strong and significant following R-CHOP therapy (cf. Figure III.2.4.3). A subtype-specific analysis of  $\nu = 44$  revealed the same but considerably weaker survival dependency (cf. Figure III.2.4.4), suggesting that not the COO induced effect, but effect  $\nu = 5$  is required to biologically understand this hierarchy.

### Effect overview

The effect has been unsupervisedly discovered in three of four cohorts with detection ranks #5<sup>(GSE10846.CHOP)</sup>, #18<sup>(GSE10846.R-CHOP)</sup> and #32<sup>(GSE31312.R-CHOP)</sup>. It has been supervisedly validated on GEP level in all four cohorts (see 44\* files in  *C=Consensus Effects\* (cohort subfolder)).


Its signal is one-sided<sup>(Figure III.3.3.5)</sup> with approximately 100 top probesets (with relative correlation  $\geq 0.5$ ).

Higher expressions of this effect correspond to higher risk (for a subset of DLBCL patients that shows higher risk with respect to effect  $\nu = 5$ ).

### Clinical associations

Besides associations to survival, it is interesting that in all three cohorts with ECOG data, ECOG performance is significantly and consistently associated with this effect for ABC DLBCL patients, but not so for GCB DLBCL patients. More precisely, significantly more ABC DLBCL patients with ECOG state  $\leq 1$  are found in the lower expression and lower risk partition of  $\nu = 44$  ( $p_{GSE10846.CHOP.ABC} = 0.01$ ,  $p_{GSE10846.R-CHOP.ABC} = 0.02$  and  $p_{GSE31312.R-CHOP.ABC} = 0.03$ ;  $\chi^2$  tests based on contingency tables), whereas this is not the case for GCB DLBCL ( $p_{GSE10846.CHOP.GCB} = 0.73$ ,  $p_{GSE10846.R-CHOP.GCB} = 0.10$  and  $p_{GSE31312.R-CHOP.GCB} = 0.99$ ). This is consistent with the effect's hierarchical prediction of survival, as the majority of samples in the higher risk partition of the stromal effect  $\nu = 5$  are classified as ABC DLBCL. However and again, only partitioning by  $\nu = 5$  and not by the COO induced effect can reveal this hierarchical survival dependency in a clear-cut way<sup>(compare Figure III.2.4.3 with Figure III.2.4.4)</sup>.

### Genomic associations

Gene ontology analyses showed some significantly and specifically overrepresented biological processes, but no molecular functions or cellular components. The two most specific GO terms are positive regulation of macrophage derived foam cell differentiation<sup>(cf. GO:0010744)</sup> (3/14 top genes belong to this term, while it is comprised of 14/20370 measured genes;  $p = 9.4 \cdot 10^{-8}$  via hypergeometric test) and triglyceride catabolic process<sup>(cf. GO:0019433)</sup> (3/14 top genes belong to this term, while it is comprised of 19/20370 measured genes;  $p = 2.5 \cdot 10^{-7}$  via hypergeometric test). All gene ontology analyses for this effect are available at  *D=Interpretation\genomic\44\GOA*.

Gene set enrichment analyses<sup>(Table III.3.3.5.a)</sup> revealed that  $\nu = 44$  is the unsupervisedly rediscovered version of the stromal-2 signature<sup>(online interpretation card, [5])</sup> from a previous CHOP-based DLBCL survival predictor; this effect has been associated with tumor blood vessel density<sup>[5]</sup>.

| Signatures            |   |           |            | GSEA             |        |      | Basic Statistics              |                          |                 |              |
|-----------------------|---|-----------|------------|------------------|--------|------|-------------------------------|--------------------------|-----------------|--------------|
| Signatures DB         | Signature Name                                    | # defined | # measured | Enrichment score | $p$    | FDR  | Mean log <sub>2</sub> (ratio) | $p$ (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| StaudtSigDB_dNov2012  | <b>Stromal-2_DLBCL_survival_predictor</b>         | 62        | 60         | 0.853            | 0.0021 | 0.2% | -0.879                        | 3.1E-21                  | 100.0%          | 0.0%         |
| GeneSigDB_v4_Sept2011 | Stomach_Yu04_17genes                              | 13        | 11         | 0.687            | 0.0020 | 0.2% | -0.947                        | 0.0033                   | 86.4%           | 13.6%        |
| MolSigDBv4_0_dMay2014 | Nakayama_soft_tissue_tumors_pca2_dn               | 80        | 80         | 0.794            | 0.0020 | 0.2% | -0.586                        | 6.0E-15                  | 97.8%           | 2.2%         |
| HGNCSigDB_dMay2014    | Metallothioneins                                  | 19        | 10         | 0.782            | 0.0021 | 0.2% | -0.368                        | 0.0021                   | 87.3%           | 12.7%        |
| MolSigDBv4_0_dMay2014 | Schaeffer_prostate_development_and_cancer_box4_up | 11        | 11         | 0.679            | 0.0020 | 0.2% | -0.466                        | 0.0431                   | 91.2%           | 8.8%         |
| GeneSigDB_v4_Sept2011 | Lymphoma_Blenk08_16genes                          | 11        | 11         | -0.691           | 0.0019 | 0.2% | 0.385                         | 1.1E-07                  | 0.0%            | 100.0%       |

Table III.3.3.5.a) Top-enriched signatures by  $\nu = 44$

Signatures with  $|\text{enrichment score}| \geq 0.67$ , a mean  $\log_2(\text{ratio}) \geq 0.2$  and at least 10 measured members are listed for genes ranked by GEP effect  $\nu = 44$ . All GSEA  $p$  values are based on permutation tests; hence, they are lower-bounded by  $1/(\#\text{permutations})$  and true  $p$  values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of  $t$ -tests against zero regulation as weights).



### Top genes overview

For probesets with  $|r_{44}^g| \geq 0.4$ , 102 unique genes from 22 chromosomes participate in the effect, clearly indicating that it does not just reflect local aberrations on DNA level, as could be expected by its identified biological function. Even for  $|r_{44}^g| \geq 0.5$ , 42 unique genes from 17 chromosomes remain.

As the effect has already been identified as the rediscovered stromal-2 signature, literature screening for individual top genes is not performed.

### Revisiting a previous CHOP-based survival predictor

A previous DLBCL survival predictor utilized a trivariate Cox model<sup>[5]</sup>. It is based on genes that were *supervisedly* selected utilizing CHOP survival data<sup>[5,51]</sup>. Restricting the analysis to this gene selection, signature candidates were formed by hierarchical clustering.

All three explanatory variables of this previous predictor have been unsupervisedly rediscovered by signal dissection:

- (a) the “Germinal center B cell DLBCL survival predictor” signature is enriched with score  $-0.867$  by top genes of effect  $\nu = 129$ ,
- (b) the “stromal-1” signature is enriched with score  $0.855$  by top genes of effect  $\nu = 5$  and
- (c) the “stromal-2” signature is enriched with score  $0.853$  by top genes of this effect  $\nu = 44$ .

Interestingly, effects  $\nu \in \{5, 44\}$  already suffice to explain CHOP patient survival<sup>(Figure III.2.4.2)</sup> to an extent that the standard COO induced effect  $\nu = 129$  did not show any significant additional explanatory value ( $p = 0.26$ ). (Effect  $\nu = 134$  was still able to contribute with  $p = 1.5 \cdot 10^{-4}$  on top of  $\nu \in \{5, 44\}$  in CHOP, but was not selected because of the chosen tight significance threshold at  $10^{-4}$ .)

Because of only 181 available follow-up events for GSE10846.CHOP, the supervised gene selection procedure underlying this previous predictor might have selected many genes that are not specific to DLBCL<sup>(cf. III.2.1.2)</sup>.

| Top genes in $\nu=44$ |             |            |                |                |                |                                     |
|-----------------------|-------------|------------|----------------|----------------|----------------|-------------------------------------|
| Probeset              | Gene Symbol | HG19 locus | $a_{44}^{g,c}$ | $r_{44}^{g,c}$ | $p_{44}^{g,c}$ | $ a_{44}^{g,c}  \cdot r_{44}^{g,c}$ |
| 207175_at             | ADIPOQ      | 3q27.3     | 1.69           | 0.87           | 2.3E-53        | 1.47                                |
| 203980_at             | FABP4       | 8q21.13    | 1.64           | 0.85           | 3.3E-49        | 1.40                                |
| 209613_s_at           | ADH1B       | 4q23       | 1.49           | 0.81           | 1.9E-41        | 1.22                                |
| 209612_s_at           | ADH1B       | 4q23       | 1.15           | 0.70           | 6.2E-26        | 0.80                                |
| 228766_at             | CD36        | 7q21.11    | 1.14           | 0.67           | 9.6E-24        | 0.76                                |
| 218087_s_at           | SORBS1      | 10q24.1    | 0.89           | 0.68           | 1.0E-24        | 0.61                                |
| 235978_at             | FABP4       | 8q21.13    | 0.82           | 0.74           | 1.1E-30        | 0.61                                |
| 209555_s_at           | CD36        | 7q21.11    | 0.97           | 0.62           | 1.5E-19        | 0.60                                |
| 225207_at             | PDK4        | 7q21.3     | 0.88           | 0.66           | 8.8E-23        | 0.58                                |
| 205913_at             | PLIN1       | 15q26.1    | 0.71           | 0.76           | 1.2E-33        | 0.54                                |
| 1565162_s_at          | MGST1       | 12p12.3    | 1.02           | 0.51           | 1.0E-12        | 0.52                                |
| 201348_at             | GPX3        | 5q33.1     | 0.92           | 0.56           | 1.4E-15        | 0.51                                |
| 203548_s_at           | LPL         | 8p21.3     | 0.85           | 0.61           | 1.6E-18        | 0.51                                |
| 224918_x_at           | MGST1       | 12p12.3    | 0.95           | 0.54           | 4.5E-14        | 0.51                                |
| 203649_s_at           | PLA2G2A     | 1p36.13    | 0.90           | 0.56           | 9.6E-16        | 0.51                                |
| 214091_s_at           | GPX3        | 5q33.1     | 0.84           | 0.61           | 2.0E-18        | 0.51                                |
| 206488_s_at           | CD36        | 7q21.11    | 0.81           | 0.59           | 4.3E-17        | 0.48                                |
| 225987_at             | STEAP4      | 7q21.12    | 0.79           | 0.59           | 3.3E-17        | 0.46                                |
| 205498_at             | GHR         | 5p13.1     | 0.65           | 0.71           | 5.9E-28        | 0.46                                |
| 229476_s_at           | THRSP       | 11q14.1    | 0.69           | 0.65           | 1.3E-21        | 0.45                                |
| 231736_x_at           | MGST1       | 12p12.3    | 0.85           | 0.51           | 7.8E-13        | 0.44                                |
| 204955_at             | SRPX        | Xp11.4     | 0.69           | 0.63           | 2.5E-20        | 0.43                                |
| 203549_s_at           | LPL         | 8p21.3     | 0.73           | 0.59           | 4.1E-17        | 0.43                                |
| 222513_s_at           | SORBS1      | 10q24.1    | 0.63           | 0.66           | 1.3E-22        | 0.42                                |
| 219140_s_at           | RBP4        | 10q23.33   | 0.64           | 0.65           | 5.5E-22        | 0.42                                |
| 208383_s_at           | PCK1        | 20q13.31   | 0.75           | 0.52           | 5.5E-13        | 0.39                                |
| 204154_at             | CDO1        | 5q22.3     | 0.69           | 0.54           | 2.3E-14        | 0.37                                |
| 209763_at             | CHRD1L      | Xq23       | 0.65           | 0.57           | 2.3E-16        | 0.37                                |
| 215049_x_at           | CD163       | 12p13.31   | 0.74           | 0.48           | 3.0E-11        | 0.36                                |
| 203571_s_at           | C10orf116   | 10q23.2    | 0.61           | 0.59           | 4.0E-17        | 0.36                                |
| 201540_at             | FHL1        | Xq26.3     | 0.61           | 0.58           | 1.2E-16        | 0.35                                |
| 202992_at             | C7          | 5p13.1     | 0.70           | 0.49           | 8.7E-12        | 0.34                                |
| 49452_at              | ACACB       | 12q24.11   | 0.67           | 0.49           | 1.2E-11        | 0.33                                |
| 206157_at             | PTX3        | 3q25.32    | 0.64           | 0.49           | 1.9E-11        | 0.31                                |
| 43427_at              | ACACB       | 12q24.11   | 0.60           | 0.51           | 7.5E-13        | 0.31                                |
| 204719_at             | ABCA8       | 17q24.2    | 0.60           | 0.51           | 6.4E-13        | 0.31                                |
| 1552509_a_at          | CD300LG     | 17q21.31   | 0.45           | 0.67           | 4.0E-23        | 0.30                                |
| 204894_s_at           | AOC3        | 17q21.31   | 0.53           | 0.56           | 1.0E-15        | 0.30                                |
| 238066_at             | RBP7        | 1p36.22    | 0.54           | 0.55           | 7.9E-15        | 0.30                                |
| 214456_x_at           | SAA1        | 11p15.1    | 0.53           | 0.55           | 5.6E-15        | 0.29                                |
| 203645_s_at           | CD163       | 12p13.31   | 0.65           | 0.45           | 1.1E-09        | 0.29                                |
| 226304_at             | HSPB6       | 19q13.12   | 0.53           | 0.53           | 9.8E-14        | 0.28                                |
| 205382_s_at           | CFD         | 19p13.3    | 0.56           | 0.46           | 2.1E-10        | 0.26                                |
| 209699_x_at           | AKR1C2      | 10p15.1    | 0.45           | 0.57           | 5.0E-16        | 0.26                                |
| 210299_s_at           | FHL1        | Xq26.3     | 0.52           | 0.49           | 1.1E-11        | 0.26                                |
| 207277_at             | CD209       | 19p13.2    | 0.48           | 0.50           | 2.1E-12        | 0.24                                |
| 228854_at             |             | 11q23.2    | 0.60           | 0.41           | 3.8E-08        | 0.24                                |
| 203305_at             | F13A1       | 6p25.1     | 0.55           | 0.44           | 2.9E-09        | 0.24                                |
| 201785_at             | RNASE1      | 14q11.2    | 0.48           | 0.48           | 2.6E-11        | 0.23                                |
| 216333_x_at           | TNXA        | 6p21.33    | 0.40           | 0.57           | 7.5E-16        | 0.23                                |
| 218736_s_at           | PALMD       | 1p21.2     | 0.43           | 0.53           | 1.5E-13        | 0.23                                |
| 209614_at             | ADH1B       | 4q23       | 0.42           | 0.54           | 3.4E-14        | 0.23                                |
| 202291_s_at           | MGP         | 12p12.3    | 0.53           | 0.40           | 4.7E-08        | 0.21                                |
| 219295_s_at           | PCOLCE2     | 3q23       | 0.48           | 0.44           | 2.3E-09        | 0.21                                |
| 212097_at             | CAV1        | 7q31.2     | 0.48           | 0.43           | 3.4E-09        | 0.21                                |
| 219607_s_at           | MS4A4A      | 11q12.2    | 0.50           | 0.42           | 1.4E-08        | 0.21                                |
| 222717_at             | SDPR        | 2q32.3     | 0.41           | 0.50           | 2.7E-12        | 0.21                                |
| 205559_s_at           | PCSK5       | 9q21.13    | 0.42           | 0.48           | 4.4E-11        | 0.20                                |
| 225575_at             | LIFR        | 5p13.1     | 0.43           | 0.46           | 3.7E-10        | 0.20                                |
| 205392_s_at           | CCL14       | 17q12      | 0.41           | 0.48           | 5.0E-11        | 0.19                                |
| 201010_s_at           | TXNIP       | 1q21.1     | 0.42           | 0.44           | 1.7E-09        | 0.19                                |
| 208607_s_at           | SAA1        | 11p15.1    | 0.44           | 0.41           | 3.4E-08        | 0.18                                |
| 219519_s_at           | SIGLEC1     | 20p13      | 0.43           | 0.42           | 1.5E-08        | 0.18                                |
| 228335_at             | CLDN11      | 3q26.2     | 0.42           | 0.42           | 9.3E-09        | 0.18                                |
| 208131_s_at           | PTGIS       | 20q13.13   | 0.42           | 0.41           | 2.0E-08        | 0.17                                |
| 202409_at             | IGF2        | 11p15.5    | 0.41           | 0.40           | 6.1E-08        | 0.16                                |

Table III.3.5.b) Top genes in validated effect  $\nu=44$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx<sup>TM</sup>[97] v33)  
 $a_{44}^{g,c}$  Components of the consensus gene axis of effect  $\nu=44$  (cf. Table III.1.5); filtered  $|a_{44}^{g,c}| \geq 0.4$ .  
 $r_{44}^{g,c}$  Consensus gene correlations; filtered  $|r_{44}^{g,c}| \geq 0.4$ .  
 $p_{44}^{g,c}$  p values for the correlations (cf. II.5.2.1)

More importantly, this previous model reserved all R-CHOP treated patients for validation purposes and therefore discovered GEP effects for survival prediction that are optimal for CHOP<sup>(as confirmed by this work in III.2.4.1)</sup>, but unfortunately suboptimal for the Rituximab-added therapy<sup>(cf. III.2.2.1 and III.2.5.1)</sup>.

In contrast, signal dissection discovers effects unsupervisedly based on GEP information only. Hence, all survival associations can be regarded as independent validation of these GEP effects. More importantly, this allows using *all* available survival data via leave-one-out validation<sup>(III.2.5.3)</sup> for predictor construction. The unbiased predictor construction procedure<sup>(cf. III.2.5.1)</sup> consequently can identify DLBCL survival effects based on survival data from much more patients. Correspondingly low  $p$  values indicate that GEP survival effects identified in this way can explain observed survival more reliably and are relevant for both CHOP and R-CHOP therapies<sup>(cf. III.2.5.4)</sup>.

Furthermore, discovered GEP effects always include as many top genes as correlations between gene expressions exist in the signal, i.e. they are not restricted to genes preselected with limited survival information. Hence, also GEP effects could be discovered that are not associated with outcome following contemporary therapies, but that might still reveal biologically interesting molecular differences between DLBCL patients<sup>(e.g. III.3.4.3)</sup>.

### ■ Inference

Both stromal effects  $v \in \{5, 44\}$  seem to concern the microenvironment, rather than gene expressions originating from DLBCL tumor cells. Additionally, effect  $v = 5$  is associated with a strong survival dependency following CHOP therapy, but no longer following current standard R-CHOP therapy. However, both effects together predict a hierarchical survival dependency for a subset of DLBCL that is still strong following R-CHOP therapy<sup>(cf. Figure III.2.4.3)</sup>. It might be biologically interesting to understand this hierarchical relation between these stromal effects. Still, in order to investigate the molecular pathogenesis of DLBCL, other discovered survival effects may provide more direct information.

### III.3.3.6 $v = 19$ : A T cell related effect (quinivariate model, 3<sup>rd</sup> predictor variable)

---

#### ■ Role in survival analysis

Effect  $v = 19$  facilitates significant ( $p = 7.4 \cdot 10^{-5}$ ) survival prediction on top of survival already explained by effects  $v \in \{134, 131\}$ <sup>(Figure III.2.5.1.c)</sup>. Together with  $v = 75$  it additionally shows a complex hierarchical dependency for effect  $v = 3$ <sup>(as described in III.3.3.8)</sup>.

#### ■ Effect overview

Effect  $v = 19$  is another unsupervisedly quad-discovered effect. Detection ranks are #47<sup>(GSE10846.CHOP)</sup>, #29<sup>(GSE10846.R-CHOP)</sup>, #29<sup>(GSE4475.nonMBL)</sup> and #57<sup>(GSE31312.R-CHOP)</sup>. It has a one-sided gradual signal<sup>(e.g. Figure III.3.3.6.a)</sup> with approximately 75 probesets (relative correlation  $\geq 0.5$ ).

Higher expressions correspond to lower risk<sup>(Table III.2.5.2)</sup> for a subset of patients.

#### ■ Clinical associations

Clinically, it is interesting that in all three cohorts with LDH data, LDH ratios are significantly and consistently associated with effect  $v = 19$  for GCB DLBCL patients, but not so for ABC DLBCL patients. More precisely, the

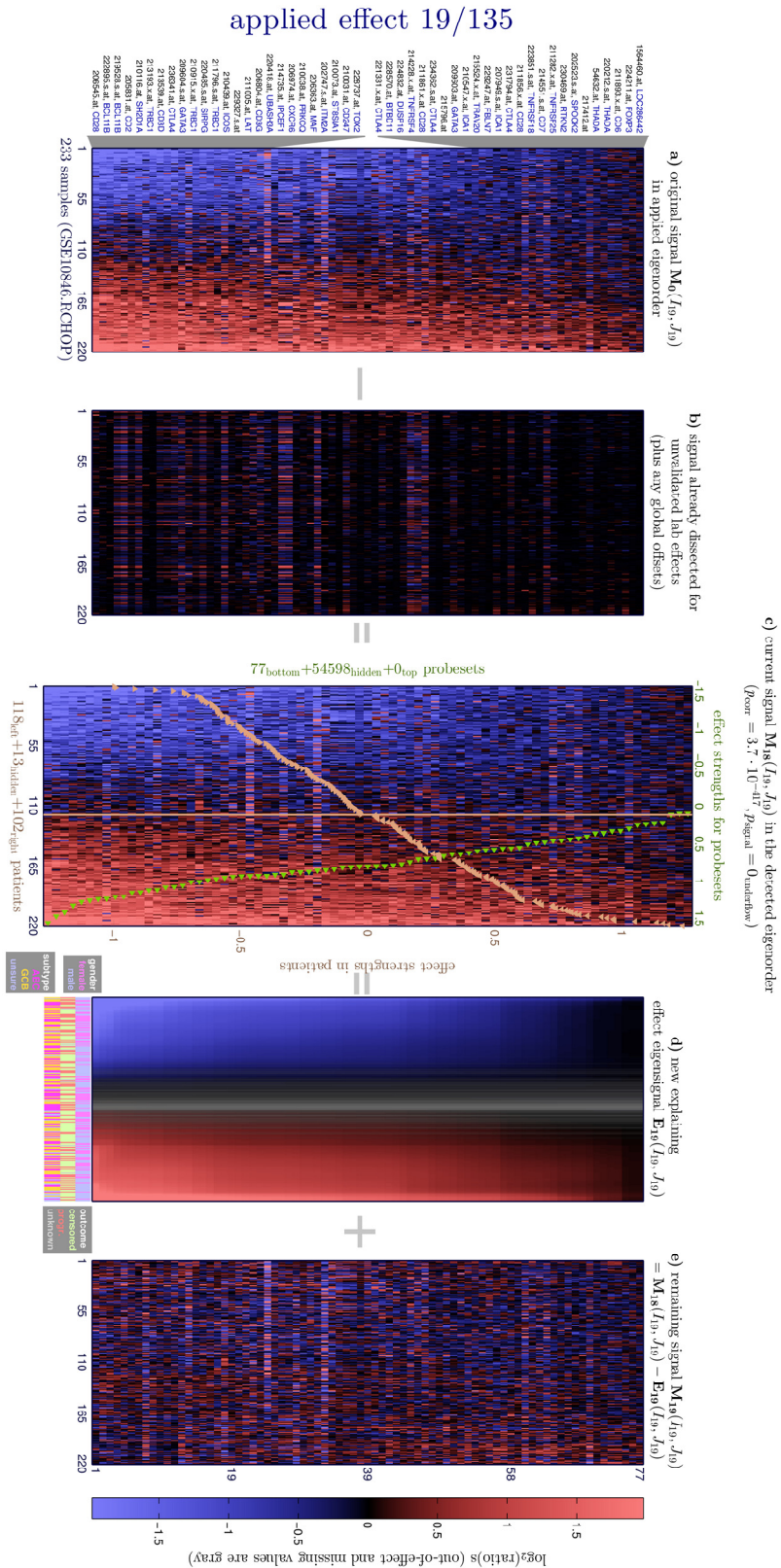


Figure III.3.3.6.a) Validated effect  $v=19$ , applied to GSE10846.R-CHOP

Third survival effect by multi-cohort based survival analysis<sup>[112.5]</sup> applied to GSE10846.R-CHOP (233 patients).

(The genomic consensus effect is applied to the cleaned signal without lab effects<sup>[cf. III.1.4.2]</sup>. Samples and probesets are ordered by their effect strengths in this cohort<sup>[cf. Table III.1.5]</sup>. Additionally, probesets are filtered by demanding a relative correlation stronger than 0.5. The effect's bimonotonic eigensignal<sup>[panel d)</sup> is grayed for samples having insufficient or insignificant correlation to this effect<sup>[II.4.2.1]</sup>.)

partition with lower average expression contains significantly more GCB DLBCL patients with LDH ratio  $\geq 1$  ( $p_{GSE10846.CHOP.GCB} = 0.01$ ,  $p_{GSE10846.R-CHOP.GCB} = 7.7 \cdot 10^{-3}$  and  $p_{GSE31312.R-CHOP.GCB} = 0.03$ ;  $\chi^2$  tests based on contingency tables). This is consistent with the overall survival trend predicted by  $v = 19$ , i.e. higher expression is associated with more favorable outcome (Table III.2.5.2). However, this association does not reliably exist for ABC DLBCL patients ( $p_{GSE10846.CHOP.ABC} = 0.06$ ,  $p_{GSE10846.R-CHOP.ABC} = 0.67$  and  $p_{GSE31312.R-CHOP.ABC} = 0.59$ ), indicating a subtype-specific survival dependency.

### Genomic associations

Gene ontology overrepresentation analyses reveal an overrepresentation of T cell co-stimulation (cf. GO:0031295) with  $p = 5.1 \cdot 10^{-9}$  (Figure III.3.3.6.b). This term contains genes responsible for antigen-independent signaling for T cell activation, i.e. an alternative to T cell receptor signaling.

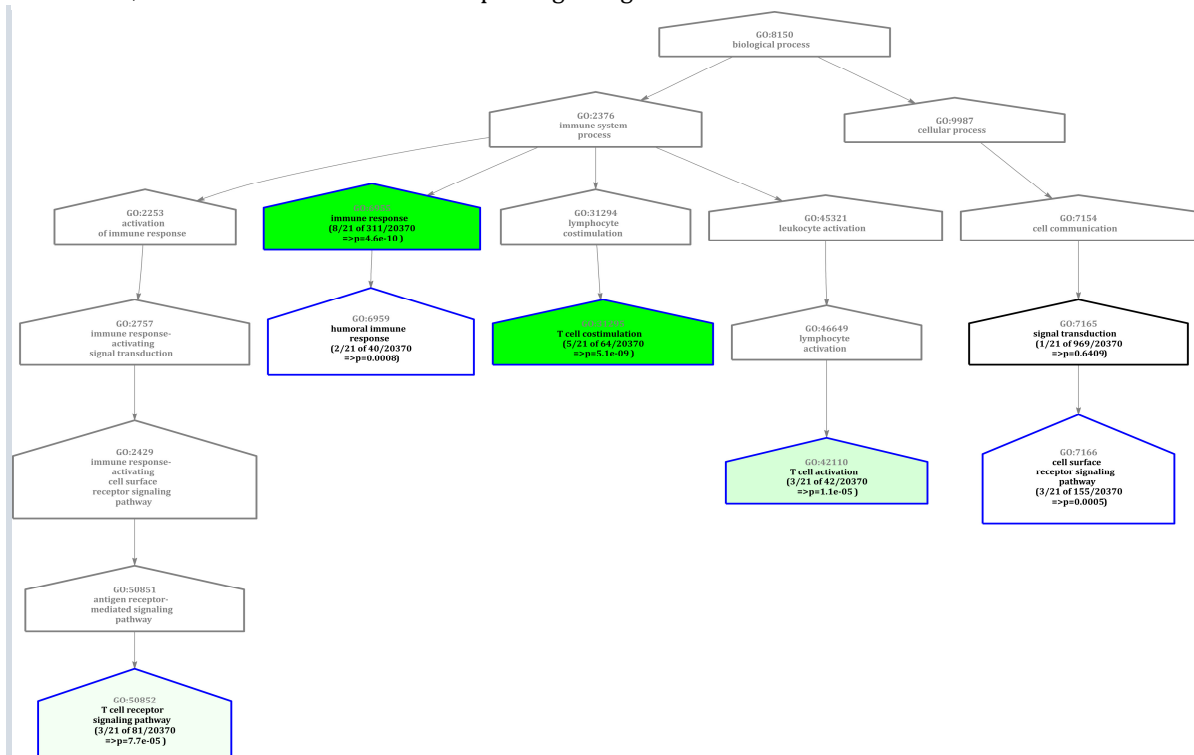


Figure III.3.3.6.b) Gene ontology overrepresentation analyses of biological processes in effect  $v = 19$

All  $p$  values are based on hypergeometric tests; cf. I.3.2.

Signature analyses (Table III.3.3.6.a) confirm this association with T cells. For example, the Biocarta pathway about surface molecules of T helper cells (online interpretation card) is significantly enriched.

| Signatures            |  |           |            | GSEA             |        |      | Basic Statistics      |                          |                 |              |
|-----------------------|--|-----------|------------|------------------|--------|------|-----------------------|--------------------------|-----------------|--------------|
| Signatures DB         | Signature Name                             | # defined | # measured | Enrichment score | $p$    | FDR  | Mean $\log_2$ (ratio) | $p$ (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| MolSigDBv4_0_dMay2014 | Biocarta_T helper_pathway                  | 14        | 12         | 0.886            | 0.0021 | 0.2% | -1.088                | 0.0002                   | 100.0%          | 0.0%         |
| StaudtSigDB_dNov2012  | T_cell                                     | 15        | 14         | 0.851            | 0.0020 | 0.2% | -1.215                | 7.9E-06                  | 93.0%           | 7.0%         |
| MolSigDBv4_0_dMay2014 | Biocarta_Tcytotoxic_pathway                | 14        | 12         | 0.802            | 0.0021 | 0.2% | -1.157                | 7.6E-05                  | 100.0%          | 0.0%         |
| MolSigDBv4_0_dMay2014 | Biocarta_TCRA_pathway                      | 13        | 11         | 0.801            | 0.0020 | 0.2% | -0.906                | 0.0001                   | 100.0%          | 0.0%         |
| MolSigDBv4_0_dMay2014 | Watanabe_ulcerative_colitis_with-cancer_dn | 14        | 14         | 0.670            | 0.0020 | 0.2% | -0.728                | 1.8E-05                  | 97.2%           | 2.8%         |
| MolSigDBv4_0_dMay2014 | GNF2_ATM                                   | 30        | 30         | 0.689            | 0.0019 | 0.2% | -0.564                | 2.3E-06                  | 96.0%           | 4.0%         |

Table III.3.3.6.a) Top-enriched signatures by  $v = 19$

Signatures with  $|\text{enrichment score}| \geq 0.67$  and at least 10 measured members are listed for genes ranked by GEP effect  $v = 19$ .

All GSEA  $p$  values are based on permutation tests; hence, they are lower-bounded by  $1/(\# \text{permutations})$  and true  $p$  values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of  $t$ -tests against zero regulation as weights).



### Top genes overview

For probesets with  $|r_{19}^g| \geq 0.4$  (cf. Table III.3.3.6.b), 70 unique genes from 18 chromosomes participate in the effect, suggesting that it reflects a genomic regulation network and not just local aberrations on DNA level, as could be expected by its identified function in T cell regulation. Even for  $|r_{19}^g| \geq 0.5$ , 33 unique genes from 12 chromosomes remain.

A brief *recherché* for individual top correlated genes (information via <http://www.ncbi.nlm.nih.gov/gene>, [107], April 2015) confirms the association with T cells: CTLA4 can cause inhibitory signals to T cells<sup>[117]</sup>, ICOS stands for “inducible T-cell co-stimulator”, CD28 belongs to one of the dominant co-stimulatory pathways<sup>[118]</sup> and TOX2 is an essential regulator of T-cell differentiation<sup>[119]</sup>.

While the top three genes are all located in 2q33.2, many other measured and expressed probesets from this locus are not correlated to the effect, making a DNA level aberration unlikely (see [DLBCL Master Table 2015, gene orders.xlsx](#)).

### Inference

Based on gene ontology analyses and confirmed by other analyses, this effect is related to T cells, probably to antigen-independent T cell co-stimulation ( $p = 5.1 \cdot 10^{-9}$ ).

To rule out a potential triggering aberration, copy number measurements (e.g. from GSE11318, [11]) for 2q33.2 could be interrogated.

The role of effect  $\nu = 19$  in DLBCL is probably easier to conceive in context of expressions of other effects. For example, it plays a key role in selecting DLBCL patients that show a strong survival dependency on effect  $\nu = 3$ , as will be described in III.3.3.8. This may also help to interpret this effect, maybe validate its association with T cells and ideally unveil their role in DLBCL.

| Top genes in $\nu=19$ |             |            |                |                |                |  |
|-----------------------|-------------|------------|----------------|----------------|----------------|--|
| Probeset              | Gene Symbol | HG19 locus | $a_{19}^{g,c}$ | $r_{19}^{g,c}$ | $p_{19}^{g,c}$ | $\left  \frac{a_{19}^{g,c}}{r_{19}^{g,c}} \right $ |
| 236341_at             | CTLA4       | 2q33.2     | 1.23           | 0.90           | 3.8E-55        | 1.10   |
| 234362_s_at           | CTLA4       | 2q33.2     | 0.90           | 0.86           | 2.1E-44        | 0.77   |
| 210439_at             | ICOS        | 2q33.2     | 0.81           | 0.82           | 3.0E-35        | 0.67   |
| 231794_at             | CTLA4       | 2q33.2     | 0.75           | 0.81           | 4.4E-37        | 0.61   |
| 221331_x_at           | CTLA4       | 2q33.2     | 0.69           | 0.83           | 2.6E-37        | 0.58   |
| 206545_at             | CD28        | 2q33.2     | 0.65           | 0.76           | 5.3E-28        | 0.50   |
| 228737_at             | TOX2        | 20q13.12   | 0.76           | 0.53           | 1.8E-12        | 0.41   |
| 229327_s_at           |             | 16q23.2    | 0.64           | 0.55           | 4.4E-13        | 0.35   |
| 211796_s_at           | TRBC1       | 7q34       | 0.53           | 0.64           | 1.8E-17        | 0.34   |
| 213193_x_at           | TRBC1       | 7q34       | 0.48           | 0.69           | 4.6E-21        | 0.33   |
| 210915_x_at           | TRBC1       | 7q34       | 0.47           | 0.70           | 9.1E-22        | 0.33   |
| 236787_at             |             | 2p11.2     | 0.66           | 0.48           | 3.1E-10        | 0.32   |
| 213135_at             | TIAM1       | 21q22.11   | 0.54           | 0.58           | 4.6E-14        | 0.32   |
| 209348_s_at           | MAF         | 16q23.2    | 0.52           | 0.58           | 4.6E-14        | 0.30   |
| 206363_at             | MAF         | 16q23.2    | 0.48           | 0.61           | 8.1E-16        | 0.29   |
| 214228_x_at           | TNFRSF4     | 1p36.33    | 0.45           | 0.64           | 1.2E-17        | 0.29   |
| 213539_at             | CD3D        | 11q23.3    | 0.41           | 0.66           | 7.8E-19        | 0.27   |
| 210116_at             | SH2D1A      | Xq25       | 0.47           | 0.56           | 7.5E-13        | 0.26   |
| 214551_s_at           | CD7         | 17q25.3    | 0.42           | 0.61           | 1.7E-15        | 0.26   |
| 211005_at             | LAT         | 16p11.2    | 0.40           | 0.65           | 7.8E-18        | 0.26   |
| 219528_s_at           | BCL11B      | 14q32.2    | 0.40           | 0.64           | 8.4E-18        | 0.26   |
| 214049_x_at           | CD7         | 17q25.3    | 0.38           | 0.67           | 3.0E-19        | 0.25   |
| 204777_s_at           | MAL         | 2q11.1     | 0.49           | 0.51           | 2.0E-10        | 0.25   |
| 230469_at             | RTKN2       | 10q21.2    | 0.53           | 0.46           | 1.8E-09        | 0.25   |
| 220485_s_at           | SIRPG       | 20p13      | 0.39           | 0.62           | 2.3E-16        | 0.24   |
| 205831_at             | CD2         | 1p13.1     | 0.38           | 0.64           | 3.0E-17        | 0.24   |
| 207949_s_at           | ICA1        | 7p21.3     | 0.42           | 0.57           | 3.4E-13        | 0.24   |
| 210547_x_at           | ICA1        | 7p21.3     | 0.39           | 0.60           | 3.6E-15        | 0.23   |
| 203828_s_at           | IL32        | 16p13.3    | 0.41           | 0.56           | 8.1E-13        | 0.23   |
| 224211_at             | FOXP3       | Xp11.23    | 0.42           | 0.52           | 5.5E-12        | 0.22   |
| 209670_at             | TRAC        | 14q11.2    | 0.36           | 0.59           | 1.0E-14        | 0.22   |
| 209671_x_at           | TRAC        | 14q11.2    | 0.37           | 0.58           | 9.2E-14        | 0.21   |
| 210972_x_at           | TRAV20      | 14q11.2    | 0.37           | 0.57           | 1.4E-13        | 0.21   |
| 227361_at             | HS3ST3B1    | 17p12      | 0.47           | 0.44           | 1.5E-08        | 0.21   |
| 213958_at             | CD6         | 11q12.2    | 0.30           | 0.68           | 4.6E-20        | 0.21   |
| 202524_s_at           | SPOCK2      | 10q22.1    | 0.36           | 0.55           | 1.1E-12        | 0.20   |
| 209604_s_at           | GATA3       | 10p14      | 0.36           | 0.55           | 2.6E-12        | 0.20   |
| 219423_x_at           | TNFRSF25    | 1p36.31    | 0.37           | 0.52           | 2.9E-11        | 0.20   |
| 211902_x_at           | YME1L1      | 14q11.2    | 0.35           | 0.56           | 8.9E-13        | 0.20   |
| 211339_s_at           | ITK         | 5q33.3     | 0.44           | 0.44           | 6.2E-08        | 0.19   |
| 54632_at              | THADA       | 2p21       | 0.41           | 0.47           | 3.7E-09        | 0.19   |
| 222895_s_at           | BCL11B      | 14q32.2    | 0.37           | 0.51           | 2.7E-11        | 0.19   |
| 229247_at             | FBLN7       | 2q13       | 0.38           | 0.49           | 1.6E-10        | 0.19   |
| 240070_at             | TIGIT       | 3q13.31    | 0.40           | 0.46           | 2.7E-09        | 0.18   |
| 236226_at             | BTLA        | 3q13.2     | 0.43           | 0.42           | 9.4E-08        | 0.18   |
| 216033_s_at           | FYN         | 6q21       | 0.38           | 0.47           | 3.6E-09        | 0.18   |
| 224832_at             | DUSP16      | 12p13.2    | 0.33           | 0.54           | 8.3E-13        | 0.18   |
| 205456_at             | CD3E        | 11q23.3    | 0.31           | 0.55           | 1.7E-12        | 0.17   |
| 220212_s_at           | THADA       | 2p21       | 0.33           | 0.52           | 7.4E-11        | 0.17   |
| 214032_at             | ZAP70       | 2q11.2     | 0.35           | 0.49           | 1.2E-09        | 0.17   |
| 210031_at             | CD247       | 1q24.2     | 0.36           | 0.47           | 6.5E-09        | 0.17   |
| 210073_at             | ST8SIA1     | 12p12.1    | 0.35           | 0.47           | 3.3E-09        | 0.17   |
| 202747_s_at           | ITM2A       | Xq21.1     | 0.35           | 0.47           | 3.1E-09        | 0.16   |
| 223377_x_at           | CISH        | 3p21.2     | 0.36           | 0.45           | 7.9E-09        | 0.16   |
| 223851_s_at           | TNFRSF18    | 1p36.33    | 0.31           | 0.52           | 8.3E-12        | 0.16   |
| 226333_at             | IL6R        | 1q21.3     | 0.40           | 0.40           | 3.5E-07        | 0.16   |
| 1555613_a_at          | ZAP70       | 2q11.2     | 0.35           | 0.45           | 4.7E-09        | 0.16   |
| 211210_x_at           | SH2D1A      | Xq25       | 0.32           | 0.50           | 5.1E-10        | 0.16   |
| 211828_s_at           | TNFK        | 3q26.2     | 0.35           | 0.45           | 3.2E-08        | 0.16   |
| 214735_at             | IPCEF1      | 6q25.2     | 0.32           | 0.48           | 1.4E-09        | 0.15   |
| 212062_at             | ATP9A       | 20q13.2    | 0.36           | 0.43           | 1.2E-07        | 0.15   |
| 202746_at             | ITM2A       | Xq21.1     | 0.38           | 0.41           | 4.8E-07        | 0.15   |
| 212473_s_at           | MICAL2      | 11p15.3    | 0.34           | 0.45           | 3.6E-08        | 0.15   |
| 230489_at             | CD5         | 11q12.2    | 0.31           | 0.48           | 2.9E-10        | 0.15   |
| 206118_at             | STAT4       | 2q32.2     | 0.30           | 0.49           | 8.7E-10        | 0.15   |
| 230488_s_at           | DBH-AS1     | 9q34.2     | 0.36           | 0.40           | 2.8E-07        | 0.14   |
| 218573_at             | MAGEH1      | Xp11.21    | 0.34           | 0.42           | 2.8E-07        | 0.14   |
| 203508_at             | TNFRSF1B    | 1p36.22    | 0.33           | 0.42           | 1.8E-07        | 0.14   |
| 222317_at             | PDE3B       | 11p15.2    | 0.33           | 0.42           | 3.0E-07        | 0.14   |
| 239288_at             | TNFK        | 3q26.31    | 0.31           | 0.40           | 3.6E-07        | 0.12   |

Table III.3.3.6.b) Top genes in validated effect  $\nu=19$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx<sup>TM</sup>[97] v33)  
 $a_{19}^{g,c}$  Components of the consensus gene axis of effect  $\nu=19$  (cf. Table III.1.5); filtered  $|a_{19}^{g,c}| \geq 0.3$ .  
 $r_{19}^{g,c}$  Consensus gene correlations; filtered  $|r_{19}^{g,c}| \geq 0.4$ .  
 $p_{19}^{g,c}$  p values for the correlations (cf. II.5.2.1)

### Role in survival analysis

Effect  $\nu = 75$  is capable of predicting significant survival differences ( $p = 3.6 \cdot 10^{-5}$ ) on top of already explained survival by effects  $\nu \in \{134, 131, 19\}$ . Together with  $\nu = 19$  it additionally reveals a complex hierarchical dependency of effect  $\nu = 3$  (as described in III.3.3.8).

### Effect overview

Two unsupervised detections underlie this consensus effect (with ranks #136 in GSE10846.R-CHOP and #43 in GSE31312.R-CHOP). It has been supervisedly validated in all four cohorts (e.g. Figure III.3.3.7). Detected effects contain several correlated genes, but after soft intersection by consensus gene axis construction (cf. III.1.3) BCL2 is the only remaining gene that is highly correlated to this effect. Eight top-correlated probesets (Table III.3.3.7.a) support this effect; all measure BCL2 or are overlapped by BCL2-annotated probesets in locus 18q21.33.

Hence it cannot be excluded that this effect is a reflection of a chromosomal feature in the GEP signal; examination of DNA measurements like array comparative genomic hybridization might clarify this.

Higher expressions of BCL2 correspond to higher risk and adverse outcome (cf. III.2.5.1) for a DLBCL subset.

### Preliminary top gene analysis

Direct literature screening for BCL2 revealed that it is an already well-known player in several lymphomas, for example in follicular lymphoma<sup>[120]</sup>.

Functionally, Bcl-2 proteins are mainly located in the outer mitochondrial membrane and bind Bax/Bak proteins that can make the mitochondrial membrane permeable as soon as they are released, thereby triggering apoptosis<sup>[121]</sup>. Consistent with adverse outcome for higher expression of  $\nu = 75$ , overexpressed BCL2 has an anti-apoptotic effect and can cause chemotherapy resistance in various human cancers; hence complementing chemotherapy with BCL2-specific agents like small molecule Bcl-2 protein inhibitors has been suggested for clinical trials<sup>[121]</sup>.

With respect to DLBCL, a review in clinical oncology<sup>[122]</sup> reported poor prognosis with standard R-CHOP therapy for patients having a  $t(14;18)$  translocation of BCL2 together with a MYC gene rearrangement (“double-hit DLBCL”). The review concluded from several other studies that a) only concurrent expression of BCL2 and MYC is important for outcome rather than MYC expression alone and that b) the incidence for double-overexpression is 20%-30% in DLBCL, if measured by immunohistochemistry. Routine evaluation of MYC and BCL2 by immunohistochemistry was recommended for clinical management. While it was clear that R-CHOP should be replaced for double-hit patients, the review concluded (in 2012) that no optimal therapy is known and double-hit patients should be referred for clinical trials wherever possible.

### Inference

The unsupervisedly (re)discovered BCL2 GEP effect plays a known important role for a subset of DLBCL patients having a  $t(14;18)$  translocation. This might be helpful when interpreting the complex survival dependency of  $\nu = 3$  in context of this effect and  $\nu = 19$  that only exists for low BCL2 expression (cf. III.3.3.8). Another study has already revealed significantly more favorable outcome for GCB DLBCL patients with high

| Top genes in $\nu = 75$ |             |            |                |                |                |  |
|-------------------------|-------------|------------|----------------|----------------|----------------|--|
| Probeset                | Gene Symbol | HG19 locus | $a_{75}^{g,c}$ | $r_{75}^{g,c}$ | $p_{75}^{g,c}$ | $\left  \frac{a_{75}^{g,c}}{r_{75}^{g,c}} \right $ |
| 244035_at               |             | 18q21.33   | 1.31           | 0.90           | 1.8E-68        | 1.17   |
| 203685_at               | BCL2        | 18q21.33   | 1.33           | 0.80           | 5.1E-44        | 1.06   |
| 232614_at               |             | 18q21.33   | 1.12           | 0.90           | 3.0E-69        | 1.01   |
| 232210_at               |             | 18q21.33   | 1.06           | 0.87           | 1.3E-60        | 0.93   |
| 237837_at               |             | 18q21.33   | 0.79           | 0.65           | 9.3E-24        | 0.51   |
| 203684_s_at             | BCL2        | 18q21.33   | 0.61           | 0.73           | 8.9E-33        | 0.45   |
| 207005_s_at             | BCL2        | 18q21.33   | 0.62           | 0.72           | 2.3E-31        | 0.44   |
| 207004_at               | BCL2        | 18q21.33   | 0.39           | 0.45           | 8.6E-11        | 0.17   |
| 211352_s_at             | NCOA3       | 20q13.12   | 0.33           | 0.36           | 3.0E-07        | 0.12   |
| 1554636_at              |             | 19q13.43   | 0.32           | 0.30           | 2.5E-05        | 0.10   |
| 206951_at               | HIST1H4I    | 6p22.2     | -0.32          | -0.32          | 8.1E-06        | -0.10  |
| 1554878_a_at            | ABCD3       | 1p21.3     | -0.37          | -0.31          | 1.0E-05        | -0.12  |

Table III.3.3.7.a) Top genes in validated effect  $\nu = 75$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx<sup>TM</sup>[97] v33)  
 $a_{75}^{g,c}$  Components of the consensus gene axis of effect  $\nu = 75$  (cf. Table III.1.5); filtered  $\left| a_{75}^{g,c} \right| \geq 0.3$ .  
 $r_{75}^{g,c}$  Consensus gene correlations; filtered  $\left| r_{75}^{g,c} \right| \geq 0.3$ .  
 $p_{75}^{g,c}$   $p$  values for the correlations (cf. II.5.2.1)

BCL2 expression and high NF- $\kappa$ B expression (cf. Figure I.3.3 and [81]). It might be biologically interesting to search for a potential common molecular mechanism involving BCL2 that may link both observations.

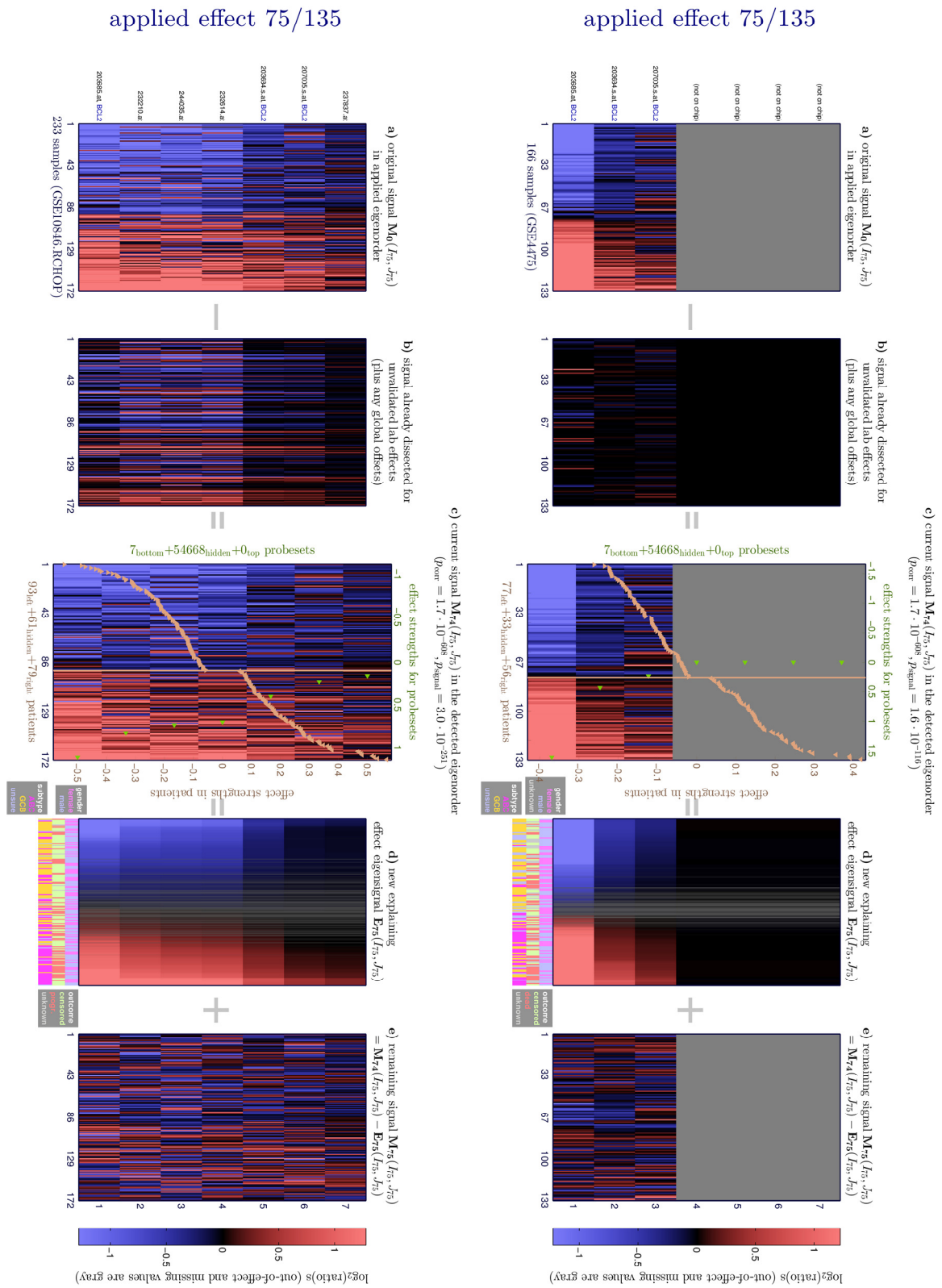


Figure III.3.7) Validated effect  $v=75$ , applied to GSE10846.R-CHOP and GSE4475

Fourth survival effect by multi-cohort based survival analysis<sup>(III.2.5)</sup> applied to GSE10846.R-CHOP (233 patients) and GSE4475 (all 166 patients, including BL patients). (The genomic consensus effect is applied to the cleaned signal without lab effects<sup>(cf. III.1.4.2)</sup>. Samples and probesets are ordered by their effect strengths in this cohort<sup>(cf. Table III.1.5)</sup>. Additionally, probesets are filtered by demanding a relative correlation  $\geq 0.5$ . The effect's bimonotonic eigensignal<sup>(panel d)</sup> is grayed for samples having insufficient or insignificant correlation to this effect<sup>(III.4.2.1)</sup>. Missing probesets for the consensus gene axis due to an older microarray design in GSE4475 are also grayed.)

### III.3.3.8 $\nu = 3$ : A zinc-related effect (quinivariate model, 5<sup>th</sup> predictor variable)

#### Role in survival analysis

The final explanatory variable for the quinivariate predictor is effect  $\nu = 3$ . With  $p = 2.5 \cdot 10^{-5}$  (III.2.5.1), it can predict significant survival differences on top of already explained survival by effects  $\nu \in \{134, 131, 19, 75\}$ . In the final quinivariate model (cf. Table III.2.5.2) it contributes with a Cox  $p$  value of  $6.3 \cdot 10^{-6}$ .

#### Effect overview

Effect  $\nu = 3$  has been unsupervisedly quad-discovered with detection ranks #204<sup>(GSE10846.CHOP)</sup>, #32<sup>(GSE10846.R-CHOP)</sup>, #24<sup>(GSE4475.nonMBL)</sup> and #42<sup>(GSE31312.R-CHOP)</sup>. It has a one-sided gradual signal with few co-regulated genes that show a clear and homogeneous folding between patients in all cohorts (Figure III.3.3.8.b).

Higher expression corresponds to higher risk (Table III.2.5.2) for a subset of DLBCL patients stratified by other predictor variables. A hierarchical survival analysis below clarifies this subset.

#### Genomic associations

Via gene ontology overrepresentation (Figure III.3.3.6.b), it was possible to locate this effect in the perinuclear region of the cytoplasm (cf. GO:0005737) ( $p = 2.0 \cdot 10^{-11}$ , hypergeometric test) where it is involved in negative regulation of growth (cf. GO:0045926) ( $p < 2.2 \cdot 10^{-16}$ ).

Interestingly, top genes of this effect also represent the majority of genes involved in cellular response to cadmium ion (cf. GO:0071276) ( $p = 1.6 \cdot 10^{-15}$ ) and in cellular response to zinc ion (cf. GO:0071294) ( $p < 2.2 \cdot 10^{-16}$ ). This zinc association seems highly specific, as this GO term is comprised of only 10/20370 measured genes and 7/8 top genes of the discovered effect belong to it.

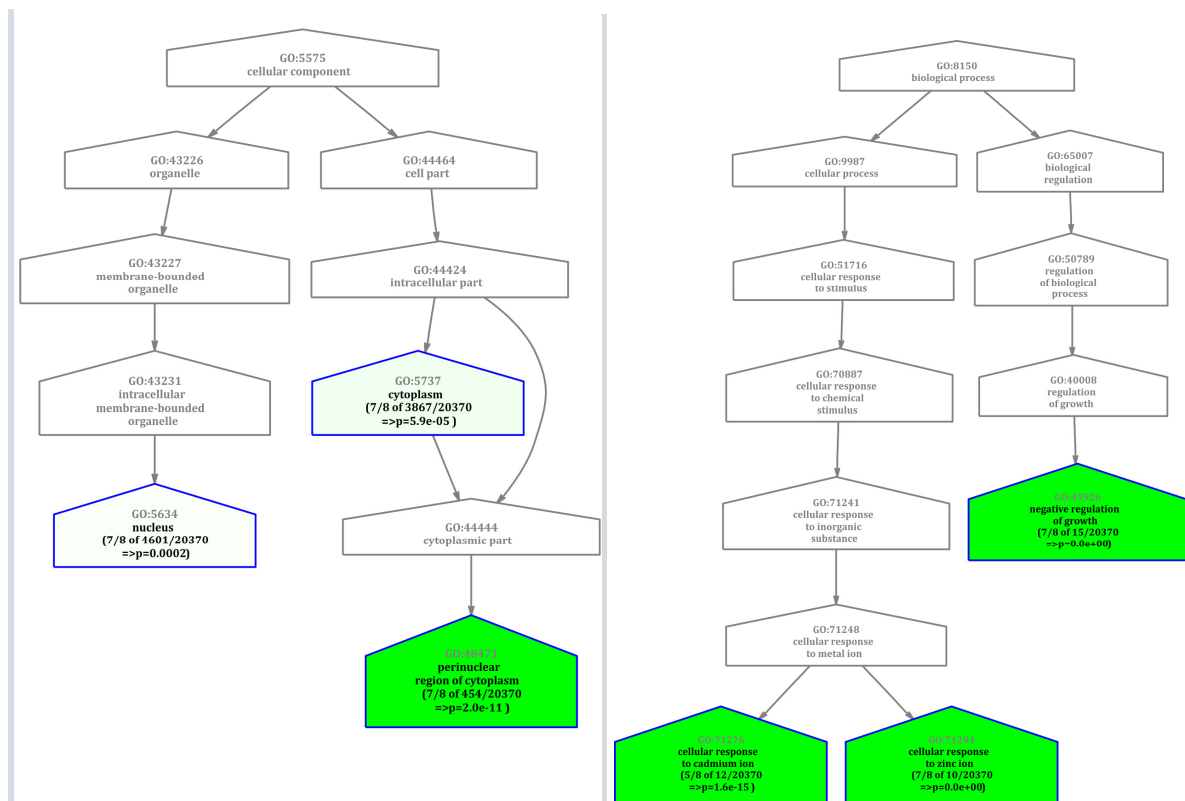


Figure III.3.3.8.a) Gene ontology overrepresentation analyses of biological processes and cellular components for effect  $\nu = 3$

All  $p$  values are based on hypergeometric tests; cf. I.3.2. (A  $p$  value of zero indicates an underflow, i.e.  $p < 2.2 \cdot 10^{-16}$ , which is the numeric resolution limit for differences to one; the true  $p$  value is never exactly zero.)

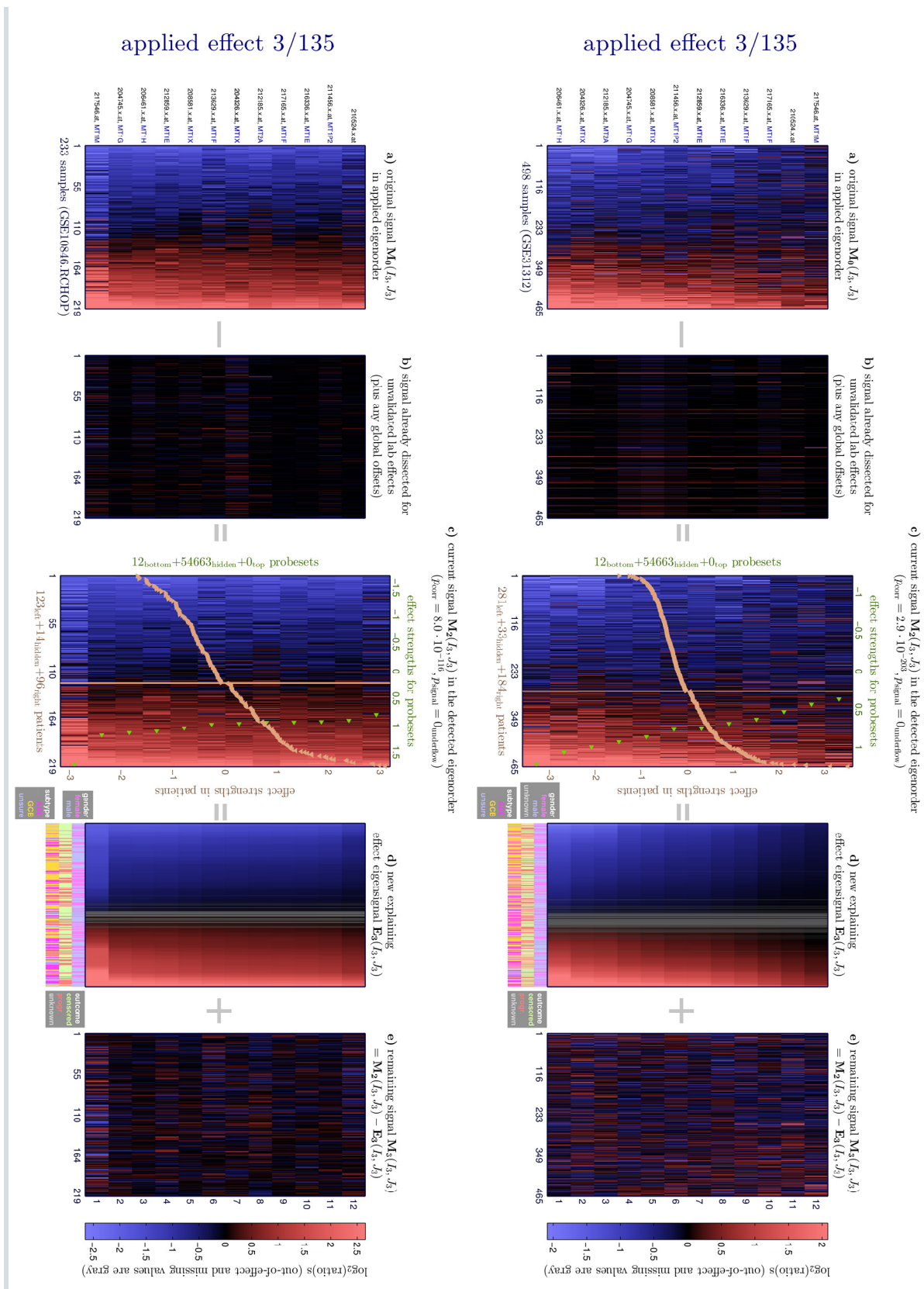


Figure III.3.3.8.b) Validated effect  $v=3$ , applied to GSE10846.R-CHOP and GSE31312.R-CHOP

The fifth and last significant survival effect by multi-cohort based survival analysis<sup>(III.2.5)</sup> is depicted for cohorts GSE10846.R-CHOP (233 patients) and GSE31312.R-CHOP (498 patients).

(The genomic consensus effect is applied to the cleaned signal without lab effects<sup>(III.1.4.2)</sup>. Samples and probesets are ordered by their effect strengths in this cohort<sup>(cf. Table III.1.5)</sup>. Additionally, probesets are filtered by demanding a relative correlation stronger than 0.5. The effect's bimonotonic eigensignal<sup>(panel d)</sup> is grayed for samples having insufficient or insignificant correlation to this effect<sup>(II.4.2.1)</sup>.)

Signature analyses reveal two significantly enriched gene signatures that contain top genes of  $\nu = 3$ ; both are strongly differentially expressed:

| Signatures            |                      |           |            | GSEA             |        |      | Basic Statistics            |                          |                 |              |
|-----------------------|----------------------|-----------|------------|------------------|--------|------|-----------------------------|--------------------------|-----------------|--------------|
| Signatures DB         | Signature Name       | # defined | # measured | Enrichment score | $p$    | FDR  | Mean $\log_2(\text{ratio})$ | $p$ (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| HGNCSigDB_dMay2014    | Metallothioneins     | 19        | 10         | 0.897            | 0.0019 | 0.2% | -1.375                      | 0.0003                   | 100.0%          | 0.0%         |
| GeneSigDB_v4_Sept2011 | Lung_Magda08_21genes | 16        | 13         | 0.610            | 0.0020 | 0.2% | -1.042                      | 0.0004                   | 97.3%           | 2.7%         |

Table III.3.3.8.a) Top-enriched signatures by  $\nu = 3$

Signatures with  $|\text{enrichment score}| \geq 0.5$ , a mean  $|\log_2(\text{ratio})| \geq 0.33$  and at least 10 measured members are listed for genes ranked by GEP effect  $\nu = 3$ . All GSEA  $p$  values are based on permutation tests; hence, they are lower-bounded by  $1/(\#\text{permutations})$  and true  $p$  values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of  $t$ -tests against zero regulation as weights).

Consistent with gene ontology findings, the Metallothioneins signature ([online interpretation card](#)) describes a family of genes that are involved in the metal metabolism of cells<sup>[123]</sup>.

Genes in signature Lung\_Magda08\_21genes ([online interpretation card](#)) were upregulated in human lung cancer cell line A549 after treatment with zinc ionophores<sup>[124]</sup>. As zinc ionophores significantly *inhibited* proliferation for these cells, they were suggested as anti-cancer agents in this context. In contrast, DLBCL patient outcome is inferior for higher expression of this effect.

### ■ Preliminary top genes analysis

The effect contains nine top-correlated genes from four genomic loci and two chromosomes.

All top genes are named MT\*, where MT abbreviates metallothionein. Thionein proteins can bind several metals; the complex is then called metallothionein. By binding physiologically important metals like zinc or by providing a metal storage buffer, they can play a role in cellular functions. A dynamic equilibrium between thionein acceptors and metallothionein donors regulates the availability of zinc<sup>[125]</sup>, for example for protein synthesis.

Seven top genes are from either 16q12.2 or from 16q13. In principle, a reflection of a chromosomal feature in the GEP signal cannot be ruled out. However, given the association of these genes to the same known function, a genetic regulation network seems more likely to be involved in their coordinate expression. Interrogating copy number measurements<sup>(e.g. from GSE11318, [11])</sup> for these loci could potentially provide more clarity.

### ■ Hierarchical survival analysis demarcates a DLBCL subset that is influenced by this effect

As standalone univariate predictor, effect  $\nu = 3$  only explains a weak survival trend with  $p = 0.0601$  on top of age and therapy (likelihood ratio test; trained with all samples). However, its additional explanatory value increased by orders of magnitudes after incorporating effect  $\nu = 19$  into the model ( $p = 1.2 \cdot 10^{-4}$ , likelihood ratio test) and increased further after incorporating  $\nu = 75$ . Hence, it should be biologically interpreted in context of these other GEP effects.

| Top genes in $\nu = 3$ |             |            |             |             |             |  |
|------------------------|-------------|------------|-------------|-------------|-------------|--|
| Probeset               | Gene Symbol | HG19 locus | $a_3^{g,c}$ | $r_3^{g,c}$ | $p_3^{g,c}$ | $\left  \frac{a_3^{g,c}}{r_3^{g,c}} \right $ |
| 206461_x_at            | MT1H        | 16q13      | 1.07        | 0.96        | 8.8E-98     | 1.03   |
| 204745_x_at            | MT1G        | 16q13      | 1.00        | 0.94        | 1.4E-79     | 0.94   |
| 208581_x_at            | MT1X        | 16q13      | 0.97        | 0.93        | 1.5E-76     | 0.91   |
| 204326_x_at            | MT1X        | 16q13      | 0.95        | 0.92        | 1.9E-69     | 0.87   |
| 212859_x_at            | MT1E        | 16q12.2    | 0.92        | 0.90        | 2.4E-62     | 0.83   |
| 211456_x_at            | MT1P2       | 1q43       | 0.87        | 0.92        | 1.9E-69     | 0.80   |
| 217165_x_at            | MT1F        | 16q12.2    | 0.85        | 0.88        | 7.0E-56     | 0.74   |
| 213629_x_at            | MT1F        | 16q12.2    | 0.83        | 0.87        | 4.4E-54     | 0.73   |
| 212185_x_at            | MT2A        | 16q12.2    | 0.81        | 0.89        | 3.9E-58     | 0.72   |
| 217546_at              | MT1M        | 16q12.2    | 0.99        | 0.73        | 3.1E-29     | 0.72   |
| 216336_x_at            | MT1E        | 1p35.1     | 0.73        | 0.87        | 8.6E-55     | 0.64   |
| 210524_x_at            |             | 17q23.3    | 0.66        | 0.81        | 5.0E-40     | 0.53   |
| 216504_s_at            | SLC39A8     | 4q24       | 0.37        | 0.31        | 4.9E-05     | 0.11   |
| 228945_s_at            | SLC39A8     | 4q24       | 0.34        | 0.31        | 2.1E-05     | 0.11   |
| 202437_s_at            | CYP1B1      | 2p22.2     | 0.30        | 0.30        | 5.5E-05     | 0.09   |

Table III.3.3.8.b) Top genes in validated effect  $\nu = 3$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx<sup>TM</sup>[97] v33)

$a_3^{g,c}$  Components of the consensus gene axis of effect  $\nu = 3$  (cf. Table III.1.5); filtered  $\left| \frac{a_3^{g,c}}{r_3^{g,c}} \right| \geq 0.3$ .

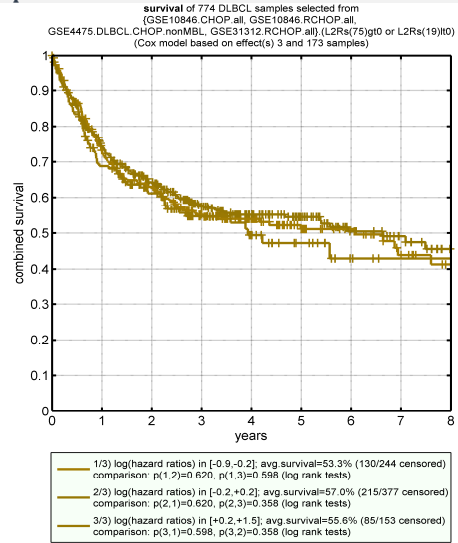
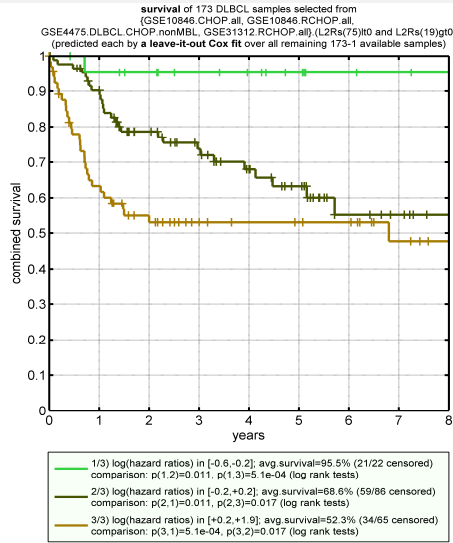
$r_3^{g,c}$  Consensus gene correlations; filtered  $\left| r_3^{g,c} \right| \geq 0.3$ .

$p_3^{g,c}$   $p$  values for the correlations (cf. II.5.2.1)

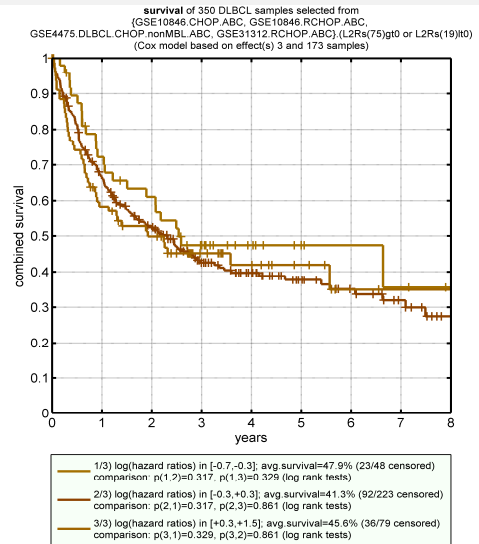
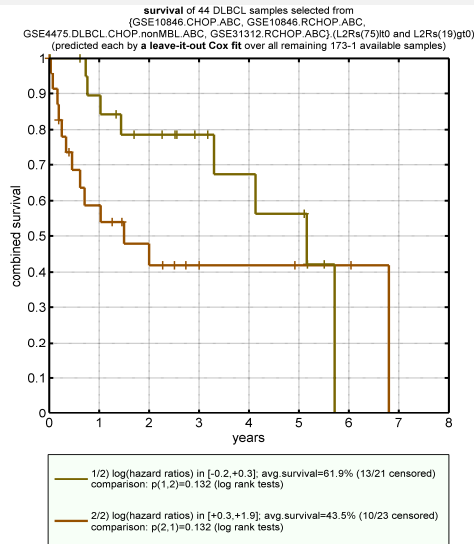
low BCL2 expression and high "T cell co-stimulation"

high BCL2 expression or low "T cell co-stimulation"

all available samples



ABC DLBCL



GCB DLBCL

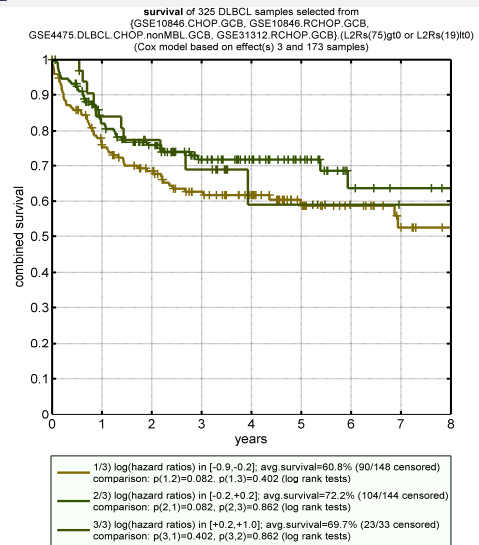
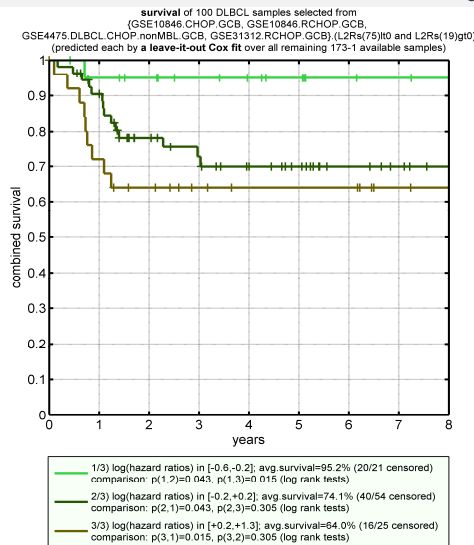


Figure III.3.3.8.c) Survival dependency on effect  $v=3$  for low BCL2 expression and high "T cell co-stimulation".

On the left, 173 samples from all four cohorts with low BCL2 expression (negative  $\log_2$ (ratios) for  $v=75$ ) and high "T cell co-stimulation" (positive  $\log_2$ (ratios) for  $v=19$ ) are split into subgroups of significantly different survival by effect  $v=3$  (hazards predicted by leave-one-out validation). On the right, hazards are predicted with the same predictor variable  $v=3$  (and with  $\beta_3 = 0.37$ ) for samples with high BCL2 expression or low "T cell co-stimulation". Here, no significant survival difference exists for identical risk intervals. (Chosen split points to present the survival spread in three risk intervals equal  $\pm \log$ (hazard ratios of 125%.)

To elucidate this complex survival dependency, further manual combinations of effects and hierarchical survival analyses have been performed. These investigations revealed a DLBCL subset of 173/947 cases (18.3%) defined by *low expression of BCL2* (i.e. the lower risk partition of effect  $\nu = 75$ <sup>(cf. III.3.3.7)</sup>) and by *high expression of the “T cell co-stimulation” effect* (i.e. also the lower risk partition of effect  $\nu = 19$ <sup>(cf. III.3.3.6)</sup>).

For this subset, a simple *univariate* Cox predictor model based only on effect  $\nu = 3$  can predict significant survival differences ( $\hat{\beta}_3 = 0.37, p = 3.2 \cdot 10^{-4}$  using all 173 samples for training), whereas for remaining DLBCL patients (having high expression of BCL2 or low expression of “T cell co-stimulation”) effect  $\nu = 3$  cannot predict any significant survival differences.

To visualize and validate this, leave-one-out validation<sup>(cf. III.2.5.3)</sup> has been applied to this subset of 173 cases, Kaplan-Meier survival estimates for risk intervals have been computed and resulting survival curves have been compared by log rank tests<sup>(Figure III.3.3.8.c)</sup>. The subset for low BCL2 expression and high “T cell co-stimulation” is clearly split into groups of significantly different survival, including a small group of *particularly favorable prognosis for lowest metallothionein expressions (95.5% average survival)*.

This trend is also visible when additionally restricting to ABC DLBCL or to GCB DLBCL. However, remaining sample counts for this three-level hierarchy are too low to reach log rank significance between all neighboring risk intervals. Additionally, no ABC DLBCL samples with very low metallothionein expressions exist (hence, only two survival curves result for identical risk cutoffs).

In contrast, for high BCL2 expression or low “T cell co-stimulation”, expressions of metallothioneins cannot predict any significant survival differences (and neither after restricting by subtype).

#### ■ *The role of zinc for B cells*

Consistent with observed favorable outcome for a subset of DLBCL patients for low levels of metallothioneins, a review on zinc related pathways in immunity<sup>[126]</sup> summarized zinc’s complex involvement in T and B cell activity:

- Zinc deficiency can lead to decreased (non-malignant) lymphocyte count and function.
- Different sensibility to zinc deprivation points to an *effect on cellular development of B cells*, rather than a functional dependency of mature B cells on zinc.
- More precisely, zinc deficiency is assumed to increase the *rate of apoptosis* during elimination of autoreactive<sup>(cf. I.1.2.2)</sup> B cells.
- Consistently, on organism level it has been observed that *loss of lymphoid tissue* during zinc deprivation exceeds that of other tissues.
- Numerous additional zinc-related pathways in context of NF- $\kappa$ B, MAPK, PI3K, NFAT and IRAK are described that can cause zinc to show complex opposing functions, depending on its concentration and on the cellular environment, especially with respect to T cells.

#### ■ *Inference*

In summary, high metallothionein expression is associated with adverse outcome for DLBCL patients with low BCL2 expression<sup>(III.3.3.7)</sup> and high “T cell co-stimulation”<sup>(III.3.3.6)</sup>. Low metallothionein expression in this subset is associated with 95.5% average survival<sup>(cf. Figure III.3.3.8.c)</sup>.

The molecular mechanisms for metallothionein overexpression are currently unclear. The gene ontology result for *cellular response to zinc ion* and published data for a lung cancer experiment using zinc ionophores<sup>[124]</sup> indicate that the expression of this effect may directly correlate with zinc concentrations. In this context, a correlation of patient zinc blood levels with average expressions of this effect could potentially be interesting.



### III.3.4 Effects without Strong Survival Impact

---

As patient outcome is dependent on therapy, effects may be relevant to understand causes of DLBCL, even if their differential expression cannot predict significant survival differences. Ideally, some of these discovered effects might be utilized therapeutically by novel agents in the future. Other effects without survival association might be DLBCL-unspecific.

Three more effects with interesting statistical properties have been selected and are briefly evaluated here.

Analyses for several more effects with significant associations to gene ontology terms or to gene signatures can be browsed via [DLBCL Master Table 2015, main overview.xlsx](#). Additionally, some effects with clearly differential signal between DLBCL patients but without any known associations can be found there.

#### III.3.4.1 $\nu = 20$ : Another perspective on DLBCL subtypes

---

##### ■ Role in survival analysis

Sorting patients by effect  $\nu = 20$  reproduces published DLBCL subtypes with 88% agreement and with  $p = 3.4 \cdot 10^{-97}$  (cf. Figure III.3.2.2.a). Kaplan-Meier survival estimates for sets of patients that result from cutting effect  $\nu = 20$  at its average expression (i.e. at zero eigensignal) are significantly different for both R-CHOP treated cohorts (with  $p_{\text{GSE10846.R-CHOP.KM}} = 7.8 \cdot 10^{-3}$  and  $p_{\text{GSE31312.R-CHOP.KM}} = 0.05$ , log rank tests). This is consistent with known differences in average survival of ABC DLBCL and GCB DLBCL.

However, the effect does not predict any consistent survival *trend* on top of age and therapy. With  $p = 0.065$  (likelihood ratio test) it only ranks 68/135 (cf. Figure III.2.5.1.a) in multivariate analysis. Fitting univariate Cox survival models for both R-CHOP-treated cohorts results in  $p_{\text{GSE10846.R-CHOP.Cox}} = 0.03$  and  $p_{\text{GSE31312.R-CHOP.Cox}} = 0.71$ . For comparison, the same univariate Cox survival analyses for effect  $\nu = 134$  result in  $p_{\nu=134, \text{GSE10846.R-CHOP.Cox}} = 1.1 \cdot 10^{-5}$  and  $p_{\nu=134, \text{GSE31312.R-CHOP.Cox}} = 1.6 \cdot 10^{-7}$ . Cox models test for a consistent survival *trend* over the effect's average expression (quantified by its eigensignal strengths (cf. Table III.1.5)), rather than comparing the average survival of sets of patients.

These results indicate that compatible survival trends over average effect expression *within* subsets of ABC DLBCL or GCB DLBCL are predicted by effect  $\nu = 134$ , but not so by effect  $\nu = 20$ . Visually consistent (cf. Figure III.3.4.1.a), there is no overrepresentation of ABC DLBCL samples with observed progressions (or deaths) on the left and neither an overrepresentation of censored GCB DLBCL samples on the right. Hence, no consistent and strong survival trend exists over sample eigensignal strengths of  $\nu = 20$ , despite the obvious arrangement by subtype.

Similar to effects  $\nu = 127$  and  $\nu = 131$  (cf. III.3.3.2), this effect gains in predictive value after including effect  $\nu = 134$  in the multivariate predictor, indicating that expression of  $\nu = 20$  is associated with two opposing survival trends for two overlapping subsets of DLBCL. However, it still only ranks 9<sup>th</sup> (cf. Figure III.2.5.1.b) with  $p = 2.1 \cdot 10^{-5}$ . After including  $\nu = 131$  or  $\nu = 127$  in the predictor model as well, no significant additional explanatory value remains for  $\nu = 20$ .



### Effect overview

Effect  $\nu = 20$  has been unsupervisedly quad-discovered with detection ranks #6<sup>(GSE10846.CHOP)</sup>, #27<sup>(GSE10846.R-CHOP)</sup>, #4<sup>(GSE4475.nonMBL)</sup> and #15<sup>(GSE31312.R-CHOP)</sup>. The effect has a well-defined signal of moderate size (e.g. Figure III.3.4.1.a) in all four cohorts (approximately 70 top probesets for relative correlation  $\geq 0.5$ ). A detailed list of top genes is available in [DLBCL Master Table 2015, gene orders.xlsx](#).

Its top genes are partially correlated to the hierarchical survival effect  $\nu \in \{127, 131\}$  (cf. III.3.3.2) (with  $r_{(20;131)}^c = 0.57$  and  $r_{(20;127)}^c = 0.31$  (cf. Eqn. III.1.3.2.b)). It is also partially correlated to the relatively large rediscovered COO induced effect (III.3.2.2) (with  $r_{(129;20)}^c = -0.36$ ). This indicates that several of its top genes may be biologically relevant in contexts of these other GEP effects. However, described survival characteristics of effect  $\nu = 20$  indicate that the average expression of its particular composition of top genes is not as specific to true biological effects as these partially correlated other effects.

### Genomic associations

The same ABC-versus-GCB signatures that were significantly enriched for the COO induced effect  $\nu = 129$  (Figure III.3.2.2.c) are nearly as strongly enriched for top genes of  $\nu = 20$  (Figure III.3.4.1.b):

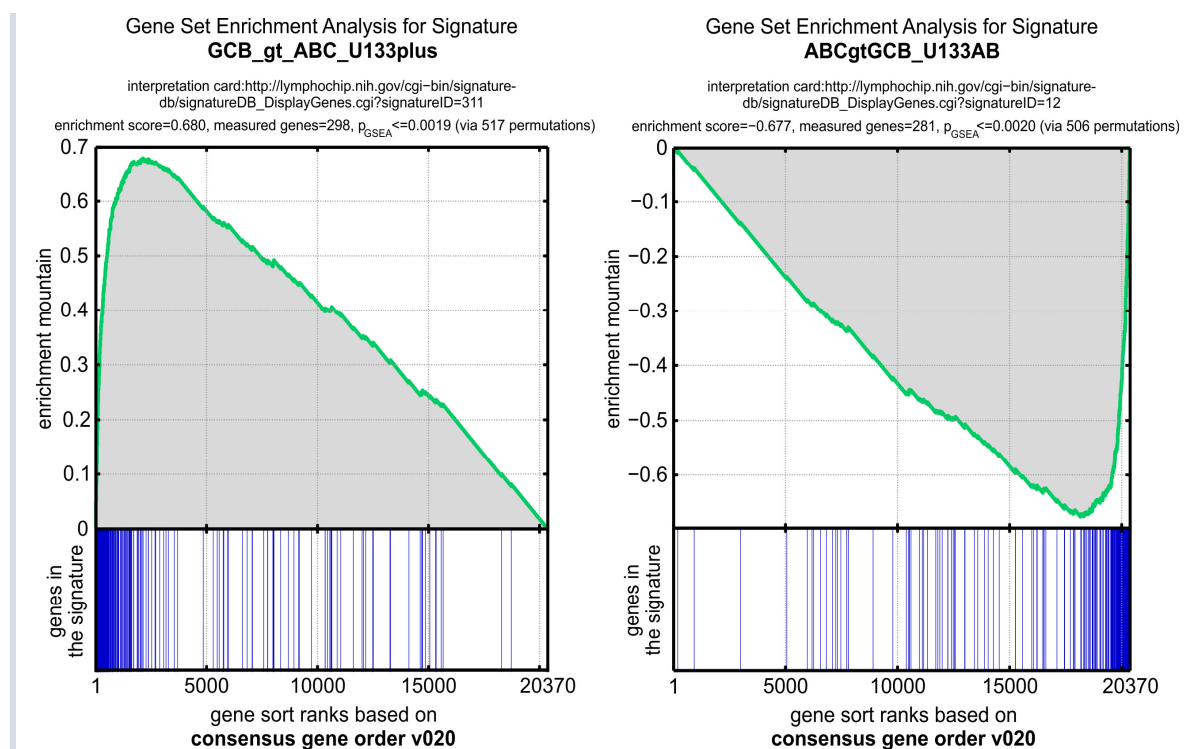


Figure III.3.4.1.b) Significant enrichment of known ABC-versus-GCB DLBCL signatures by effect  $\nu = 20$

### Inference

Results for this effect demonstrate that significant association with and high agreement with binary DLBCL subtypes is not sufficient information to reliably infer *consistent trends* of patient survival over average effect expression. (Significantly different *average* survival for corresponding patient *partitions* may be inferred.)

Hence, depicted signatures ABCgtGCB\_U133AB ([online interpretation card](#)) and GCB\_gt\_ABC\_U133plus ([online interpretation card](#)) may contain genes that are not necessarily associated with a consistent survival trend, as these genes were identified based on *previously assigned subtype classes* (cf. page 180 for details),

This is confirmed by the enrichment of these signatures for both the COO induced effect  $v = 129$  (cf. III.3.2.2) (that is associated with a strong survival trend with  $p = 1.1 \cdot 10^{-12}$ ) and likewise enrichment for the current effect  $v = 20$ , although it only has  $p = 0.065$  for the same test (cf. Figure III.2.5.1.a).

Consequently and in general, significant enrichment of the same signatures for two different effects does not necessarily indicate that these effects represent biologically similar functions. Hence, results from signature enrichment analyses should be interpreted with care. In particular, the biological specificity of enriched signatures should be independently verified in the respective context, if possible.

Furthermore, to quantify whether a small gene signature is associated with a particular effect, individual correlations of signature genes to the effect's sample axis might provide a biologically more specific answer. These correlations may complement enrichment or overrepresentation analyses on signature level. (However, they can only be computed in context of a concrete patient cohort and are not available for pure genomic association analyses.)

### III.3.4.2 $v = 7$ : Presumably the main blood concentration effect

---

#### ■ Effect overview

This one-sided effect (Figure III.3.4.2) consists of approximately 800 top probesets (relative correlation  $\geq 0.5$ ) and is one of the largest discovered effects. It has been unsupervisedly quad-discovered with detection ranks #2<sup>(GSE10846.CHOP)</sup>, #2<sup>(GSE10846.R-CHOP)</sup>, #3<sup>(GSE4475.nonMBL)</sup> and #1<sup>(GSE31312.R-CHOP)</sup>.

It is not associated with patient outcome following (R-)CHOP therapy.

#### ■ Effect identification

Due to the effect's size, many gene ontology terms are overrepresented and many signatures are enriched, including many signatures from specific leukemia and immune contexts. All of them are of much smaller size and hence *are not representative for the full effect*. Furthermore, they describe different biology. There should be a *common explanation* for such high correlations between so many genes.

While unsupervised quad-discovery rules lab-specific technical effects out, it may be speculated that the effect represents *differences in frequency of another cell type* from the microenvironment that has been inadvertently included in measured tumor samples. This could affect all four patient cohorts and could explain the large number of correlated genes, as simply all genes expressed in this other cell type would be ordered by relative frequencies of this cell type in measured samples. However, this is just a hypothesis that seems plausible because of the large size of this effect that I have only seen between different cell types so far. Instead of a single cell type, it might also be an indirect measurement of a common mixture of cell types.



Screening significant signatures with this hypothesis in mind revealed that *several different blood modules are significantly enriched*:

| Signatures           |                                       |           |            | GSEA             |        |      | Basic Statistics              |                          |                 |              |
|----------------------|---------------------------------------|-----------|------------|------------------|--------|------|-------------------------------|--------------------------|-----------------|--------------|
| Signatures DB        | Signature Name                        | # defined | # measured | Enrichment score | $p$    | FDR  | Mean log <sub>2</sub> (ratio) | $p$ (t-test versus zero) | Down-reg. ratio | Upreg. ratio |
| StaudtSigDB_dNov2012 | Dendritic_cell_CD16pos_blood          | 42        | 41         | 0.659            | 0.0020 | 0.2% | -0.635                        | 1.7E-09                  | 91.9%           | 8.1%         |
| StaudtSigDB_dNov2012 | Blood_Module-1.5_Myeloid_lineage-1    | 110       | 108        | 0.583            | 0.0020 | 0.2% | -0.611                        | 6.3E-25                  | 92.3%           | 7.7%         |
| StaudtSigDB_dNov2012 | Blood_Module-3.1_Interferon_inducible | 94        | 93         | 0.694            | 0.0020 | 0.2% | -0.526                        | 5.4E-18                  | 91.8%           | 8.2%         |
| StaudtSigDB_dNov2012 | Blood_Module-2.1_Cytotoxic_cells      | 160       | 155        | 0.586            | 0.0021 | 0.2% | -0.551                        | 1.5E-25                  | 87.1%           | 12.9%        |
| StaudtSigDB_dNov2012 | Blood_Module-2.6_Myeloid_lineage-2    | 145       | 143        | 0.543            | 0.0020 | 0.2% | -0.461                        | 5.3E-22                  | 87.8%           | 12.2%        |
| StaudtSigDB_dNov2012 | Blood_Module-1.3_B_cells              | 55        | 53         | -0.577           | 0.0020 | 0.2% | 0.496                         | 2.0E-11                  | 8.8%            | 91.2%        |

Table III.3.4.2) Blood module signatures that are significantly enriched for  $\nu = 7$

All GSEA  $p$  values are based on permutation tests; hence, they are lower-bounded by  $1/(\# \text{permutations})$  and true  $p$  values might be considerably lower in this case. 1000 permutations have been computed for each signature, i.e. approximately 500 for each enrichment sign. Percentages of down- and upregulated genes in a signature are weighted averages of gene regulation signs (using  $1 - p$  values of  $t$ -tests against zero regulation as weights).

This might indicate that this effect is just *an indirect measurement of blood concentrations* in tumor samples and that these concentration differences cause the discovered broad correlation effect. Only moderate enrichment of these signatures could be caused by slightly changing mixture ratios of different blood cells from patient to patient.

Interestingly, *Blood\_Module-1.3\_B\_cells*<sup>(cf. online interpretation card,<sup>[127]</sup>)</sup> is anti-correlated to all other enriched blood modules, indicating that the effect does not only measure different levels of blood concentration, but a relative concentration of B cells and other blood cells.

### Inference

Assuming that this effect indeed measures relative concentrations of B cells to other blood modules, it may be interesting to find out whether these concentration ratios have already been introduced by tumor sampling. If this can be excluded, the effect might allow indirect insights on how different forms of DLBCL influence their tumor microenvironment.

## III.3.4.3

### $\nu = 4$ : A strong immunoglobulin effect

#### Effect overview

Effect  $\nu = 4$  shows a strongly differential one-sided signal (Figure III.3.4.3, p222) and was also unsupervisedly discovered in all four DLBCL cohorts (with detection ranks #9<sup>(GSE10846.CHOP)</sup>, #9<sup>(GSE10846.R-CHOP)</sup>, #12<sup>(GSE4475.nonMBL)</sup> and #14<sup>(GSE31312.R-CHOP)</sup>). It is not associated with patient outcome following (R-)CHOP therapy.

#### Top genes overview

Interestingly, most of this effect's top-correlated genes originate from only three loci: 2p11.2, 14q32.33 and 22q11.22<sup>(cf. Table III.3.4.3.a)</sup>. Such overrepresented and specific loci might indicate a reflection of chromosomal features. To test for potential triggering aberrations, copy number measurements<sup>(e.g. from GSE11318, [11])</sup> for these loci could be interrogated.

| Top genes in $\nu=4$ |              |            |             |             |             |  |
|----------------------|--------------|------------|-------------|-------------|-------------|--|
| Probeset             | Gene Symbol  | HG19 locus | $a_4^{g,c}$ | $r_4^{g,c}$ | $p_4^{g,c}$ | $ a_4^{g,c} $<br>$\cdot \frac{1}{r_4^{g,c}}$ |
| 215176_x_at          |              | 2p11.2     | 2.21        | 0.95        | 1.6E-99     | 2.10   |
| 211645_x_at          |              | 2p11.2     | 2.18        | 0.96        | 1.4E-109    | 2.09   |
| 216576_x_at          | IGKC         | 2p11.2     | 2.04        | 0.95        | 8.5E-98     | 1.93   |
| 234764_x_at          | IGLC1        | 22q11.22   | 2.11        | 0.80        | 8.9E-50     | 1.68   |
| 216401_x_at          |              | 2p11.2     | 1.66        | 0.94        | 1.9E-92     | 1.56   |
| 216207_x_at          | IGKC         | 2p11.2     | 1.57        | 0.94        | 3.0E-93     | 1.48   |
| 217157_x_at          | IGKC         | 2p11.2     | 1.60        | 0.92        | 2.2E-80     | 1.47   |
| 217378_x_at          | LOC100130100 | 2q13       | 1.55        | 0.93        | 1.2E-85     | 1.44   |
| 216510_x_at          | IGHA1        | 14q32.33   | 1.65        | 0.79        | 1.5E-42     | 1.30   |
| 217148_x_at          | IGLC1        | 22q11.22   | 1.81        | 0.69        | 8.1E-29     | 1.25   |
| 211644_x_at          | IGKC         | 2p11.2     | 1.73        | 0.72        | 5.2E-32     | 1.24   |

However, as most top genes from all three loci are coding immunoglobulins, their consistent regulation might also be associated with a common function or cell type.

### Genomic associations

The gene ontology term for antigen binding<sup>(cf. GO:0003823)</sup> is significantly overrepresented with  $p < 2.2 \cdot 10^{-16}$  (7 of the effect's top 17 genes belong to this term, while only 25 genes of all 20370 measured genes belong to it; hypergeometric test).

Signature analyses revealed several significantly yet moderately enriched signatures, including several immunoglobulin gene families and `Blood_Module-1.1_Plasma_cells`<sup>(cf. online interpretation card.[127])</sup> (enrichment score 0.591,  $p \leq 0.0021$ , 486 permutations).

### Inference

Signature enrichment suggests that correlated expressions of genes in this effect might reflect concentrations of blood plasma cells that might produce and secrete soluble antibodies (immunoglobulins).

In this case and assuming that no bias with respect to plasma cell concentrations has been introduced by tumor sampling,  $\nu = 4$  might measure an indirect effect of different forms of DLBCL on the micro-environment. (Biases due to the way of sampling are unlikely here, because in all four independent cohorts approximately one third of patients overexpresses this effect relative to baseline and approximately two thirds have lower expression than baseline<sup>(cf. Figure III.3.4.3)</sup>. In case of random fluctuations of plasma cell concentrations by tumor sampling, a baseline that is near the median would be more likely.)

However, the link to plasma cells by enrichment requires further biological validation, because only approximately 25% of this signature's genes are strongly differentially expressed by  $\nu = 4$ . Hence, also a more direct role of this effect in DLBCL cannot be excluded.

|             |              |            |      |      |         |      |
|-------------|--------------|------------|------|------|---------|------|
| 214768_x_at | IGKC         | 2p11.2     | 1.57 | 0.79 | 2.2E-43 | 1.24 |
| 211430_s_at | IGHG1        | 14q32.33   | 1.72 | 0.71 | 1.6E-31 | 1.23 |
| 217480_x_at | LOC100287723 | (multiple) | 1.35 | 0.91 | 3.1E-76 | 1.23 |
| 216557_x_at | IGHA1        | 14q32.33   | 1.43 | 0.77 | 2.9E-39 | 1.10 |
| 217281_x_at | IGH@         | 14q32.33   | 1.40 | 0.77 | 1.2E-39 | 1.08 |
| 216984_x_at | IGLC1        | 22q11.22   | 1.52 | 0.70 | 4.4E-30 | 1.07 |
| 211643_x_at | IGKC         | 2p11.2     | 1.40 | 0.76 | 3.1E-38 | 1.06 |
| 217022_s_at | IGH@         | 14q32.33   | 1.55 | 0.64 | 5.1E-24 | 1.00 |
| 211650_x_at | IGH@         | 14q32.33   | 1.28 | 0.78 | 1.1E-40 | 0.99 |
| 211798_x_at | IGLJ3        | 22q11.22   | 1.33 | 0.74 | 4.1E-35 | 0.98 |
| 214973_x_at | IGHD         | 14q32.33   | 1.35 | 0.73 | 2.8E-33 | 0.98 |
| 214777_at   |              | 2p11.2     | 1.52 | 0.64 | 5.2E-24 | 0.97 |
| 216491_x_at | IGHM         | 14q32.33   | 1.40 | 0.67 | 3.4E-27 | 0.94 |
| 217258_x_at | IGLV1-44     | 22q11.22   | 1.27 | 0.71 | 9.9E-32 | 0.91 |
| 224342_x_at | LOC96610     | 22q11.22   | 1.23 | 0.72 | 9.7E-36 | 0.88 |
| 217227_x_at | IGLV1-44     | 22q11.22   | 1.20 | 0.72 | 4.7E-33 | 0.87 |
| 211637_x_at | IGH@         | 14q32.33   | 1.23 | 0.66 | 2.7E-26 | 0.82 |
| 217179_x_at |              | 22q11.22   | 1.24 | 0.66 | 2.5E-25 | 0.81 |
| 214669_x_at | IGKC         | 2p11.2     | 1.26 | 0.63 | 9.2E-23 | 0.79 |
| 217235_x_at | IGLL5        | 22q11.22   | 1.20 | 0.65 | 5.6E-25 | 0.78 |
| 224795_x_at | IGKC         | 15q21.3    | 1.31 | 0.59 | 4.5E-22 | 0.78 |
| 211881_x_at | IGLJ3        | 22q11.22   | 1.04 | 0.73 | 3.2E-34 | 0.76 |
| 221651_x_at | IGKC         | 2p11.2     | 1.30 | 0.58 | 4.5E-19 | 0.76 |
| 221671_x_at | IGKC         | 2p11.2     | 1.28 | 0.58 | 4.9E-19 | 0.74 |
| 214836_x_at | IGKC         | 2p11.2     | 1.17 | 0.63 | 3.2E-23 | 0.74 |
| 211868_x_at | IGH@         | 14q32.33   | 1.05 | 0.68 | 2.1E-28 | 0.72 |
| 211641_x_at | IGHA1        | 14q32.33   | 0.93 | 0.75 | 2.9E-37 | 0.71 |
| 215214_at   | IGLC1        | 22q11.22   | 1.00 | 0.68 | 1.0E-27 | 0.68 |
| 209138_x_at | IGLC1        | 22q11.22   | 1.28 | 0.52 | 6.4E-15 | 0.67 |
| 211908_x_at | IGK@         | 14q32.33   | 0.92 | 0.72 | 4.1E-33 | 0.67 |
| 234884_x_at | IGLC1        | 22q11.22   | 1.01 | 0.65 | 4.7E-28 | 0.66 |
| 214677_x_at | IGLC1        | 22q11.22   | 1.30 | 0.49 | 2.5E-13 | 0.64 |
| 211634_x_at | IGHM         | 14q32.33   | 1.00 | 0.61 | 1.8E-21 | 0.61 |
| 216542_x_at | IGHA1        | 14q32.33   | 0.83 | 0.73 | 7.5E-34 | 0.61 |
| 216560_x_at | IGLC1        | 22q11.22   | 1.05 | 0.57 | 3.4E-18 | 0.60 |
| 211639_x_at | IGH@         | 14q32.33   | 0.98 | 0.58 | 2.6E-19 | 0.57 |
| 215121_x_at | IGLC1        | 22q11.22   | 1.14 | 0.49 | 2.7E-13 | 0.56 |
| 234366_x_at | IGLC1        | 22q11.22   | 0.92 | 0.61 | 2.1E-23 | 0.56 |
| 214916_x_at | IGHA1        | 14q32.33   | 0.92 | 0.61 | 5.2E-21 | 0.56 |
| 216829_at   | IGKC         | 2p11.2     | 0.75 | 0.71 | 4.0E-31 | 0.54 |
| 215379_x_at | IGLV1-44     | 22q11.22   | 1.06 | 0.50 | 1.5E-13 | 0.52 |
| 216853_x_at | IGLC1        | 22q11.22   | 0.89 | 0.56 | 2.1E-17 | 0.50 |
| 217360_x_at | IGHA1        | 14q32.33   | 0.69 | 0.70 | 3.0E-30 | 0.49 |
| 215949_x_at | IGHM         | 14q32.33   | 0.73 | 0.66 | 6.0E-26 | 0.49 |
| 211635_x_at | IGHA1        | 14q32.33   | 0.83 | 0.58 | 3.2E-19 | 0.49 |
| 211640_x_at | IGHG1        | 14q32.33   | 0.78 | 0.62 | 2.3E-22 | 0.48 |
| 212592_at   | IGJ          | 4q13.3     | 1.18 | 0.41 | 3.5E-09 | 0.48 |
| 215946_x_at | IGLL3P       | 22q11.23   | 0.89 | 0.50 | 6.4E-14 | 0.45 |
| 216412_x_at | IGLC1        | 22q11.22   | 0.71 | 0.60 | 9.2E-21 | 0.43 |
| 211633_x_at |              | 14q32.33   | 0.70 | 0.59 | 5.5E-20 | 0.42 |
| 237625_s_at |              | 2p11.2     | 0.94 | 0.43 | 2.2E-11 | 0.41 |
| 213502_x_at | GUSBP11      | 22q11.23   | 0.84 | 0.48 | 1.7E-12 | 0.40 |
| 216430_x_at | IGLV1-44     | 22q11.22   | 0.62 | 0.61 | 4.9E-21 | 0.38 |
| 217384_x_at | IGHV3-48     | 14q32.33   | 0.56 | 0.66 | 5.7E-26 | 0.37 |
| 216365_x_at | IGLC1        | 22q11.22   | 0.72 | 0.51 | 3.1E-14 | 0.37 |
| 234851_at   | IGLC1        | 22q11.22   | 0.60 | 0.58 | 2.7E-21 | 0.35 |
| 234792_x_at | IGHA1        | 14q32.33   | 0.57 | 0.58 | 6.7E-21 | 0.33 |
| 216541_x_at | IGHG1        | 14q32.33   | 0.60 | 0.52 | 4.0E-15 | 0.32 |
| 235965_at   |              | 7q21.3     | 0.64 | 0.47 | 3.3E-13 | 0.30 |
| 234707_x_at | IGLV1-44     | 22q11.22   | 0.68 | 0.43 | 4.2E-11 | 0.29 |
| 217236_x_at | IGH@         | 14q32.33   | 0.47 | 0.59 | 1.3E-19 | 0.28 |
| 217239_x_at | LOC100508797 | 14q32.33   | 0.48 | 0.57 | 3.2E-18 | 0.27 |
| 211647_x_at | IGHG1        | 14q32.33   | 0.48 | 0.56 | 2.9E-17 | 0.27 |
| 211649_x_at | IGHA1        | 14q32.33   | 0.46 | 0.53 | 2.1E-15 | 0.24 |
| 216708_x_at | CKAP2        | 22q11.22   | 0.49 | 0.47 | 2.6E-12 | 0.23 |
| 200670_at   | XBP1         | 22q12.1    | 0.54 | 0.42 | 7.3E-10 | 0.23 |
| 217145_at   | IGKC         | 2p11.2     | 0.42 | 0.51 | 4.7E-14 | 0.21 |
| 216517_at   | IGKC         | 2p11.2     | 0.45 | 0.44 | 8.9E-11 | 0.20 |
| 229721_x_at | DERL3        | 22q11.23   | 0.45 | 0.44 | 1.2E-11 | 0.20 |
| 201287_s_at | SDC1         | 2p24.1     | 0.43 | 0.45 | 5.4E-11 | 0.19 |

Table III.3.4.3.a) Top genes in validated effect  $\nu = 4$

(probesets) from Affymetrix U133 Plus 2.0 microarrays; manufacturer annotations (NetAffx<sup>TM</sup>[97] v33)  
 $a_i^{g,c}$  Components of the consensus gene axis of effect  $\nu = 4$  (cf. Table III.1.5); filtered  $|a_i^{g,c}| \geq 0.4$ .  
 $r_i^{g,c}$  Consensus gene correlations; filtered  $|r_i^{g,c}| \geq 0.4$ .  
 $p_i^{g,c}$  p values for the correlations (cf. II.5.2.1)





# Conclusion

---

*A novel method for signal dissection into interpretable patterns has been designed, developed and successfully validated against synthetic and real-world data. Its search strategy for interactions is based on an extremum principle for correlations. Its bimonotonic effect model allows the regression of a broad class of nonlinear gene regulations. With its capability to dissect even partially correlated effects precisely, it goes conceptually beyond standard methods like principle components analysis and hierarchical clustering.*

*All known major GEP effects for DLBCL that are significantly associated with patient survival have been rediscovered. Additionally, novel genetic effects with greater predictive power have been discovered. They can predict significant survival differences within known disease subtypes and within clinical risk classes by international prognostic index. Comprehensive biostatistical evaluations for discovered survival effects reveal hierarchical dependencies and pinpoint molecular heterogeneities. Effect correlations identify potential oncogenes or tumor suppressor genes. Together, these results may help to clarify the molecular pathogenesis of DLBCL.*

*Signal dissection can be readily applied to other cancer entities as well. Moreover, its concept of interaction may have the potential to lead to more interpretable insights into signals from many other fields of science, for instance, into spectral energy distributions of stars.*

## ■ Key concepts and scope of application

Mathematically, signal dissection is applicable to any high-dimensional multi-sample signal  $\in \mathbb{R}^{m \times n}$  (cf. II.2.1.1) consisting of  $n$  samples (e.g. tumor biopsies) of a system with  $m$  dimensions (e.g. genes). However, for resulting effects to be interpretable, its concept of interaction and detection must be compatible with the analyzed system. There are three key concepts driving signal dissection that determine its scope of application and distinguish it from principal components analysis (PCA).

**Correlation maximization principle:** Initially, the search strategy<sup>(II.3)</sup> detects effect axes similar to principal components, but it utilizes a different generic concept of interactions for detection<sup>(cf. I.1.3.4)</sup>. Rather than looking for maximal signal variance and minimal cross-effect covariance as done by PCA<sup>(cf. I.2.2.2)</sup>, it *maximizes the within-effect correlation*<sup>(II.3.1.6)</sup>. Searching for maximal correlation instead of maximal variance optimizes specificity of resulting effects. Generally, for signal dissection to be applicable, it must make sense to ask for *non-local* correlations between arbitrary system dimensions and between arbitrary samples. For genes, this is the case, as their extrinsic order by genomic sequence does not prevent non-neighboring genes from being coordinately expressed by a shared pathway that is active in measured cells. Discovered effect axes then summarize laws of gene regulation mediated by such pathways (or by other causes) as linear combinations of genes.

**Generic bimonotonic effect model:** A bimonotonic effect model<sup>(II.2.1.2)</sup> and a corresponding bimonotonic regression algorithm<sup>(II.4.1)</sup> is utilized to empirically estimate an particular effect's own contributions to the measured signal sum of superposed effects<sup>(II.2.1.1)</sup>. The resulting effect eigensignal can be parameterized as either the effect's gene curve in gene space or the effect's sample curve in sample space<sup>(II.2.2.3)</sup>. Effect curves *extend the linear concept* of gene and sample axes. This also extends the method's applicability to a broad class of nonlinear effects that are monotonic with respect to projections on effect axes, e.g. biological activation

thresholds or saturations<sup>(II.1.1.2)</sup>. In contrast, usual projections along gene axes can only explain the linear component of an effect's law of gene regulation. Hence, a naïve dissection by projection would split the effect by leaving its nonlinear parts in the signal. This would trigger later discoveries of hard-to-interpret secondary effects.

**Effect focusing and precise dissection:** Weighted uncentered correlations<sup>(cf. II.2.3.1)</sup> to converged gene and sample axes and their statistical significance<sup>(II.5.2.1)</sup> define the final focus of a discovered effect, i.e. its participating genes and affected samples. Utilizing this focus as dissection strengths and together with regressed effect curves, this allows the precise and exclusive dissection of the effect<sup>(II.4.2)</sup>. Signal parts from other potentially overlapping effects (that may regulate the same genes in the same samples by other laws) are left untouched for their later separate discovery<sup>(e.g. Figure II.4.2.2.b)</sup>. This makes the method even applicable in context of *partially correlated effects*, for instance to all four effects in the 3D concept example<sup>(II.6.1)</sup>. In contrast, the dissection of such effects is not possible with PCA or with other methods that are equivalent to projections or to orthogonal coordinate transformations of the gene or sample space. (This is conceptually impossible for these methods, because after three full projections in 3D only a point signal remains, but there are four distinct effect axes in this signal<sup>(cf. II.6.1)</sup>.)

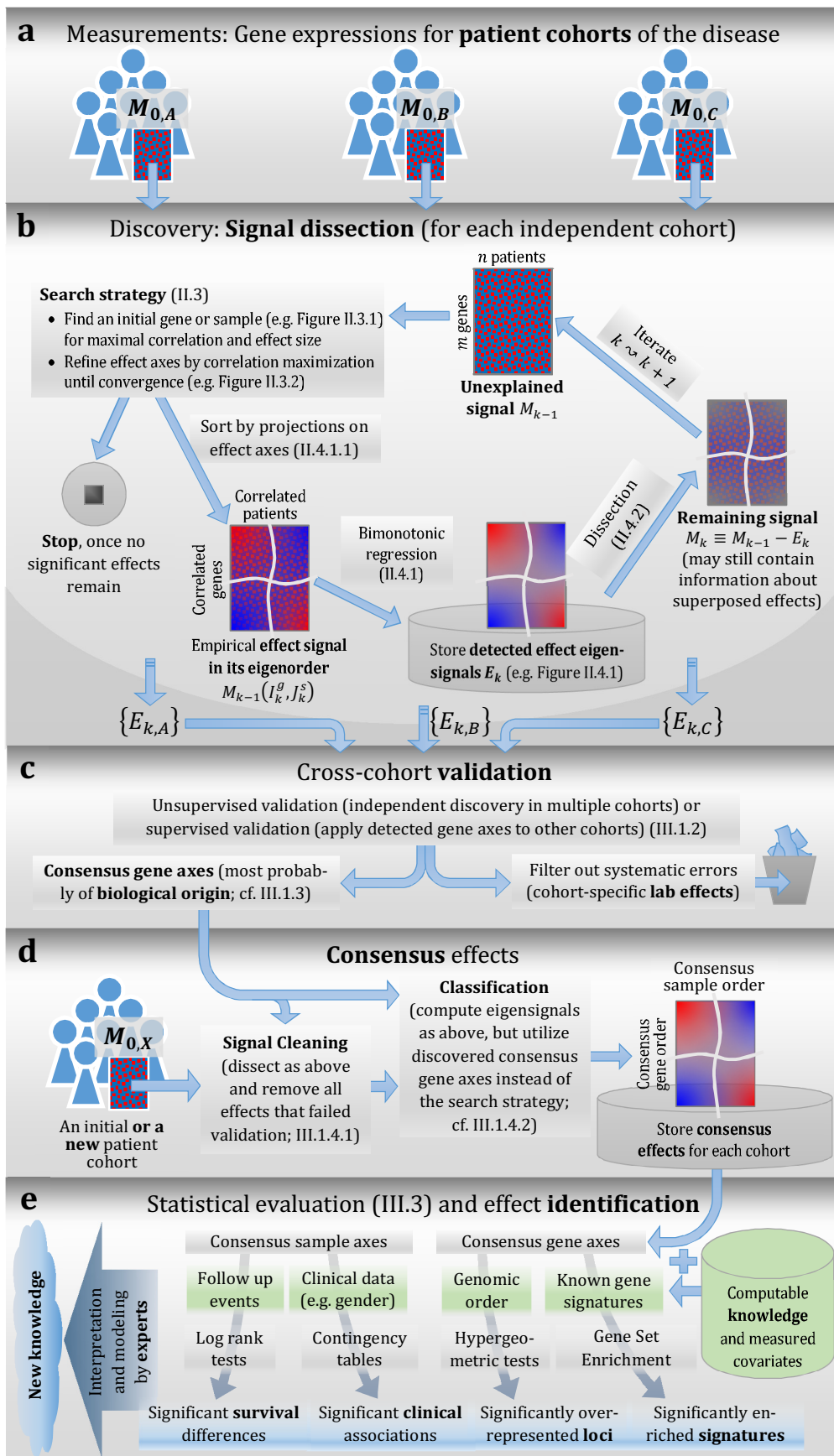
Signal dissection utilizes these concepts iteratively to detect, extract, quantify and summarize distinct laws of gene regulation by effect curves. Other than traditional techniques like hierarchical clustering that can just reorganize genes and samples into groups once, dissection modifies the signal itself by “peeling off” correlated and potentially overlapped signal parts. The sum of all detected and dissected effects *reconstructs the complete signal*, except for noise<sup>(cf. II.2.1.1)</sup>. If needed, a traditional clustering can be readily derived from *each* effect by choosing a cutoff. Hence, signal dissection effectively also realizes and describes a set of alternate clusterings of genes and samples by shared gene regulation effects.

### ■ Solution overview

As introduced<sup>(I.1)</sup>, the practical aim is to bridge the increasing gap between overwhelmingly detailed signals based on modern measurement technology and expert modeling of underlying (and typically complex) systems. To this end, signal dissection contributes an interpretable summarization of measured signals. More precisely, it contributes a superposition<sup>(II.2.1.1)</sup> of specific effects of interaction that are observable by correlations<sup>(II.2.1.2)</sup>. However, a complete solution for this aim needs some additional steps.

For validation purposes, two or more independent sample sets have to be measured for the analyzed system; this is depicted by three exemplary patient cohorts in the solution scheme<sup>(unterhalb)</sup>. First, the signal for each cohort is dissected independently. Important steps of signal dissection are illustrated again. Next, resulting sets of effects are validated across cohorts; this enables filtering out systematic errors like lab-specific effects. Hence, validated effects most probably originate from true interactions in the analyzed system. Finally and as optimal preparation for expert interpretation and modeling, several statistical methods are utilized to associate validated effects with available sources of computable knowledge.

This solution has been applied to more than 1000 tumor samples from DLBCL patients<sup>(III.1.1.1)</sup>. All gene expression effects that can distinguish patients of DLBCL have been unsupervisedly discovered and validated<sup>(III.1.2)</sup> across four independent patient cohorts. As signal dissection is complete<sup>(II.2.1.1)</sup>, there are no significant<sup>(II.5)</sup> GEP correlation effects left in remaining signals. Several validated effects are significantly associated with patient outcome. Genetically novel survival effects are summarized below. To the best of my knowledge, this is the most comprehensive unsupervised gene expression analysis of DLBCL to date.



Solution scheme (exemplary for gene expressions in bioscience)

(a) Gene expressions have been measured for independent patient cohorts. (b) Signal dissection and its exploratory search strategy detect effects. (c) Discovered effects are validated across independent cohorts. (d) Applying validated consensus gene effects classifies samples. (e) Genomic associations to consensus gene axes and clinical associations to consensus sample axes provide the basis for expert assessment and interpretation of effects.

### ■ *Method validation and comparison*

Several synthetic data scenarios have been designed to validate signal dissection thoroughly: A versatility test<sup>(II.6.2)</sup> has simulated overlapping effects of various size, shape and signal strength that mimic known biological or technical real world effects. A superposition scenario<sup>(II.6.3)</sup> has tested the maximum dissectible amount of overlapping effects. Several tests have been designed to test diverse detection limits<sup>(II.6.4)</sup>: the limit of weak signals relative to simulated noise, the limit of acceptable missing values and the limit of an acceptable number of noise genes for the detection of small effects.

A systematic comparison with PCA has proven that signal dissection leads to more interpretable results, i.e. the method is able to rediscover more simulated effects with significantly higher correlations to true simulated effect axes<sup>(cf. II.6.2.5)</sup>. In particular, signal dissection has clear advantages for overlapping effects of similar size, which are common for real world gene expression signals. Here, PCA's interaction concept of maximal variance guides to genes that are expressed by the highest number of overlapping effects. Consequently, resulting principal components represent uninterpretable linear combinations of overlapping yet distinct effects, rather than to dissect the signal into true simulated effects<sup>(cf. Figure II.6.3.1.a)</sup>.

Furthermore, signal dissection can still detect effects for high ratios of missing values<sup>(e.g. Figure II.6.4.4.b)</sup>. To a certain extent, it can even restore missing signals from correlations<sup>(cf. Figure II.6.4.4.d)</sup>. (PCA does not support signals with missing values.)

### ■ *Real-world validation and rediscovered biological effects*

Signal dissection has also been successfully validated against real-world data. For instance, the gender GEP effect has been unsupervisedly discovered in all four dissected DLBCL patient cohorts<sup>(III.3.2.1)</sup>. This is already an independent control of detection, validation, annotation and interpretation pipelines.

Due to the completeness of signal dissection<sup>(II.2.1.1)</sup>, additionally all major previously discovered DLBCL GEP effects have been unsupervisedly rediscovered. All are known for their association with patient outcome. The biologically most important one is the cell-of-origin (COO) induced effect. It identifies two patient subgroups as molecularly distinct DLBCL subtypes<sup>(III.3.2.2, index v=129)</sup>. These subtypes are associated with significantly different survival and are hard to distinguish on morphological level. Hence, this molecular effect is possibly needed for precise therapy decisions, when subtype-specific agents become available. Secondly, a stromal effect has been rediscovered that is associated with the extracellular matrix<sup>(III.3.3.4, index v=5)</sup>. Thirdly, another stromal effect has been rediscovered that has previously been associated with the tumor blood-vessel density<sup>(III.3.3.5, index v=44)</sup>.

### ■ *Genetically novel effects and survival prediction*

Several more genetically distinct GEP effects have been discovered that are significantly associated with survival<sup>(III.2.5.1)</sup>. In particular, one genetically novel effect<sup>(III.3.3.1, index v=134)</sup> can predict observed survival even more consistent<sup>(Figure III.2.5.1.a)</sup> than the COO induced effect.

Via iterative selection of GEP effects that provide the highest additional explanatory value for observed patient outcome, a quivariate Cox survival predictor has been constructed<sup>(III.2.5)</sup>. Based on leave-one-out validation, predicted survival probabilities show a homogeneous predictor performance from 29% to 89% average survival<sup>(cf. Figure III.2.5.3.b)</sup>. Comparison with survival dependencies that are predicted by known DLBCL subtypes<sup>(III.2.1.8)</sup> shows an obvious and strong increase in the predicted survival spread, both for the former

standard CHOP and the current standard R-CHOP chemotherapy<sup>(III.2.5.4)</sup>. Furthermore, it can predict significant survival differences *within* known DLBCL subtypes<sup>(III.2.5.6)</sup> and *within* risk classes by international prognostic index (IPI)<sup>(III.2.5.8)</sup>. This suggests that discovered GEP effects register so far unknown DLBCL biology and identify relevant genetic heterogeneity. Additionally, this demonstrates that macroscopic clinical observables underlying the IPI score presumably can no longer serve as surrogates for molecular predictors, as soon as different therapies for specific molecular subtypes become clinically available. The latter is a concrete goal towards precision medicine<sup>(I.1.2.4)</sup>.

### ■ *Biostatistical evaluation of major novel survival effects*

It is not yet clear which precise molecular mechanisms cause these survival differences. Towards understanding them, all discovered major survival effects<sup>(III.3.3)</sup> have been biostatistically evaluated. Genomic associations of effects with existing knowledge, clinical associations with patient covariates and in particular hierarchical survival dependencies between effects have been analyzed. These results and suggested biological validation experiments may help to advance the investigation of DLBCL's molecular pathogenesis. Selected results and inferred biological hypotheses are summarized below.

Effect  $\nu = 134$ <sup>(III.3.3.1)</sup> is the primary survival factor in DLBCL, as it can predict survival most consistently with  $p = 4.5 \cdot 10^{-17}$ . The COO induced effect<sup>(cf. III.3.2.2)</sup> follows at rank #2 with  $p = 1.1 \cdot 10^{-12}$ <sup>(Figure III.2.5.1.a)</sup>. Top genes of these two effects are only weakly correlated ( $r_{(134;129)}^c = -0.29$ ). With only 69 unique correlated top genes, effect  $\nu = 134$  is more specific than the COO induced effect with 151 unique top genes for the same correlation cutoff. One of the novel effect's top genes is KLHL6. As KLHL6 proteins contain a domain that is known for transcriptional repression activity and might be involved in the germinal center B-cell differentiation pathway<sup>[109]</sup>, it could potentially be a tumor suppressor gene. This would be consistent with significantly adverse patient outcome for lower KLHL6 expression. As they are linked by high GEP correlation, the same biological functions might be associated with FGD6 and other top-correlated genes of  $\nu = 134$ . These hypotheses about potential tumor suppressor genes could possibly be investigated by overexpression experiments in DLBCL cell lines that show low expression of effect  $\nu = 134$ . To identify these cell lines, they could be screened for their protein levels of KLHL6 and FGD6. Ideally, the proliferation of cells with low or nonexistent levels of these proteins can be stopped by corresponding overexpression experiments.

Effect  $\nu = 127$ <sup>(III.3.3.2)</sup> is the best secondary predictor variable with  $p = 5.6 \cdot 10^{-8}$  on top of the primary effect  $\nu = 134$  when training with samples from all R-CHOP treated patients<sup>(III.2.2.1)</sup>. It predicts hierarchical survival dependencies<sup>(III.2.2.3)</sup> that exclusively exist in the lower risk partition of the primary effect<sup>(Figure III.2.2.3.a)</sup>. To elucidate the biological pathway underlying  $\nu = 127$ , further biological experiments might potentially be helpful. Analytically, BACH2 seems to be a promising oncogene candidate for the lower risk partition of  $\nu = 134$ , i.e. predominantly for GCB DLBCL patients. To test this hypothesis, GCB DLBCL cell lines with high BACH2 protein expression could be selected for BACH2 knockdown experiments.

Effect  $\nu = 3$ <sup>(III.3.3.8)</sup> contributes to the final quivariate model<sup>(cf. Table III.2.5.2)</sup> with a Cox  $p$  value of  $6.3 \cdot 10^{-6}$ . It measures metallothionein expressions and predicts significant survival differences in context of two other effects. High metallothionein expression is associated with adverse outcome for DLBCL patients with low BCL2 expression<sup>(III.3.3.7)</sup> and high "T cell co-stimulation"<sup>(III.3.3.6)</sup>. Low metallothionein expression in this subset is associated with 95.5% average survival<sup>(cf. Figure III.3.3.8.c)</sup>. The molecular mechanisms for metallothionein overexpression are currently unclear. The gene ontology result for cellular response to zinc ion and published data for a lung cancer experiment using zinc ionophores<sup>[124]</sup> indicate that the expression of this effect may

directly correlate with zinc concentrations. In this context, a correlation of patient zinc blood levels with average expressions of this effect could potentially be interesting.

Besides major survival effects<sup>(III.3.3)</sup>, several additional effects with differential gene expressions have been discovered and validated across DLBCL patient cohorts. Some of these effects describe ordinary human gene expression differences like the quad-discovered gender effect<sup>(III.3.2.1)</sup> or a presumed blood concentration effect<sup>(III.3.4.2)</sup>. Further disease-specific effects might also be among these effect, because current chemotherapies may have no or only a constant impact on their underlying pathways. All biostatistical analyses have been performed for these effects as well and are provided on disc<sup>(page 231)</sup>.

## Research Perspectives

---

### ■ *Dissecting other cancer entities*

Given validated and biologically relevant results for DLBCL, a promising research perspective is the dissection of gene expression signals for other cancer entities. With a multitude of gene expression cohorts for various cancers already publicly available via the NCBI Gene Expression Omnibus<sup>([63], <http://www.ncbi.nlm.nih.gov/geo>)</sup> or via the Cancer Genome Atlas<sup>(see TCGA Research Network, <http://cancergenome.nih.gov>)</sup>, this perspective has enormous scope.

Furthermore, the recent<sup>(January 2015)</sup> United States Government Precision Medicine Initiative has the aim to measure genomic data for *up to one-million samples*<sup>[15]</sup>, an unprecedented amount of genomic data. This is ideal for signal dissection and may even allow the discovery of effects that concern only tiny fractions of samples for a particular disease. This initiative also underlines the priority of this research field<sup>[14]</sup> and the potential of signal dissection in particular.

### ■ *Biologically more specific genomic associations*

The conceptual problem of representing effects by flat sets<sup>(cf. II.1.2.6)</sup> has been demonstrated for DLBCL subtypes: Several genetically distinct GEP effects are significantly associated with ABC DLBCL and GCB DLBCL, but they show strikingly different predictive power with respect to patient outcome<sup>(e.g. III.3.4.1)</sup>. The same problem does not only concern sets of samples. Information is likewise lost when representing genomic effects by sets of genes, as is commonly done for gene signatures today. This can result in significant enrichments of the same signature for genetically distinct effects with distinct biological characteristics. For instance, the same ABC-versus-GCB signatures are significantly enriched for validated effects  $v = 129$ ,  $v = 134$  and  $v = 20$ , although these effects have clearly different top genes and again strikingly different predictive power. In brief,

flat sets for sample clusters or gene signatures do not encode enough information about biological effects to enable *specific* associations.

In contrast, signal dissection quantifies effects with much more information as is stored by flat sets. For example, gene axes store regulation strengths mediated by the underlying pathway for each single gene. Additionally, gene correlations describe the consistency of these regulations over all samples for each single gene. Based on these information and the associated effect foci, a precise cross-cohort validation of effects can be facilitated<sup>(III.1.2.1)</sup>. More precisely, this task is realized by weighted uncentered correlations<sup>(II.2.3.1)</sup> that can be complemented with measures of statistical significance<sup>(II.5.2)</sup>.

The same effect comparison method could also be utilized to replace gene set enrichment analyses in order to facilitate biologically more specific associations. An effects database similar to large public gene signature databases does not yet exist, but discovered DLBCL effects could provide a start. Many gene signatures are based on supervised analyses<sup>(I.2.1)</sup> or on hierarchical clustering<sup>(I.2.2.1)</sup> of gene expression signals that are already stored in public databases. Hence and in principle, such an effects database could be built semi-automatically by dissection of these stored gene expression signals, although this would require a tremendous amount of computation.

#### ■ *Towards standalone microarray classification for clinical applications*

To utilize results like survival predictors<sup>(III.2)</sup> for clinical applications, e.g. for therapy decisions and towards precision medicine, standalone microarray measurements for tissue samples from single patients should be robustly classifiable. The technological problem here is that absolute gene expression intensities may vary strongly by chip design and by lab, for example due to different measurement protocols or due to different sequences probed for the same genes. Not all sources for these differences are known. Typically, this problem is solved by measuring sufficiently large patient cohorts with exactly the same microarrays in exactly the same lab. Then gene expression ratios *relative* to average gene intensities in this cohort are analyzed to avoid problems originating from technology-specific gene offset intensities. For a clinical application based on only one standalone microarray however, this is not possible (all ratios would equal one). Hence, a way to dissect all technological offsets precisely from this standalone microarray is required in order to compare the remaining biological signal with known and validated biological effects (e.g. with the COO effect for DLBCL subtype classification).

Assuming that lab effects and offset effects are systematic errors and only a finite number of them exists, it may be possible to achieve this by signal dissection. First, raw gene expression signals for many patient cohorts that were measured by various microarray designs in different labs need to be dissected. Resulting systematic lab effects, i.e. all effects that do *not* validate across biologically equal cohorts, but do validate across technologically equal and biologically different cohorts, are recorded in a database. Now the standalone microarray could be tested for similar effects by computing correlations with gene axes of recorded and validated technical effects. Top-correlated offset effects or other lab effects could then be dissected from its signal, which should result in effective  $\log_2(\text{ratio})$ s that can subsequently be classified with validated biological effects. To quantify the confidence of such classifications,  $p$  values for weighted uncentered correlations could be utilized.

Such a normalization by dissection of known technical effects may be much more precise compared to global normalization methods like quantile normalization. This could provide a tool for precision medicine that can utilize existing molecular knowledge for future therapy decisions in clinical settings, even for cost-effective

standalone microarray measurements. Furthermore, this lab effects database could help to identify technological causes of lab effects in order to improve experimental reproducibility in the first place.

### ■ *Beyond bioscience*

Throughout this work, DLBCL was selected as concrete biological application for signal dissection. Therefore and lastly, let the following analogy illustrate the method's general applicability and potential beyond biosciences, whenever intrinsic data orders by correlations are of interest.

| System   | Bioscience example  | Astrophysics example  |
|--|---|---|
| Samples or system instances                                | tumor biopsies from patients  | stars or quasars (i.e. centers of active galaxies)  |
| Dimensions or system parts                                 | all known genes   | e.g. known spectral lines for all elements or other light-emitting entities (or simply an equidistantly discretized light spectrum) |
| Extrinsic order that tolerates non-local correlations      | order of genes by genomic sequence  | order of spectral lines by energy   |
| Observable signal  | gene expressions (i.e. mRNA molecule concentrations in cells)                             | light intensities in spectral intervals (spectral energy distributions)   |
| <b>Signal dissection</b>                                   |   |   |
| Discovered effect curves                                   | empirical laws of coordinated and specific gene regulation                                | empirical laws of coordinated and specific light emission   |
| Classification of samples by their correlations to effects | disease subtypes, i.e. groups of patients that may share the same drivers of pathogenesis | classes of stars that maintain common reactions and may share a similar element composition   |
| <b>Modeling by experts</b>                                 |   |   |
| Underlying system interactions                             | gene-regulating pathways (e.g. for the process of cell division)                          | light-emitting physical reaction pathways (e.g. for nuclear fusion)   |

#### General applicability of signal dissection by analogy

Similar to the explorative detection of so far unknown biological processes in cells, signal dissection could yield so far unknown light-emitting reactions in stars. Patients belonging to the same disease subtype correspond to stars that maintain a common reaction. This could be utilized for star classification, for example. Finally, reverse engineering of biological pathways by molecular biologists corresponds to reverse engineering of physical reaction chains in stars by astrophysicists.

Similar to bioscience, large and growing spectral data volumes are already publicly available, for example via the Sloan Digital Sky Survey<sup>(SDSS, see <http://www.sdss.org>)</sup>. Consistent with results for synthetic test scenarios<sup>(e.g. Figure II.6.3.1.a)</sup>, principal components in astrophysics are already known to be hard to interpret in terms of their physical properties<sup>(cf. I.1.4.1)</sup>. Hence, application of signal dissection to light spectra of measured stars or quasars might be a promising research perspective as well.



# Appendix: Digital Content

---

All dissection results for DLBCL, all 135 validated consensus GEP effects<sup>(III.1)</sup> and all performed biostatistical analyses for each consensus effect<sup>(see III.3 for examples)</sup> are provided on disc in both graphical and tabular format. Files of biological interest can be located quickly via master tables. Additionally, dissection results for all method validation scenarios<sup>(cf. II.6)</sup> are provided.

## ■ Master tables

Master tables in the root folder provide a convenient access to all results for DLBCL by linking to them.

The *main overview* table lists all 135 validated DLBCL consensus effects, links to their signal plots for each patient cohort, followed by links to their genomic analyses like signature analyses tables and gene ontology analyses folders. The next column group lists and links to survival analyses and to associations with clinical covariates like gender or IPI scores. These analyses are provided separately for all patients and for ABC DLBCL and GCB DLBCL subsets. The last column group shows related effects having similar top genes in form of their consensus gene axis correlations.

The *gene orders* master table lists consensus gene axes and consensus gene correlations for all measured probesets and for each effect. Annotation columns like genomic alignments and probeset IDs allow comparing genomic loci of top genes and allow retrieving exact sequences that underlie correlated gene expressions.

In the *sample orders* master table, patients from all four cohorts are listed together with their consensus sample axes and consensus sample correlations for each effect. These columns allow sorting all patients by



Blu-ray disc) All dissection results and biostatistical analyses in digital form (ca. 18.2 GB, ca. 65.000 files)

their involvement in a given effect. Another column group in this table lists all available clinical data about those samples, including available follow-up information.

(Master tables were tested to work on a Windows® 7 PC with Excel® 2013 and 8GB RAM. Weaker PCs or older versions of Excel® might have problems due to the large table size. In this case, the directory structure described below allows opening smaller result files for individual effects manually.)

## ■ Directors structure

### 📁 A=Detection

- Contains *cohort subfolders* for GSE10846.CHOP, GSE10846.R-CHOP, GSE4475 and GSE31312.
  - For each discovered and dissected effect, several files exist, named with the effect's dissection rank. Most importantly, files named like *007, effect overview.eps* show the effect's heatmap, its regressed eigensignal and the remaining signal for further dissection. Files named like *007, definition.xlsx* contain converged gene and sample axes, correlations and *p* values for an effect, both before and after dissection to control dissection efficacy.

### 📁 B=Validation and Consensus Eigenorders

- *\A=cohorts vs cohorts*: Contains gene order correlations between detected effects of all cohorts based on product gene scores<sup>(cf. III.1.3.2)</sup>. These correlations are the basis for unsupervised cross-cohort validation of detected GEP effects.
- *\validatedEffects*: Contains scatter plots for validated effect tuples.

### 📁 C=Consensus Effects

- Contains *cohort subfolders* for GSE10846.CHOP, GSE10846.R-CHOP, GSE4475 and GSE31312.
  - For each validated effect, similar files exist as for detection. However, this time each validated effect is applied to and dissected from the cohort's initial GEP signal (except for cleaned lab effects).

### 📁 D=Interpretation

- *\genomic*
  - Contains *effect subfolders* named like *v007*.
    - *\SA*: These subfolders contain signature analyses, together with enrichment plots and heatmaps for strongly enriched signatures (genes ranked by product gene scores<sup>(cf. III.1.3.2)</sup>).
    - *\GOA*: Contains gene ontology overrepresentation analyses for effect top gene signatures of various size. Analyses are available for term trees of biological processes, molecular functions and cellular components. (Not available if the effect has less than five top genes.)
- *\clinical*
  - Contains *effect subfolders* named like *v007*.
    - Clinical cohort subfolders like *GSE10846\_RCHOP.GCB.PFS*
      - Kaplan-Meier survival estimates for patient subsets cut by the effect's sample scores<sup>(cf. III.3.1.1)</sup>. Excel files contain plots in tabular form.
      - Excel files for clinical correlations contain contingency sub tables for each covariate.
    - *\comparisons*: Contains Kaplan-Meier survival estimates for standard ABC DLBCL, unclassified and GCB DLBCL patient subsets for comparison.

### 📁 Method Validation

- Contains subfolders for various synthetic data scenarios for method validation<sup>(II.6)</sup>.

# Indexes

## Symbols

### ■ Spaces and the initial signal

|                           |  |
|---------------------------|--|
| $m$                       | Number of measured dimensions or <b>genes</b>  |
| $n$                       | Number of measured <b>samples</b>  |
| $\mathbb{R}^{m \times n}$ | <b>Signal space</b> , i.e. the matrix space $\mathbb{R}^{m \times n} \equiv \{(X_{i,j})_{i=1 \dots m, j=1 \dots n}   X_{i,j} \in \mathbb{R}\}$ . The complete gene expression signal measured for a patient cohort is one element in this signal space, for example.                     |
| $M_0$                     | <b>Initial signal matrix</b> $\in \mathbb{R}^{m \times n}$   |
| $V^g$                     | <b>Gene vector space</b> over $\mathbb{R}^m$ , spanned by $m$ gene basis vectors $\{ e_i^g\rangle   i = 1 \dots m\}$ . Contains measured <b>samples</b> $ s_j\rangle$ with their expressions $\langle e_i^g   s_j \rangle$ for all genes $i = 1 \dots m$ . See II.2.2.1 for details.     |
| $V^s$                     | <b>Sample vector space</b> over $\mathbb{R}^n$ , spanned by $n$ sample basis vectors $\{ e_j^s\rangle   j = 1 \dots n\}$ . Contains measured <b>genes</b> $ g_i\rangle$ with their expressions $\langle e_j^s   g_i \rangle$ for all samples $j = 1 \dots n$ . See II.2.2.1 for details. |
| $I_0^g$                   | External measurement or <b>reference order of genes</b> ; without loss of generality $I_0^g \equiv (1, 2, \dots, m)$   |
| $J_0^s$                   | External measurement or <b>reference order of samples</b> ; without loss of generality $J_0^s \equiv (1, 2, \dots, n)$   |
| $ X(I_0^g, j)\rangle$     | Abbreviation for the vector $ x\rangle \in V^g$ with components $\forall i = 1 \dots m: \langle e_i^g   x \rangle \equiv X(I_0^g(i), j)$   |
| $ X(i, J_0^s)\rangle$     | Abbreviation for the vector $ x\rangle \in V^s$ with components $\forall j = 1 \dots n: \langle e_j^s   x \rangle \equiv X(i, J_0^s(j))$   |
| $X(I, J)$                 | Permuted matrix for <b>sort vectors</b> $(I, J)$ , i.e. $X(I, J) \equiv (X(I(i), J(j)))_{i=1 \dots m, j=1 \dots n}$ . Sort vectors are permutations of row indices $I_0^g$ or column indices $J_0^s$ , respectively. In particular, $M_0(I_0^g, J_0^s) = M_0$ .                          |

### ■ Basic operations and functions

|                               |   |
|-------------------------------|---|
| $ x\rangle, \mathbf{x}$       | Vector $ x\rangle$ and its coordinate array, e.g. for $ x\rangle \in V^g$ the column array $\mathbf{x} = (\langle e_i^g   x \rangle)_{i,1} \in \mathbb{R}^{m \times 1}$   |
| $\mathbf{x}\mathbf{y}$        | Matrix multiplication, defined as $(\sum_k x_{i,k} y_{k,j})_{i,j}$ . E.g. for a row vector $\mathbf{x} \in \mathbb{R}^{1 \times m}$ and a column vector $\mathbf{y} \in \mathbb{R}^{m \times 1}$ the scalar $\sum_{k=1 \dots m} x_{1,k} y_{k,1} \in \mathbb{R}$ . In case of a column vector $\mathbf{x} \in \mathbb{R}^{m \times 1}$ and a row vector $\mathbf{y} \in \mathbb{R}^{1 \times n}$ the matrix $(x_i y_j)_{i=1 \dots m, j=1 \dots n} \in \mathbb{R}^{m \times n}$ . |
| $\langle x  , \mathbf{x}^T$   | Dual vector for $ x\rangle$ and transpose operation for its coordinate array $\mathbf{x}$ . E.g. for $ x\rangle \in V^g$ the dual vector is $\langle x   \in V^{g*}$ ; it is computed on coordinate level via transposition and hence equals the row vector $\mathbf{x}^T = (\langle e_i^g   x \rangle)_{i,1} \in \mathbb{R}^{1 \times m}$ .  |
| $\mathbf{x} \cdot \mathbf{y}$ | Hadamard product (i.e. component-wise multiplication); yields a vector or matrix of the same size.  |
| $\langle x   a \rangle$       | Scalar product aka dot product; for $ x\rangle,  a\rangle \in V^g$ defined on coordinate level as $\mathbf{x}^T \mathbf{a} \in \mathbb{R}$ .  |
| $\ x\ $                       | Euclidean vector norm $\ x\  \equiv \sqrt{\langle x   x \rangle} = (\sum_i x_i^2)^{1/2}$  |
| $ x\rangle \otimes  y\rangle$ | Tensor product aka outer product; e.g. for $ x\rangle \in V^g,  y\rangle \in V^s$ defined as $\mathbf{x}\mathbf{y}^T \in \mathbb{R}^{m \times n}$ .   |
| $\hat{\sigma}_{uc}(X)$        | Uncentered standard deviation of $X$ , i.e. $\sqrt{\hat{E}[X^2 - 0]}$ , where $\hat{E}$ denotes the expectation estimator.  |

### ■ Central measures of interaction

|                                       |  |
|---------------------------------------|--|
| $\langle x   a \rangle_{ w\rangle}^0$ | <b>Weighted projections</b> of a vector $ x\rangle$ in direction of an axis $ a\rangle$ using dimension weights $ w\rangle$ . Equals the normalized weighted scalar product defined as $\langle w \cdot x   w \cdot a \rangle / \ w \cdot a\ $ . See II.2.3.2 for details. |
| $[x   a]_{ w\rangle}$                 | <b>Weighted uncentered correlation</b> aka the weighted cosine distance between $ x\rangle$ and $ a\rangle$ using dimension weights $ w\rangle$ . Defined as $\langle w \cdot x   w \cdot a \rangle / (\ w \cdot x\  \ w \cdot a\ )$ . See II.2.3.1 for details.           |

### ■ Search strategy and effect axes convergence

- $k, \hat{k}$  Index of the **detection iteration** and of the effect detected by it. The total number of detected effects before termination is denoted by  $\hat{k}$ .
- $M_{k-1}$  Input signal for detection iteration  $k$  that ultimately yields the remaining signal  $M_k$  after dissecting effect  $k$ . (See the signal model in II.2.1.1.)
- $M_{k-1}^S$  Standardized signal matrix with uncentered variances equaling one for all rows and columns (cf. II.3.1.1).
- $i_{k,1}, j_{k,1}$  Gene index  $i_{k,1} \in [1, m] \subset \mathbb{N}$  respectively sample index  $j_{k,1} \in [1, n] \subset \mathbb{N}$  selected as **initial representative** for effect  $k$  by the search strategy (cf. II.3.1).
- $|a^g\rangle, |a^s\rangle$  Gene and sample **effect axes**. Based on the selected initial gene or sample and its twin axis (cf. II.3.1.3).
- $|w_{\text{initial}}^g\rangle, |w_{\text{initial}}^s\rangle$  Initial gene and sample weights based on the standardized signal (cf. II.3.1.2).
- $|r^g\rangle, |r^s\rangle$  Gene and sample **correlations** to the respective effect axis (cf. II.3.1.4).
- $|p^g\rangle, |p^s\rangle$  Gene and sample  $p$  values for correlations (cf. II.5.2.1).
- $|w^g\rangle, |w^s\rangle$  Gene and sample weights aka the **effect focus**. Based on correlations and their significance (cf. II.3.1.4).
- $s_k$  Scalar **effect score** based on correlations of genes and samples with the effect and on the effect size (II.3.1.6).
- $l, \hat{l}$  Iteration index of effect axes convergence (cf. II.3.2); equals the number of representatives utilized so far for definition of the effect's axes.  $\hat{l}$  denotes the number of representatives considered sufficient (cf. II.3.2.3).
- $|a_l^g\rangle, |a_l^s\rangle$  Gene axis and sample axis for representative  $l$  (cf. II.3.1.3).
- $|b_l^g\rangle, |b_l^s\rangle$  Accumulated gene axis and accumulated sample axis for representatives  $1 \dots l$  (cf. II.3.2.3).
- $|b_l^g\rangle, |b_l^s\rangle, |r_l^g\rangle, |r_l^s\rangle, |p_l^g\rangle, |p_l^s\rangle, |w_l^g\rangle, |w_l^s\rangle$  **Converged effect axes** based on  $\hat{l}$  selected representative genes or samples, **final correlations** to these axes and their significance, and the **final effect focus** for effect  $k$ . (The index  $k$  is clear from the context and suppressed for readability reasons.)

### ■ Bimonotonic regression, effect eigensignal and its dissection

- $i, \hat{i}$  Index of outer regression iterations (cf. II.4.1) and the converged iteration  $\hat{i}$  (cf. II.4.1.5).
- $|u_{k,i}^g\rangle, |u_{k,i}^s\rangle$  **Effect strengths** for effect  $k$  in regression iteration  $i$ . Defined as projections of all genes and samples on the respective final effect axis or on the regressed effect curves (cf. II.4.1.1).
- $I_{k,i}^g, J_{k,i}^s$  Empirical effect **eigenorder** based on effect strengths (cf. II.4.1.2 and also see the effect model in II.2.1.2).
- $M_{k-1}(I_{k,i}^g, J_{k,i}^s)$  Current signal in the empirical eigenorder.
- $\hat{j}, \hat{j}$  Index of inner bimonotonic regression iterations (cf. II.4.1.3) (and convergence iteration  $\hat{j}$ ).
- $\mathcal{M}_{i,\hat{j}}$  Result of the converged iterative **bimonotonic regression** of the signal in empirical eigenorder (cf. II.4.1.3).
- $S(\mathcal{M}_{i,\hat{j}})$  **Adaptive smoothing** of the result from bimonotonic regression using rescaling and 2D Fourier transformations (cf. II.4.1.4).
- $D$  **Dissection strengths** of the effect, defined based on the product effect focus (cf. II.4.2.1).
- $E_k$  **Eigensignal** of detected effect  $k$  (cf. II.4.2.1 and also see the effect model in II.2.1.2).

### ■ Effect validation and scores for biostatistical association analyses

- $\hat{C}$  Number of independently dissected cohorts that are available for validation (cf. III.1.2.2).
- $|a^g\rangle, |r^g\rangle$  **Gene scores** based on the gene axis and gene correlations; defined as  $\sum_{i=1}^m | \langle e_i^g | a^g \rangle | | \langle e_i^g | r^g \rangle | | e_i^g \rangle$ . Used as basis for cross-cohort comparison of effects (cf. III.1.2.1), to compare consensus gene effects (cf. III.1.3.2) and to associate effects with genomic knowledge (cf. Table III.1.5).

- $r_{(C_1, k_1, C_2, k_2)}$  Correlations between gene scores of effects  $k_1$  from cohort  $C_1$  and  $k_2$  from cohort  $C_2$  (cf. III.1.2.1).
- $\mathcal{C}, \mathcal{R}, \mathcal{V}$  For each possible effects tuple, the counts of significant pairwise correlations between effects from different cohorts, their average correlation to each other and a validation score (cf. III.1.2.2).
- $\nu$  **Validation index** for all unsupervisedly (cf. III.1.2.2) and selected supervisedly (cf. III.1.2.3) validated effects.
- $\{a_v^{g,c}, r_v^{g,c}, w_v^{g,c}\}$  Consensus gene axes, correlations and weights for detected effects for validation index  $\nu$  (III.1.3.1).
- $M_{C,0}^{\text{cleaned}}$  Cleaned signal of cohort  $C$  (cf. III.1.4.1).
- $E_{C,\nu}^C$  Eigensignal of consensus effect  $\nu$  in cohort  $C$  (cf. III.1.4.2).
- $\{u_{C,\nu}^S\}$  **Sample eigensignal strengths** of consensus effect  $\nu$  in cohort  $C$  (cf. III.1.4.2). Used to associate effects with clinical knowledge (cf. Table III.1.5).

## Named Equations

|  |    |  |     |
|--|----|--|-----|
| Eqn. II.2.1.1) Signal model.....   | 46 | Eqn. II.4.1.4.b) Application of the convolution theorem reduces the smoothing task to 2D Fourier transforms and component-wise multiplication .....                | 69  |
| Eqn. II.2.1.2) Effect Model .....  | 47 | Eqn. II.4.2.1.a) Dissection strengths.....   | 71  |
| Eqn. II.2.3.1) Weighted uncentered correlation .....   | 50 | Eqn. II.4.2.1.b) Eigensignal of the discovered effect .....  | 71  |
| Eqn. II.2.3.2) Weighted projections.....   | 51 | Eqn. II.4.2.1.c) Remaining signal after dissection of the discovered effect $k$ .....  | 71  |
| Eqn. II.3.1.2) Initial effect focus based on the standardized signal .....                   | 55 | Eqn. II.5.1.2.a) $t$ statistic for the difference between the mean signal amplitude in the focus of an effect and the mean noise signal amplitude .....            | 80  |
| Eqn. II.3.1.3) Initial twin axes for symmetrization .....                                    | 55 | Eqn. II.6.2.3) Correlations of detected and simulated effect axes.....   | 93  |
| Eqn. II.3.1.4.a) Initial correlations.....   | 56 | Eqn. III.1.2.1.a) Gene scores (combining signal strength and correlation information) .....  | 129 |
| Eqn. II.3.1.4.b) Refined effect focus based on correlations .....                            | 56 | Eqn. III.1.2.1.b) Comparison of effects from two cohorts by correlation .....  | 129 |
| Eqn. II.3.1.6) Scalar effect score (average correlation times effect size) .....             | 57 | Eqn. III.1.3.1) Consensus gene axis, consensus gene correlations and consensus gene weights for a validated effect $\nu$ describe a genomic consensus effect ..... | 132 |
| Eqn. II.3.2.2) Accumulated effect axes over $l$ effect representatives .....                 | 61 | Eqn. III.1.3.2.a) Consensus gene scores for a validated effect $\nu$ (combining signal strength and correlation information).....                                  | 133 |
| Eqn. II.4.1.1.a) Effect strengths by weighted projections on effect axes for $i = 1$ .....   | 64 | Eqn. III.1.3.2.b) Correlation of consensus gene scores .....   | 133 |
| Eqn. II.4.1.1.b) Effect strengths by weighted projections on effect curves for $i > 1$ ..... | 64 | Eqn. III.1.4.1) Gene expression signal after dissecting cohort-specific effects .....  | 135 |
| Eqn. II.4.1.2) Empirical effect eigenorder based on effect strengths.....                    | 64 | Eqn. III.1.4.2) Sample eigensignal strengths.....  | 136 |
| Eqn. II.4.1.3.a) The product effect focus and initial weights for 1D regressions.....        | 65 | Eqn. III.2.1.1) Cox proportional hazard models .....   | 139 |
| Eqn. II.4.1.3.b) Interweaved regression weights for genes and samples.....                   | 67 |  |     |
| Eqn. II.4.1.3.c) Bimonotonic regression step .....   | 67 |  |     |
| Eqn. II.4.1.4.a) Rescaled and downscaled gene and sample space for smoothing.....            | 69 |  |     |

## Tables

|   |     |  |     |
|---|-----|--|-----|
| Table I.1.3.4) Example concepts of generic summarizations by interaction.....   | 11  | Table III.2.4.2) Bivariate Cox model $\nu \in \{5,44\}$ , CHOP training set.....   | 156 |
| Table II.3.1.8) Qualification thresholds for effect candidates .....  | 59  | Table III.2.5.2) Quinivariate Cox model, final statistics for the complete training set based on available samples and follow-up data for all DLBCL patients ..... | 163 |
| Table II.6.2.5) Versatility test scenario with 13 effects, 49 runs, comparison between signal dissection and PCA results..... | 96  | Table III.3.2.1.a) Top genes in validated effect $\nu = 2$ .....   | 173 |
| Table III.1.5) Overview of available gene and sample scores for validated effects .....                                       | 137 | Table III.3.2.1.b) Gender effect, contingency with clinical annotations.....   | 174 |
| Table III.2.2.2) Bivariate Cox model $\nu \in \{134,127\}$ , final statistics for the R-CHOP training set.....                | 147 | Table III.3.2.2.a) Top genes in validated effect $\nu = 129$ .....   | 178 |
| Table III.2.3.2) Bivariate Cox model $\nu \in \{129,105\}$ , R-CHOP training set.....   | 153 | Table III.3.2.2.b) Top-enriched signatures by $\nu = 129$ .....  | 180 |
|   |     | Table III.3.3.1.a) Top-enriched signatures by $\nu = 134$ .....  | 186 |

|  |     |   |     |
|--|-----|---|-----|
| Table III.3.3.1.b) Top genes in validated effect $\nu = 134$ ..... | 188 | Table III.3.3.6.b) Top genes in validated effect $\nu = 19$ ..... | 207 |
| Table III.3.3.2.a) Top-enriched signatures by $\nu = 127$ .....    | 191 | Table III.3.3.7.a) Top genes in validated effect $\nu = 75$ ..... | 208 |
| Table III.3.3.2.b) Top genes in validated effect $\nu = 127$ ..... | 193 | Table III.3.3.8.a) Top-enriched signatures by $\nu = 3$ .....     | 212 |
| Table III.3.3.3.a) Top genes in validated effect $\nu = 105$ ..... | 195 | Table III.3.3.8.b) Top genes in validated effect $\nu = 3$ .....  | 212 |
| Table III.3.3.4.a) Top genes in validated effect $\nu = 5$ .....   | 196 | Table III.3.4.2) Blood module signatures that are significantly   |     |
| Table III.3.3.4.b) Top-enriched signatures by $\nu = 5$ .....      | 198 | enriched for $\nu = 7$ .....                                      | 220 |
| Table III.3.3.5.a) Top-enriched signatures by $\nu = 44$ .....     | 201 | Table III.3.4.3.a) Top genes in validated effect $\nu = 4$ .....  | 221 |
| Table III.3.3.5.b) Top genes in validated effect $\nu = 44$ .....  | 203 |   |     |
| Table III.3.3.6.a) Top-enriched signatures by $\nu = 19$ .....     | 206 |   |     |

## Figures

|   |    |   |    |
|---|----|---|----|
| Figure I.1.2.1) Definition of subtypes ABC DLBCL and GCB<br>DLBCL <sup>[4]</sup> .....  | 5  | Figure II.3.1) 3D concept example, first detection iteration, initial<br>representative for the blue effect and its associated initial gene<br>axis .....     | 52 |
| Figure I.1.2.2) Classical NF- $\kappa$ B pathway <sup>[7]</sup> .....   | 6  | Figure II.3.1.1.a) Standardization results $M_0^S$ in gene space for<br>the 3D example via equalization of uncentered variances .....                         | 53 |
| Figure I.1.2.3) PTEN-stained DLBCL cells <sup>[12]</sup> .....  | 7  | Figure II.3.1.1.b) Standardization results in gene space for the 3D<br>example when equalizing absolute norms .....   | 53 |
| Figure I.1.3.5) Typical stages of systems research with examples,<br>mostly from biosciences .....  | 13 | Figure II.3.2) 3D concept example, first detection iteration,<br>generalizing representatives and converged gene axis for the<br>blue effect .....            | 60 |
| Figure I.1.4.1) Principal components do not point into effect<br>directions.....  | 14 | Figure II.4.1) 3D example, blue effect curve in gene space .....  | 63 |
| Figure I.1.4.2.a) 3D concept example with four effects.<br>Misleading principal components and precise effect curves by<br>signal dissection (preview).....                           | 16 | Figure II.4.1.3) Illustration of bimonotonic regression (inner<br>convergence loop) .....   | 66 |
| Figure I.1.4.2.b) Example of a roughly bimonotonic real-world<br>effect <sup>[29]</sup> .....   | 16 | Figure II.4.1.4) Illustration of adaptive signal smoothing<br>(realized by rescaling and 2D Fourier transformation).....                                      | 68 |
| Figure I.2.1.1) Example for a supervised gene expression<br>analysis that determines significantly differentially expressed<br>genes between two known settings .....                 | 19 | Figure II.4.2.2.a) 3D example, remaining signal $M_1$ after<br>dissection of the blue effect $E_1$ .....  | 71 |
| Figure I.2.1.2) Application of an identified NF- $\kappa$ B signature (left)<br>to gene expressions following treatment with a PI3K<br>inhibitor .....                                | 20 | Figure II.4.2.2.b) Dissection with and without using the effect<br>focus as dissection strengths.....   | 72 |
| Figure I.2.2.1.a) Example result from bi-hierarchical clustering  | 23 | Figure II.4.2.3.a) Coordinate view for high-dimensional<br>visualization; exemplary dissection of the green effect for the<br>3D example .....                | 73 |
| Figure I.2.2.1.b) Example result from bi-hierarchical clustering<br>for a more complex gene expression signal.....  | 24 | Figure II.4.2.3.a) 3D example, remaining signal after all four<br>dissection iterations.....  | 75 |
| Figure I.2.2.2.a) Simple 3D illustration of PCA.....  | 25 | Figure II.5.1.2) Signal significance in the 3D example .....  | 79 |
| Figure I.2.2.2.b) Principal components do not point into effect<br>directions.....  | 26 | Figure II.5.2.1) Comparison of $p$ value computation methods for<br>weighted correlations .....   | 82 |
| Figure I.2.3) Exemplary aCGH analyses .....   | 27 | Figure II.6.1.1) 3D concept example, all detected gene curves, all<br>dissection steps and the remaining signal .....   | 84 |
| Figure I.2.4) Examples of IC50 fits for an anti-CD79B ADC.....  | 28 | Figure II.6.1.2) 3D concept example, comparison of final<br>detected gene curves and principal components .....   | 85 |
| Figure I.3.1.1) Significant enrichment of a proliferation related<br>signature.....   | 30 | Figure II.6.1.3.a) 3D concept example, results from hierarchical<br>clustering (Euclidean distance metric, average linkage).....                              | 86 |
| Figure I.3.1.2) A immediate early immune response signature<br>identified by gene set enrichment analyses.....  | 31 | Figure II.6.1.3.b) 3D concept example, results from hierarchical<br>clustering (Correlation distance metric, average linkage).....                            | 86 |
| Figure I.3.3) Kaplan-Meier survival analyses show significantly<br>more favorable outcome for high NF- $\kappa$ B expressions in a GCB<br>DLBCL subset with high BCL2 expression..... | 33 | Figure II.6.2.1) Versatility test, 7 simulated effect patterns and<br>the superposed signal.....  | 87 |
| Figure II.1.1.1) 3D example; red effect: a linear law of gene<br>regulation .....   | 37 | Figure II.6.2.2.a) Versatility test with 7 effects, detection and<br>dissection of the large overlapping lab effect (pattern #1,<br>iteration $k = 1$ ) ..... | 88 |
| Figure II.1.1.2) 3D example; blue, green and magenta effects:<br>monotonic non-linear laws of gene regulation.....  | 38 | Figure II.6.2.2.b) Versatility test with 7 effects, detection and<br>dissection of pattern #3 (in iteration $k = 2$ ).....                                    | 90 |
| Figure II.1.1.3) Merged 3D example with four effects.....   | 39 |   |    |
| Figure II.2.1.2) Example of a roughly bimonotonic real-world<br>effect <sup>[29]</sup> .....  | 47 |   |    |

|   |     |   |     |
|---|-----|---|-----|
| Figure II.6.2.2.c) Versatility test with 7 effects, detection and dissection of pattern #5 (in iteration $k = 6$ ).....   | 91  | Figure II.6.4.2.d) Example of a false positive discovery (from a versatility test embedded in a noise signal with 51000 genes) .....                              | 117 |
| Figure II.6.2.2.d) Versatility test with 7 effects, detection and dissection of pattern #6 (in iteration $k = 7$ ).....   | 92  | Figure II.6.4.3.a) Versatility test with 7 effects, 20...2000 samples, correlations of gene axes and comparison of signal dissection with PCA.....                | 119 |
| Figure II.6.2.3.a) Versatility test with 7 effects, correlation of detected and simulated gene axes.....  | 93  | Figure II.6.4.3.b) Versatility test with 7 effects simulated for only 20 samples, detection and dissection of pattern #4 with several false positive genes.....   | 120 |
| Figure II.6.2.3.b) Versatility test with 7 effects, correlation of detected and simulated sample axes.....  | 94  | Figure II.6.4.4.a) Versatility test with 7 effects, 0%...100% missing values, correlations of detected gene and sample axes with simulated ones.....              | 122 |
| Figure II.6.2.3.c) Detected effect eigensignals in their true simulated eigenorder for the versatility test with 7 effects ....   | 94  | Figure II.6.4.4.b) Detection of effect pattern #3 for the versatility test with 7 effects and 80% missing values.....   | 123 |
| Figure II.6.2.4.a) Versatility test with 7 effects, 49 runs, correlations of effect axes and comparison of signal dissection with PCA .....                                   | 95  | Figure II.6.4.4.c) Versatility test with 80% missing values.....  | 124 |
| Figure II.6.2.5.a) Versatility test with 13 effects, 49 runs, correlations of gene axes and comparison of signal dissection with PCA .....                                    | 97  | Figure II.6.4.4.d) Imputed eigensignals (versatility test, 80% missing values) .....  | 124 |
| Figure II.6.2.5.b) Versatility test with 13 effects, 49 runs, correlations of sample axes and comparison of signal dissection with PCA.....                                   | 98  | Figure III.1.3.2.a) Pairwise correlation $r_{(v_1;v_2)}^c$ of consensus gene scores for all validated effects.....  | 134 |
| Figure II.6.2.6) Versatility test with 7 effects, results from hierarchical clustering (corr. distance, average linking).....   | 100 | Figure III.2.1.8) Survival spread for standard classifications in ABC DLBCL, unclassified and GCB DLBCL in all four analyzed patient cohorts for comparison ..... | 143 |
| Figure II.6.3.1.a) Superposition tests based on pattern #3, 1 to 20 instances, correlations of gene axes and comparison of signal dissection with PCA.....                    | 102 | Figure III.2.2.1.a) Selection of the best primary GEP effect for survival prediction based on all 690 available R-CHOP-treated patients .....                     | 145 |
| Figure II.6.3.1.b) Superposition tests based on pattern #3, 1 to 20 instances, correlations of sample axes and comparison of signal dissection with PCA.....                  | 103 | Figure III.2.2.1.b) Selection of the best secondary GEP effect for survival prediction based on all 690 available R-CHOP-treated patients .....                   | 146 |
| Figure II.6.3.1.c) Superposition test based on pattern #3, dissection of the signal with 20 superposed effect instances, detection iteration 20 .....                         | 104 | Figure III.2.2.2) Bivariate predictor $v \in \{134,127\}$ ; performance in the R-CHOP training set and generalization to the CHOP validation set.....             | 147 |
| Figure II.6.3.2.a) Superposition tests based on pattern #4, 1 to 20 instances, correlations of gene axes and comparison of signal dissection with PCA.....                    | 106 | Figure III.2.2.3.a) Univariate predictor based on $v = 127$ applied to risk partitions of $v = 134$ .....   | 148 |
| Figure II.6.3.2.b) Superposition tests based on pattern #4, 20 instances, detection and dissection of the accumulated offset effect.....                                      | 107 | Figure III.2.2.3.b) Univariate predictor based on $v = 127$ separately applied to GCB DLBCL and ABC DLBCL.....  | 149 |
| Figure II.6.3.3) Superposition tests based on pattern #6, 1 to 20 instances, correlations of gene axes and comparison of signal dissection with PCA.....                      | 109 | Figure III.2.2.4) Bivariate predictor $v \in \{134,127\}$ , predictor performance within risk partitions of $v = 134$ .....                                       | 150 |
| Figure II.6.4.1.a) Detection limit with respect to signal strength, effect pattern #6, signal dissection results .....  | 110 | Figure III.2.2.5) Bivariate predictor $v \in \{134,127\}$ , performance in subtypes ABC DLBCL and GCB DLBCL .....   | 151 |
| Figure II.6.4.1.b) Detection limit with respect to signal strength, effect pattern #6, PCA results.....   | 111 | Figure III.2.3.1) Selection of the best secondary GEP effect for survival prediction based on all 690 available samples from R-CHOP-treated patients .....        | 152 |
| Figure II.6.4.1.c) Detection limit with respect to signal strength, effect pattern #3, signal dissection and PCA results .....  | 112 | Figure III.2.3.2) Bivariate predictor $v \in \{129, 105\}$ , performance in the R-CHOP training set and generalization to the CHOP validation set.....            | 154 |
| Figure II.6.4.2.a) Versatility test with 7 effects, 1000...55000 genes, correlations of gene axes and comparison of signal dissection with PCA.....                           | 113 | Figure III.2.4.1.a) Selection of the best primary GEP effect for survival prediction based on all 257 available samples from CHOP-treated patients .....          | 155 |
| Figure II.6.4.2.b) Detection of the narrow pattern #7 with many false positive genes (in a versatility test with 7 effects embedded in a noise signal with 39000 genes) ..... | 115 | Figure III.2.4.1.b) Selection of the best secondary GEP effect for survival prediction based on all 257 available samples from CHOP-treated patients .....        | 156 |
| Figure II.6.4.2.c) Versatility test with 7 effects embedded in a noise signal with 55000 genes, detection and dissection of the small pattern #5 .....                        | 116 | Figure III.2.4.2) Bivariate predictor $v \in 5,44$ ; performance in the CHOP training set and its generalization to the R-CHOP validation set.....                | 157 |

|  |     |   |     |
|--|-----|---|-----|
| Figure III.2.4.3) Univariate predictor based on $\nu = 44$ applied to risk partitions of $\nu = 5$ .....                             | 158 | Figure III.3.3.3) Validated effect $\nu = 105$ , applied to GSE31312 .....  | 194 |
| Figure III.2.4.4) Univariate predictor based on $\nu = 44$ applied to DLBCL subtypes .....   | 159 | Figure III.3.3.4.a) Gene ontology overrepresentation analyses of cellular components in effect $\nu = 5$ .....  | 196 |
| Figure III.2.5.1.a) Selection of the primary GEP effect for survival prediction based on all 947 available events .....              | 161 | Figure III.3.3.4.b) Gene ontology overrepresentation analyses of biological processes and molecular functions in effect $\nu = 5$ .....                   | 197 |
| Figure III.2.5.1.b) Selection of the secondary GEP effect for survival prediction based on all 947 available events.....             | 161 | Figure III.3.3.4.c) Significant enrichment of the stromal-1 signature <sup>[5]</sup> .....  | 198 |
| Figure III.2.5.1.c) Selection of the tertiary GEP effect for survival prediction based on all 947 available events .....             | 162 | Figure III.3.3.4.d) Validated effect $\nu = 5$ , applied to GSE10846.CHOP .....   | 199 |
| Figure III.2.5.3.a) Distributions of $\beta$ , over all 947 leave-one-out fits.....  | 163 | Figure III.3.3.5) Validated effect $\nu = 44$ , exemplary application to GSE10846.R-CHOP .....  | 202 |
| Figure III.2.5.3.b) Quinvariate predictor performance for all 947 DLBCL samples based on leave-one-out validation.....               | 164 | Figure III.3.3.6.a) Validated effect $\nu = 19$ , applied to GSE10846.R-CHOP.....   | 205 |
| Figure III.2.5.4) Quinvariate predictor performance for all CHOP-treated and R-CHOP-treated cohorts .....                            | 165 | Figure III.3.3.6.b) Gene ontology overrepresentation analyses of biological processes in effect $\nu = 19$ .....  | 206 |
| Figure III.2.5.5) Quinvariate predictor performance, separately for frozen cell material and FFPE based R-CHOP cohorts.....          | 166 | Figure III.3.3.7) Validated effect $\nu = 75$ , applied to GSE10846.R-CHOP and GSE4475 .....  | 209 |
| Figure III.2.5.6) Quinvariate model, predicted survival dependencies within standard ABC- and GCB-like DLBCL subtypes.....           | 167 | Figure III.3.3.8.a) Gene ontology overrepresentation analyses of biological processes and cellular components for effect $\nu = 3$ .....                  | 210 |
| Figure III.2.5.7) Quinvariate model, predicted survival differences within risk partitions of $\nu = 134$ .....                      | 168 | Figure III.3.3.8.b) Validated effect $\nu = 3$ , applied to GSE10846.R-CHOP and GSE31312.R-CHOP .....   | 211 |
| Figure III.2.5.8) Quinvariate model, predicted survival differences within IPI risk classes .....                                    | 170 | Figure III.3.3.8.c) Survival dependency on effect $\nu = 3$ for low BCL2 expression and high “T cell co-stimulation” .....                                | 213 |
| Figure III.3.2.1) Detected gender effects .....  | 175 | Figure III.3.4.1.a) The quad-discovered effect $\nu = 20$ is significantly associated with DLBCL subtypes, but not with a consistent survival trend ..... | 216 |
| Figure III.3.2.2.a) Validated GEP effects that are top-associated with public sample classifications as ABC DLBCL or GCB DLBCL ..... | 176 | Figure III.3.4.1.b) Significant enrichment of known ABC-versus-GCB DLBCL signatures by effect $\nu = 20$ .....  | 217 |
| Figure III.3.2.2.b) Rediscovered cell of origin effect ( $\nu = 129$ ) .....   | 179 | Figure III.3.4.2) One of the largest quad-discovered GEP effects that presumably represents blood concentrations in tumor samples .....                   | 219 |
| Figure III.3.2.2.c) Significant enrichment of known ABC-versus-GCB DLBCL signatures by effect $\nu = 129$ .....                      | 181 | Figure III.3.4.3) A quad-discovered immunoglobulin effect.....  | 222 |
| Figure III.3.3.1.a) Validated effect $\nu = 134$ , applied to GSE31312.R-CHOP.....   | 184 |   |     |
| Figure III.3.3.1.b) Validated effect $\nu = 134$ , applied to GSE10846.R-CHOP.....   | 185 |   |     |
| Figure III.3.3.2) Validated effect $\nu = 127$ , applied to GSE31312 .....   | 190 |   |     |



# Bibliography

- [1] R.A. Meyers, Ed., "Encyclopedia of Complexity and Systems Science." New York, NY: Springer New York, 2009. (Public preface, available from: <http://link.springer.com/10.1007/978-0-387-30440-3>.)
- [2] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers and D.M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008", *Int. J. Cancer*, vol. 127, no. 12, pp. 2893–917, Dec. 2010.
- [3] S.-S. Wenzel, M. Grau, C. Mavis, S. Hailfinger, ... G. Lenz, "MCL1 is deregulated in subgroups of diffuse large B-cell lymphoma", *Leukemia*, vol. 27, no. 6, pp. 1381–90, Dec. 2013.
- [4] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, ... L.M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, vol. 403, no. 6769, pp. 503–11, Feb. 2000.
- [5] G. Lenz, G. Wright, S.S. Dave, W. Xiao, ... L.M. Staudt, "Stromal gene signatures in large-B-cell lymphomas", *N. Engl. J. Med.*, vol. 359, no. 22, pp. 2313–23, Nov. 2008.
- [6] P. Rubin, J.P. Williams, S.S. Devesa, L.B. Travis and L.S. Constine, "Cancer genesis across the age spectrum: associations with tissue development, maintenance, and senescence", *Semin. Radiat. Oncol.*, vol. 20, no. 1, pp. 3–11, Jan. 2010.
- [7] T.D. Gilmore, "Introduction to NF-kappaB: players, pathways, perspectives", *Oncogene*, vol. 25, no. 51, pp. 6680–4, Oct. 2006.
- [8] C.M. Grimaldi, R. Hicks and B. Diamond, "B Cell Selection and Susceptibility to Autoimmunity", *J. Immunol.*, vol. 174, no. 4, pp. 1775–1781, Feb. 2005.
- [9] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, "Molecular Biology of the Cell, Fourth Edition." New York, NY: Garland Science, 2002. (Chapter 24, The Adaptive Immune System. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK21070>.)
- [10] L.M. Staudt, "Oncogenic activation of NF- $\kappa$ B", *Cold Spring Harb. Perspect. Biol.*, vol. 2, no. 6, p. a000109, Jun. 2010.
- [11] G. Lenz, G.W. Wright, N.C.T. Emre, H. Kohlhammer, ... L.M. Staudt, "Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways", *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 36, pp. 13520–5, Sep. 2008.
- [12] M. Pfeifer, M. Grau, D. Lenze, S.-S. Wenzel, ... G. Lenz, "PTEN loss defines a PI3K/AKT pathway-dependent germinal center subtype of diffuse large B-cell lymphoma", *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 30, pp. 12420–5, Mar. 2013.
- [13] M.G. Kharas, R. Okabe, J.J. Ganis, M. Gozo, ... K. Gritsman, "Constitutively active AKT depletes hematopoietic stem cells and induces leukemia in mice", *Blood*, vol. 115, no. 7, pp. 1406–15, Feb. 2010.
- [14] F.S. Collins and H. Varmus, "A New Initiative on Precision Medicine", *N. Engl. J. Med.*, pp. 2012–2014, 2015.
- [15] M. a Cantrell and C.J. Kuo, "Organoid modeling for cancer precision medicine", *Genome Med.*, vol. 7, no. 1, pp. 158–160, 2015.
- [16] D. Rudnicka, A. Oszmiana, D.K. Finch, I. Strickland, ... D.M. Davis, "Rituximab causes a polarization of B cells that augments its therapeutic function in NK-cell-mediated antibody-dependent cellular cytotoxicity", *Blood*, vol. 121, no. 23, pp. 4694–702, Jun. 2013.
- [17] F. Oszolak and P.M. Milos, "RNA sequencing: advances, challenges and opportunities", *Nat. Rev. Genet.*, vol. 12, no. 2, pp. 87–98, Feb. 2011.
- [18] T.R. Mercer, M.E. Dinger, C.P. Bracken, G. Kolle, ... J.S. Mattick, "Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome", *Genome Res.*, vol. 20, no. 12, pp. 1639–50, Dec. 2010.
- [19] J. Beane, J. Vick, F. Schembri, C. Anderlind, ... A. Spira, "Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq", *Cancer Prev. Res. (Phila.)*, vol. 4, no. 6, pp. 803–17, Jun. 2011.
- [20] T. Steijger, J.F. Abril, P.G. Engström, F. Kokocinski, ... P. Bertone, "Assessment of transcript reconstruction methods for RNA-seq", *Nat. Methods*, vol. 10, no. 12, pp. 1177–84, Dec. 2013.
- [21] E. de Klerk, J.T. den Dunnen and P.A.C. 't Hoen, "RNA sequencing: from tag-based profiling to

- resolving complete transcript structure”, *Cell. Mol. Life Sci.*, vol. 71, no. 18, pp. 3537–51, May 2014.
- [22] J.N. Rouder, R.D. Morey, N. Cowan, C.E. Zwilling, C.C. Morey and M.S. Pratte, “An assessment of fixed-capacity models of visual working memory”, *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 16, pp. 5975–9, Apr. 2008.
- [23] A. Baddeley, “Working memory: looking back and looking forward”, *Nat. Rev. Neurosci.*, vol. 4, pp. 829–839, 2003.
- [24] J. Walcher, B. Groves, T. Budavári and D. Dale, “Fitting the integrated spectral energy distributions of galaxies”, *Astrophys. Space Sci.*, vol. 331, no. 1, pp. 1–51, Aug. 2010.
- [25] C.P. Ahn, R. Alexandroff, C.A. Prieto, F. Anders, ... G. Zhu, “The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment”, *Astrophys. J. Suppl. Ser.*, vol. 211, no. 2, p. 17, Jul. 2013.
- [26] I. Pâris, P. Petitjean, É. Aubourg, N.P. Ross, ... D.G. York, “The Sloan Digital Sky Survey quasar catalog: tenth data release”, *Astron. Astrophys.*, vol. 563, p. A54, Mar. 2014.
- [27] A.B. Olshen, E.S. Venkatraman, R. Lucito and M. Wigler, “Circular binary segmentation for the analysis of array-based DNA copy number data”, *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.
- [28] M. Grau, “Time-Dependent Flow around an Oscillating Sphere for Small Reynolds Numbers”, Diploma thesis, Department of Physics, Philipps-Universität Marburg, 2009.
- [29] C. Visco, Y. Li, Z.Y. Xu-Monette, R.N. Miranda, ... K.H. Young, “Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: a report from the International DLBCL Rituximab-CHOP Consortium”, *Leukemia*, vol. 26, no. November 2011, pp. 2103–2113, 2012.
- [30] A. Flörcken, M. Grau, A. Wolf, A. Weilemann, ... J. Westermann, “Gene expression profiling of peripheral blood mononuclear cells during treatment with a gene-modified allogeneic tumor cell vaccine in advanced renal cell cancer: Tumor-induced immunosuppression and a possible role for NF- $\kappa$ B”, *Int. J. Cancer*, Sep. 2014.
- [31] B. Kloo, D. Nagel, M. Pfeifer, M. Grau, ... D. Krappmann, “Critical role of PI3K signaling for NF- $\kappa$ B-dependent survival in a subset of activated B-cell-like diffuse large B-cell lymphoma cells”, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 1, pp. 272–7, Jan. 2011.
- [32] D. Nagel, S. Spranger, M. Vincendeau, M. Grau, ... D. Krappmann, “Pharmacologic inhibition of MALT1 protease by phenothiazines as a therapeutic approach for the treatment of aggressive ABC-DLBCL”, *Cancer Cell*, vol. 22, no. 6, pp. 825–37, Dec. 2012.
- [33] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: an overview”, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, Jan. 2012.
- [34] R.L. Graham and P. Hell, “On the history of the minimum spanning tree problem”, *Annals Of The History Of Computing*, vol. 7, pp. 43–57, 1985.
- [35] K. Florek, J. Łukaszewicz, J. Perkal, H. Steinhaus and S. Zubrzycki, “Sur la liaison et la division des points d’un ensemble fini”, *Colloq. Math.*, vol. 2, no. 3–4, pp. 282–285, 1951.
- [36] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, “Cluster analysis and display of genome-wide expression patterns”, *Proc. Natl. Acad. Sci. USA*, vol. 95, no. 25, pp. 14863–8, Dec. 1998.
- [37] A. Schmidt, M. Beck, J. Malmström, H. Lam, ... R. Aebersold, “Absolute quantification of microbial proteomes at different states by directed mass spectrometry”, *Mol. Syst. Biol.*, vol. 7, no. 1, p. 510, Jan. 2011.
- [38] J. Bohlin, E. Skjerve and D.W. Ussery, “Analysis of genomic signatures in prokaryotes using multinomial regression and hierarchical clustering”, *BMC Genomics*, vol. 10, p. 487, Jan. 2009.
- [39] S. Huang, N.L. Taylor, R. Narsai, H. Eubel, J. Whelan and A.H. Millar, “Experimental analysis of the rice mitochondrial proteome, its biogenesis, and heterogeneity”, *Plant Physiol.*, vol. 149, no. 2, pp. 719–34, Feb. 2009.
- [40] J.-Y. Zhang, Y.-C. Lee, I. Torres-Jerez, M. Wang, ... M.K. Udvardi, “Development of an integrated transcript sequence database and a gene expression atlas for gene discovery and analysis in

- switchgrass (*Panicum virgatum* L)", *The Plant Journal*, vol. 74, no. 1, pp. 160–73, Apr. 2013.
- [41] P.R. Wadia, N.J. Cabaton, M.D. Borrero, B.S. Rubin, ... A.M. Soto, "Low-dose BPA exposure alters the mesenchymal and epithelial transcriptomes of the mouse fetal mammary gland", *PLoS One*, vol. 8, no. 5, p. e63902, Jan. 2013.
- [42] M. Zhu, M. Yi, C.H. Kim, C. Deng, ... J.E. Green, "Integrated miRNA and mRNA expression profiling of mouse mammary tumor models identifies miRNA signatures associated with mammary tumor lineage", *Genome Biol.*, vol. 12, no. 8, p. R77, Jan. 2011.
- [43] C.B. Bigger, B. Guerra, K.M. Brasky, G. Hubbard, ... R.E. Lanford, "Intrahepatic gene expression during chronic hepatitis C virus infection in chimpanzees", *J. Virol.*, vol. 78, no. 24, pp. 13779–92, Dec. 2004.
- [44] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, ... D. Botstein, "Repeated observation of breast tumor subtypes in independent gene expression data sets", *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 14, pp. 8418–23, Jul. 2003.
- [45] A. Holleman, M.H. Cheok, M.L. den Boer, W. Yang, ... W.E. Evans, "Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment", *N. Engl. J. Med.*, vol. 351, no. 6, pp. 533–42, Aug. 2004.
- [46] A. Spira, J. Beane, V. Shah, G. Liu, ... J.S. Brody, "Effects of cigarette smoke on the human airway epithelial cell transcriptome", *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 27, pp. 10143–8, Jul. 2004.
- [47] J. Lapointe, C. Li, J.P. Higgins, M. van de Rijn, ... J.R. Pollack, "Gene expression profiling identifies clinically relevant subtypes of prostate cancer", *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 3, pp. 811–6, Jan. 2004.
- [48] Y. Pawitan, J. Bjöhle, L. Amler, A.-L. Borg, ... J. Bergh, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts", *Breast Cancer Res.*, vol. 7, no. 6, pp. R953–64, Jan. 2005.
- [49] G.Y. Lee, W.I. Yang, H.C. Jeung, S.C. Kim, ... S.Y. Rha, "Genome-wide genetic aberrations of thymoma using cDNA microarray based comparative genomic hybridization", *BMC Genomics*, vol. 8, p. 305, Jan. 2007.
- [50] S. Sridhar, F. Schembri, J. Zeskind, V. Shah, ... A. Spira, "Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium", *BMC Genomics*, vol. 9, p. 259, Jan. 2008.
- [51] E. Bignotti, A. Ravaggi, R. a Tassi, S. Calza, ... a D. Santin, "Trefoil factor 3: a novel serum marker identified by gene expression profiling in high-grade endometrial carcinomas", *Br. J. Cancer*, vol. 99, no. 5, pp. 768–73, Sep. 2008.
- [52] A.H. Beck, C.-H. Lee, D.M. Witten, B.C. Gleason, ... M. van de Rijn, "Discovery of molecular subtypes in leiomyosarcoma through integrative molecular profiling", *Oncogene*, vol. 29, no. 6, pp. 845–54, Feb. 2010.
- [53] L. Bullinger, M. Ehrich, K. Döhner, R.F. Schlenk, ... D. van den Boom, "Quantitative DNA methylation predicts survival in adult acute myeloid leukemia", *Blood*, vol. 115, no. 3, pp. 636–42, Jan. 2010.
- [54] D. Lindgren, A. Frigyesi, S. Gudjonsson, G. Sjö Dahl, ... M. Höglund, "Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome", *Cancer Res.*, vol. 70, no. 9, pp. 3463–72, May 2010.
- [55] A. Broyl, D. Hose, H. Lokhorst, Y. de Knecht, ... P. Sonneveld, "Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients", *Blood*, vol. 116, no. 14, pp. 2543–53, Oct. 2010.
- [56] E. Gardiner, N.J. Beveridge, J.Q. Wu, V. Carr, ... M.J. Cairns, "Imprinted DLK1-DIO3 region of 14q32 defines a schizophrenia-associated miRNA signature in peripheral blood mononuclear cells", *Mol. Psychiatry*, vol. 17, no. 8, pp. 827–40, Jul. 2012.
- [57] C. Hicks, L. Miele, T. Koganti, L. Young-Gaylor, ... G. Megason, "Analysis of Patterns of Gene Expression Variation within and between Ethnic Populations in Pediatric B-ALL", *Cancer Informatics*, vol. 12, pp. 155–73, Jan. 2013.
- [58] N. Dawany, L.C. Showe, A. V Kossenkova, C. Chang, ... L.J. Montaner, "Identification of a 251 gene expression signature that can accurately detect M tuberculosis in patients with and without HIV co-infection", *PLoS One*, vol. 9, no. 2, p. e89925, Jan. 2014.

- [59] F. Schreiber, D.J. Lynn, A. Houston, J. Peters, ... M.A. Gordon, "The human transcriptome during nontyphoid Salmonella and HIV coinfection reveals attenuated NFkappaB-mediated inflammation and persistent cell cycle disruption", *J. Infect. Dis.*, vol. 204, no. 8, pp. 1237–45, Oct. 2011.
- [60] S. Nagalla, J.W. Chou, M.C. Willingham, J. Ruiz, ... L.D. Miller, "Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis", *Genome Biol.*, vol. 14, no. 4, p. R34, Apr. 2013.
- [61] F.C. Chan, A. Telenius, S. Healy, S. Ben-Neriah, ... C. Steidl, "An RCOR1 loss-associated gene expression signature identifies a prognostically significant DLBCL subgroup", *Blood*, Nov. 2014.
- [62] R. Whitaker, M.P. Gil, F. Ding, M. Tatar, S.L. Helfand and N. Neretti, "Dietary switch reveals fast coordinated gene expression changes in *Drosophila melanogaster*", *Aging*, vol. 6, no. 5, pp. 355–68, May 2014.
- [63] R. Edgar, M. Domrachev and A.E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository", *Nucleic Acids Res.*, vol. 30, pp. 207–210, 2002.
- [64] K. Pearson, "On lines and planes of closest fit to systems of points in space", *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, pp. 559–572, 1901.
- [65] R. Bro and A.K. Smilde, "Principal component analysis", *Anal. Methods*, vol. 6, no. 9, p. 2812, Apr. 2014.
- [66] M. Ringnér, "What is principal component analysis?", *Nat. Biotechnol.*, vol. 26, no. 3, pp. 303–304, 2008.
- [67] H.-J. Kowalsky and G.O. Michler, "Lineare Algebra", 11. Auflage ed. Berlin: Walter de Gruyter GmbH, 1998.
- [68] O. Alter, P.O. Brown and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling", *Proc. Natl. Acad. Sci.*, vol. 97, no. 18, pp. 10101–10106, Aug. 2000.
- [69] M. Pfeifer, B. Zheng, T. Erdmann, H. Koeppen, ... G. Lenz, "Anti-CD22 and anti-CD79B antibody drug conjugates are active in different molecular diffuse large B-cell lymphoma subtypes", *Leukemia*, 2015.
- [70] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo and J.P. Mesirov, "Molecular signatures database (MSigDB) 30", *Bioinformatics*, vol. 27, no. 12, pp. 1739–40, Jun. 2011.
- [71] A.C. Culhane, M.S. Schröder, R. Sultana, S.C. Picard, ... J. Quackenbush, "GeneSigDB: a manually curated database and resource for analysis of gene expression signatures", *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1060–6, Jan. 2012.
- [72] A.L. Shaffer, G. Wright, L. Yang, J. Powell, ... L.M. Staudt, "A library of gene expression signatures to illuminate normal and pathological lymphoid biology", *Immunol. Rev.*, vol. 210, pp. 67–85, Apr. 2006.
- [73] R.L. Seal, S.M. Gordon, M.J. Lush, M.W. Wright and E.A. Bruford, "genenames.org: the HGNC resources in 2011", *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D514–9, Jan. 2011.
- [74] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, ... J.P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 15545–15550, 2005.
- [75] K. Heinig, M. Gätjen, M. Grau, V. Stache, ... U.E. Höpken, "Access to Follicular Dendritic Cells Is a Pivotal Step in Murine Chronic Lymphocytic Leukemia B-cell Activation and Proliferation", *Cancer Discov.*, vol. 4, no. 12, pp. 1448–65, Dec. 2014.
- [76] A. Weilemann, M. Grau, T. Erdmann, O. Merkel, ... G. Lenz, "Essential role of IRF4 and MYC signaling for survival of anaplastic large cell lymphoma", *Blood*, Oct. 2014.
- [77] M. Ashburner, C. a Ball, J. a Blake, D. Botstein, ... G. Sherlock, "Gene ontology: tool for the unification of biology The Gene Ontology Consortium", *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000.
- [78] E. Dimmer, R. Huntley, D. Barrell, D. Binns, ... R. Lovering, "The Gene Ontology - Providing a Functional Role in Proteomic Studies", vol. 25, no. 22, pp. 3045–3046, 2008.
- [79] B.R. Rosner, "Fundamentals of Biostatistics", 7/e, Inter. Cengage Learning, Inc, 2010.
- [80] S. Grossmann, S. Bauer, P.N. Robinson and M. Vingron, "Improved detection of overrepresentation of Gene-Ontology annotations

- with parent-child analysis”, *Bioinformatics*, vol. 23, no. 22, pp. 3024–3031, 2007.
- [81] H. Jing, J. Kase, J.R. Dörr, M. Milanovic, ... S. Lee, “Opposing roles of NF- $\kappa$ B in anti-cancer treatment outcome unveiled by cross-species investigations”, *Genes & Development*, vol. 25, no. 20, pp. 2137–46, Oct. 2011.
- [82] B. Andreopoulos, A. An, X. Wang and M. Schroeder, “A roadmap of clustering algorithms: finding a match for a biomedical application”, *Brief. Bioinform.*, vol. 10, no. 3, pp. 297–314, May 2009.
- [83] G. Wright, B. Tan, A. Rosenwald, E.H. Hurt, A. Wiestner and L.M. Staudt, “A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma”, *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 17, pp. 9991–6, Aug. 2003.
- [84] O. Burdakov, O. Sysoev, A. Grimvall and M. Hussian, “An  $O(n^2)$  Algorithm for Isotonic Regression”, in *Large-Scale Nonlinear Optimization SE - 3*, vol. 83, G. Di Pillo and M. Roma, Eds. Springer US, 2006, pp. 25–33.
- [85] I.N. Bronstein, K.A. Semendjajew, G. Musiol and H. Mühlig, “Taschenbuch der Mathematik”, vol. 5, no. 0. 2001.
- [86] X. Liu, M. Tanaka and M. Okutomi, “Single-image noise level estimation for blind denoising”, *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5226–5237, 2013.
- [87] M. Moussallam, A. Gramfort, L. Daudet and G. Richard, “Blind denoising with random greedy pursuits”, *IEEE Signal Process. Lett.*, vol. 21, pp. 1341–1345, 2014.
- [88] M.A. Stephens, “Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables”, *J. R. Stat. Soc. B*, vol. 32, no. 1, pp. 115–122, 1970.
- [89] R. Chicheportiche and J.P. Bouchaud, “Weighted Kolmogorov-Smirnov test: Accounting for the tails”, *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 86, pp. 1–7, 2012.
- [90] G. Box, “Non-normality and tests on variances”, *Biometrika*, vol. 40, no. 3, pp. 318–335, 1953.
- [91] S.S. Sawilowsky and R.C. Blair, “A more realistic look at the robustness and Type II error properties of the t test to departures from population normality”, *Psychol. Bull.*, vol. 111, no. 2, pp. 352–360, 1992.
- [92] T. Lumley, P. Diehr, S. Emerson and L. Chen, “The importance of the normality assumption in large public health data sets”, *Annu. Rev. Public Health*, vol. 23, pp. 151–169, 2002.
- [93] E.M. Pugh and G.H. Winslow, “The Analysis of Physical Measurements.” London: Addison-Wesley, 1966.
- [94] T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, ... A. Soboleva, “NCBI GEO: Archive for functional genomics data sets - Update”, *Nucleic Acids Res.*, vol. 41, 2013.
- [95] M. Hummel, S. Bentink, H. Berger, W. Klapper, ... R. Siebert, “A biologic definition of Burkitt’s lymphoma from transcriptional and genomic profiling”, *N. Engl. J. Med.*, vol. 354, no. 23, pp. 2419–30, Jun. 2006.
- [96] N. Waddell, S. Cocciardi, J. Johnson, S. Healey, ... G. Chenevix-Trench, “Gene expression profiling of formalin-fixed, paraffin-embedded familial breast tumours using the whole genome-DASL assay”, *J. Pathol.*, vol. 221, no. 4, pp. 452–461, 2010.
- [97] G. Liu, A.E. Loraine, R. Shigeta, M. Cline, ... M. a. Siani-Rose, “NetAffx: Affymetrix probesets and annotations”, *Nucleic Acids Research*, vol. 31, no. 1. pp. 82–86, 2003.
- [98] D.R. Cox, “Regression models and life tables”, *J. R. Stat. Soc. Ser. B*, vol. 34, pp. 187–220, 1972.
- [99] S.S. Wilks, “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”, *Ann. Math. Stat.*, vol. 9, no. 1, pp. 60–62, Mar. 1938.
- [100] The International Non-Hodgkin’s Lymphoma Prognostic Factors Project, “A predictive model for aggressive non-Hodgkin’s lymphoma”, 1993.
- [101] T.M. Habermann, F. Hong, V.A. Morrison, S.R. Dakhil, ... S.J. Horning, “Differences in Outcomes in Males and Females with Diffuse Large B-Cell Lymphoma with Induction Rituximab and Follicular Lymphoma Treated with Maintenance Rituximab”, *ASH Annu. Meet. Abstr.*, vol. 120, no. 21, p. 3705, Nov. 2012.
- [102] M. Van Keimpema, L.J. Gr. M. Mokry, R. Van Boxtel, ... M. Spaargaren, “FOXP1 directly represses transcription of proapoptotic genes and

- cooperates with NF- $\kappa$ B to promote survival of human B cells”, vol. 124, no. 23, pp. 3431–3441, 2015.
- [103] C. Huisman, G.B. a Wisman, H.G. Kazemier, M. a T.M. van Vugt, ... M.G. Rots, “Functional validation of putative tumor suppressor gene C13ORF18 in cervical cancer by Artificial Transcription Factors”, *Mol. Oncol.*, vol. 7, no. 3, pp. 669–679, 2013.
- [104] E. Cubedo, A.J. Gentles, C. Huang, Y. Natkunam, ... I.S. Lossos, “Identification of LMO2 transcriptome and interactome in diffuse large B-cell,lymphoma”, *Blood*, vol. 119, no. 23, pp. 5478–5491, 2012.
- [105] T.M. Shams, “High expression of LMO2 in Hodgkin, Burkitt and germinal center diffuse large B cell lymphomas”, *J. Egypt. Natl. Canc. Inst.*, vol. 23, no. 4, pp. 147–153, 2011.
- [106] K.D. Pruitt, T. Tatusova and D.R. Maglott, “NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”, *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D501–4, Jan. 2005.
- [107] D. Maglott, J. Ostell, K.D. Pruitt and T. Tatusova, “Entrez Gene: gene-centered information at NCBI”, *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D54–8, Jan. 2005.
- [108] S.K. Low, S. Chung, A. Takahashi, H. Zembutsu, ... Y. Nakamura, “Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan”, *Cancer Sci.*, vol. 104, no. 8, pp. 1074–1082, 2013.
- [109] N. Gupta-Rossi, S. Storck, P.J. Griebel, C.A. Reynaud, J.C. Weill and A. Dahan, “Specific over-expression of deltex and a new Kelch-like protein in human germinal center B cells”, *Mol. Immunol.*, vol. 39, no. 13, pp. 791–799, 2003.
- [110] D.E.K. Tan, J.N. Foo, J.-X. Bei, J. Chang, ... J. Liu, “Genome-wide association study of B cell non-Hodgkin lymphoma identifies 3q27 as a susceptibility locus in the Chinese population”, *Nat. Genet.*, vol. 45, no. 7, pp. 804–807, 2013.
- [111] R. Watanabe, N. Tomita, C. Matsumoto, Y. Hattori, ... Y. Ishigatsubo, “Prognostic value of the 3q27 and 18q21 translocations for diffuse large B-cell lymphoma and follicular lymphoma in the rituximab era”, *J. Clin. Oncol.*, vol. 29, no. 15, pp. 3–9, 2011.
- [112] L. a Ramkissoon, P.M. Horowitz, J.M. Craig, S.H. Ramkissoon and B.E. Rich, “Genomic analysis of diffuse pediatric low-grade gliomas identifies recurrent oncogenic truncating rearrangements in the transcription factor MYBL1”, pp. 1–6.
- [113] S. Ichikawa, N. Fukuhara, H. Katsushima, T. Takahashi, ... H. Harigae, “Association between BACH2 expression and clinical prognosis in diffuse large B-cell lymphoma”, *Cancer Sci.*, vol. 105, no. 4, pp. 437–444, 2014.
- [114] W.-J. Cao, H.-L. Wu, B.-S. He, Y.-S. Zhang and Z.-Y. Zhang, “Analysis of long non-coding RNA expression profiles in gastric cancer”, *World J. Gastroenterol.*, vol. 19, no. 23, pp. 3658–64, 2013.
- [115] Q. Pang, J. Ge, Y. Shao, W. Sun, ... J. Guo, “Increased expression of long intergenic non-coding RNA LINC00152 in gastric cancer and its clinical significance”, *Tumor Biol.*, vol. 35, no. 6, pp. 5441–5447, 2014.
- [116] H. Tani and M. Torimura, “Identification of short-lived long non-coding RNAs as surrogate indicators for chemical stress response”, *Biochem. Biophys. Res. Commun.*, vol. 439, no. 4, pp. 547–551, 2013.
- [117] F. Lühder, C. Chambers, J.P. Allison, C. Benoist and D. Mathis, “Pinpointing when T cell costimulatory receptor CTLA-4 must be engaged to dampen diabetogenic T cells”, *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 22, pp. 12204–12209, 2000.
- [118] D.J. Lenschow, T.L. Walunas and J. a Bluestone, “CD28/B7 system of T cell costimulation”, *Annu. Rev. Immunol.*, vol. 14, pp. 233–258, 1996.
- [119] Q.P. Vong, W.-H. Leung, J. Houston, Y. Li, ... W. Leung, “TOX2 regulates human natural killer cell development by controlling T-BET expression”, *Blood*, vol. 124, no. 26, pp. 3905–3913, 2014.
- [120] C. Correia, P.A. Schneider, H. Dai, A. Dogan, ... S.H. Kaufmann, “BCL2 mutations are associated with increased risk of transformation and shortened survival in follicular lymphoma”, *Blood*, vol. 125, no. 4, pp. 658–668, 2014.
- [121] M.H. Kang and C.P. Reynolds, “Bcl-2 inhibitors: targeting mitochondrial apoptotic pathways in cancer therapy”, *Clin. Cancer Res.*, vol. 15, no. 4, pp. 1126–1132, 2009.
- [122] J.W. Friedberg, “Double-hit diffuse large B-cell lymphoma”, *J. Clin. Oncol.*, vol. 30, no. 28, pp. 3439–3443, 2012.

- [123] M. Karin, R.L. Eddy, W.M. Henry, L.L. Haley, M.G. Byers and T.B. Shows, "Human metallothionein genes are clustered on chromosome 16", *Proc. Natl. Acad. Sci. USA*, vol. 81, no. 17, pp. 5494–5498, 1984.
- [124] D. Magda, P. Lecane, Z. Wang, W. Hu, ... J.L. Sessler, "Synthesis and anticancer properties of water-soluble zinc ionophores", *Cancer Res.*, vol. 68, no. 13, pp. 5318–5325, 2008.
- [125] A. Krężel and W. Maret, "Dual nanomolar and picomolar Zn(II) binding properties of metallothionein", *J. Am. Chem. Soc.*, vol. 129, no. 35, pp. 10911–10921, 2007.
- [126] H. Haase and L. Rink, "Functional significance of zinc-related signaling pathways in immune cells", *Annu. Rev. Nutr.*, vol. 29, pp. 133–152, 2009.
- [127] D. Chaussabel, C. Quinn, J. Shen, P. Patel, ... V. Pascual, "A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus", *Immunity*, vol. 29, no. 1, pp. 150–164, 2008.

#### ■ Software resources

All analyses were performed with and all programs were implemented in MATLAB® (R2013b, The MathWorks®, Natick, Massachusetts, United States). Tabular results have been exported using Office Excel® (2013, Microsoft®, Redmond, Washington, United States). Most plots were also produced with MATLAB® and sometimes post-processed using CorelDRAW® (X7, Corel Corporation, Ottawa, ON, Canada). This dissertation was authored using Office Word (2013, Microsoft®, Redmond, Washington, United States). Citations were managed using Mendeley Desktop (v1.13.8, Mendeley Ltd., London, United Kingdom).





# Scientific Profile



## ■ University Education

2015-08-03 **Final doctoral examination** and defense of this dissertation

Thesis Dissection of Complex Genetic Correlations into Interaction Effects

Supervisor Prof. Dr. Peter Lenz, Complex Systems Group, Department of Physics, Philipps-Universität Marburg

Final grade Summa cum laude

2010-2011 Expanding into biostatistics, bioinformatics and molecular biology (see publications below)

2009-09-25 **Diploma in Physics**

Thesis Time-Dependent Flow around an Oscillating Sphere for Small Reynolds Numbers

Supervisor Prof. Dr. Bruno Eckhardt, Complex Systems Group, Department of Physics, Philipps-Universität Marburg

Exam subjects Quantum mechanics of many-particle systems and statistical physics; electrodynamics; atom, nuclear and solid state physics; fluid dynamics and computational physics

Final grade Very good, with distinction

2007-07-25 **Diploma in Computer Science**

Thesis System zur Definition und Online-Erkennung handschriftlicher Symbole und Strukturen

Supervisor Prof. Dr. Bernhard Seeger, Database Research Group, Department of Mathematics and Computer Science, Philipps-Universität Marburg

Exam subjects Knowledge discovery and temporal data mining; universal co-algebra and computer-supported mathematical proving systems; database systems; classical mechanics and single-particle quantum mechanics

Final grade Very good, with distinction

## ■ Publications (chronological)

1. B. Kloo, D. Nagel, M. Pfeifer, M. Grau, M. Düwel, M. Vincendeau, B. Dörken, P. Lenz, G. Lenz and D. Krappmann, "Critical role of PI3K signaling for NF- $\kappa$ B-dependent survival in a subset of activated B-cell-like diffuse large B-cell lymphoma cells", *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 1, pp. 272–7, Jan. 2011.
2. B.A. Ratsch, M. Grau, B. Dörken, P. Lenz and G. Lenz, "The use of microarray technologies in mantle cell lymphoma", *Semin. Hematol.*, vol. 48, no. 3, pp. 166–71, Jul. 2011.
3. S. Hailfinger, H. Nogai, C. Pelzer, M. Jaworski, K. Cabalzar, J.-E. Charton, M. Guzzardi, C. Décaillet, M. Grau, B. Dörken, P. Lenz, G. Lenz and M. Thome, "Malt1-dependent RelB cleavage promotes canonical NF- $\kappa$ B activation in lymphocytes and lymphoma cell lines", *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 35, pp. 14596–601, Aug. 2011.
4. H. Jing, J. Kase, J.R. Dörr, M. Milanovic, D. Lenze, M. Grau, G. Beuster, S. Ji, M. Reimann, P. Lenz, M. Hummel, B. Dörken, G. Lenz, C. Scheidereit, C.A. Schmitt and S. Lee, "Opposing roles of NF- $\kappa$ B in anti-cancer treatment outcome unveiled by cross-species investigations", *Genes & Development*, vol. 25, no. 20, pp. 2137–46, Oct. 2011.
5. D. Nagel, S. Spranger, M. Vincendeau, M. Grau, S. Raffegerst, B. Kloo, D. Hlahla, M. Neuenschwander, J. Peter von Kries, K. Hadian, B. Dörken, P. Lenz, G. Lenz, D.J. Schendel and D. Krappmann, "Pharmacologic inhibition of MALT1 protease by phenothiazines as a therapeutic approach for the treatment of aggressive ABC-DLBCL", *Cancer Cell*, vol. 22, no. 6, pp. 825–37, Dec. 2012.
6. M. Pfeifer, M. Grau, D. Lenze, S.-S. Wenzel, A. Wolf, B. Wollert-Wulf, K. Dietze, H. Nogai, B. Storek, H. Madle, B. Dörken, M. Janz, S. Dirnhofer, P. Lenz, M. Hummel, A. Tzankov and G. Lenz, "PTEN loss defines a PI3K/AKT pathway-dependent germinal center subtype of diffuse large B-cell lymphoma", *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 30, pp. 12420–5, Mar. 2013.

7. H. Nogai, S.S. Wenzel, S. Hailfinger, M. Grau, E. Kaergel, V. Seitz, B. Wollert-Wulf, M. Pfeifer, A. Wolf, M. Frick, K. Dietze, H. Madle, A. Tzankov, M. Hummel, B. Dörken, C. Scheidereit, M. Janz, P. Lenz, M. Thome and G. Lenz, "I $\kappa$ B-zeta controls the constitutive NF- $\kappa$ B target gene network and survival of ABC DLBCL", *Blood*, vol. 122, no. 13, pp. 2242–2250, Sep. 2013.
8. S.-S. Wenzel, M. Grau, C. Mavis, S. Hailfinger, A. Wolf, H. Madle, G. Deeb, B. Dörken, M. Thome, P. Lenz, S. Dirnhofer, F.J. Hernandez-Ilizaliturri, A. Tzankov and G. Lenz, "MCL1 is deregulated in subgroups of diffuse large B-cell lymphoma", *Leukemia*, vol. 27, no. 6, pp. 1381–90, Dec. 2013.
9. J.C. Costello, L.M. Heiser, E. Georgii, M. Gönen, M.P. Menden, N.J. Wang, M. Bansal, M. Ammad-Ud-Din, P. Hintsanen, S.A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, NCI DREAM Community, J.J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S. Kaski, J.W. Gray and G. Stolovitzky, "A community effort to assess and improve drug sensitivity prediction algorithms", *Nat. Biotechnol.*, Jun. 2014 (as member of the NCI-DREAM Community).
10. A. Flörcken\*, M. Grau\*, A. Wolf\*, A. Weilemann, J. Kopp, B. Dörken, T. Blankenstein, A. Pezzutto, P. Lenz, G. Lenz\* and J. Westermann\*, "Gene expression profiling of peripheral blood mononuclear cells during treatment with a gene-modified allogeneic tumor cell vaccine in advanced renal cell cancer: Tumor-induced immunosuppression and a possible role for NF- $\kappa$ B", *Int. J. Cancer*, Sep. 2014. (\*)contributed equally.
11. S. Kreher, M.A. Bouhlel, P. Cauchy, B. Lamprecht, S. Li, M. Grau, F. Hummel, K. Köchert, I. Anagnostopoulos, K. Jöhrens, M. Hummel, J. Hiscott, S.-S. Wenzel, P. Lenz, M. Schneider, R. Küppers, C. Scheidereit, M. Giefing, R. Siebert, K. Rajewsky, G. Lenz, P.N. Cockerill, M. Janz, B. Dörken, C. Bonifer and S. Mathas, "Mapping of transcription factor motifs in active chromatin identifies IRF5 as key regulator in classical Hodgkin lymphoma", *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 42, pp. E4513–22, Oct. 2014.
12. A. Weilemann, M. Grau, T. Erdmann, O. Merkel, U. Sobhiafshar, I. Anagnostopoulos, M. Hummel, A. Siegert, C. Hayford, H. Madle, B. Wollert-Wulf, I. Fichtner, B. Dörken, S. Dirnhofer, S. Mathas, M. Janz, N.C.T. Emre, A. Rosenwald, G. Ott, P. Lenz, A. Tzankov and G. Lenz, "Essential role of IRF4 and MYC signaling for survival of anaplastic large cell lymphoma", *Blood*, Oct. 2014.
13. K. Heinig, M. Gätjen, M. Grau, V. Stache, I. Anagnostopoulos, K. Gerlach, R.A. Niesner, Z. Cseresnyes, A.E. Hauser, P. Lenz, T. Hehlhans, R. Brink, J. Westermann, B. Dörken, M. Lipp, G. Lenz, A. Rehm and U.E. Höpken, "Access to Follicular Dendritic Cells Is a Pivotal Step in Murine Chronic Lymphocytic Leukemia B-cell Activation and Proliferation", *Cancer Discov.*, vol. 4, no. 12, pp. 1448–65, Dec. 2014.
14. M. Pfeifer, B. Zheng, T. Erdmann, H. Koeppen, R. McCord, M. Grau, A. Staiger, A. Chai, T. Sandmann, H. Madle, B. Dörken, Y.W. Chu, a I. Chen, D. Lebovic, G. a Salles, M.S. Czuczman, M.C. Palanca-Wessels, O.W. Press, R. Advani, F. Morschhauser, B.D. Cheson, P. Lenz, G. Ott, a G. Polson, K.E. Mundt and G. Lenz, "Anti-CD22 and anti-CD79B antibody drug conjugates are active in different molecular diffuse large B-cell lymphoma subtypes", *Leukemia*, Feb. 2015 (advanced online publication).

# Acknowledgements

---

By being there for me, by teaching me or by providing motivating surrounding conditions, many people made this work in physics and cancer research possible. Collectively, they enabled me to reach my personal aims attached to it. Namely, to produce not “just” a sound theoretical result but also one that is and *feels* useful for practice and, thus, can motivate me and hopefully others for future research. For that, I am very grateful to all these people, although only some of them can be named here.

First, I profoundly want to thank my PhD supervisor Peter Lenz for his guidance, his feedback and his calmness, especially when certain hard-to-solve mathematical optimization problems emerged during algorithm development. In particular, I am grateful for opened doors towards participation in the 62<sup>nd</sup> Lindau Nobel Laureate Meeting dedicated to Physics. This meeting was even more special than could be expected because of the discovery of a new boson that was published and discussed with Nobel laureates in a CERN panel discussion during this meeting. This boson turned out to be the long sought-after Higgs Boson that was theoretically predicted *in 1964*. A unique moment no theoretical physicist would want to miss, many thanks! Furthermore, I am thankful for his guidance when I applied for a three-year doctoral scholarship and want to use this opportunity to also thank my home university and the Marburg University Research Academy for considering me worthy of it.

Likewise, I deeply thank Georg Lenz for biological insight, his feedback and his support. I am especially thankful for providing me with the opportunity to take part in many interesting research projects in molecular biology, thereby pushing me nearer towards nature. I am particularly grateful for my lab time in the Charité that was stimulating and I am looking forward to getting even nearer towards cells and nature in the near future. Together with Peter Lenz, he realized a close and fruitful collaboration between theory and experiment. Both quickly guided me towards taking part in ongoing research tasks. Additionally, both provided me with the time and freedom to develop a substantial contribution of my own in form of this work. It is exciting to analyze biological data in a context that is not far away from medicine and that works towards meaningful clinical applications; thank you for this opportunity!

There are far too many teachers in my life to name them all, but a few more deserve to be named here because they bestowed me with major qualifications during my education before venturing towards biology. In chronological order: I thank Peter Klar, who showed me first steps in an experimental semiconductors lab and let me experiment with quantum dots, lasers and Raman spectroscopy. I thank Alfred Ultsch, who laid the foundation for my skills in clustering, knowledge discovery and data mining. I thank Hans-Peter Gumm, especially for showing me a bird’s eye view of math via higher-order logic and computer-assisted proofing. I thank Bernhard Seeger for the opportunity to realize my ideas of mathematical handwriting recognition as Diploma thesis in computer science. Last but not least, I want to express profound thanks to Bruno Eckhardt for sharpening my analytic eye in nonlinear dynamics and my physical eye for good approximations during my Diploma thesis in Physics.

Whenever I was in our Complex Systems group in Marburg and whenever I was in our molecular biology labs in Berlin or Münster there was not only a productive and cooperative but also an amicable atmosphere. For that I am thankful to André, Claudia, Hannelore, Kristian, Lisa, Nicole and Tabea as well as to Alex, Kevin, Konstanze, Kristof, Lennart, Marcus, Myroslav, Patrick and Shuxin and finally to Annika, Felix, Florian, Hannes,

Jan, Jens, Konstantin, Lars, Lisa, Marina, Michael, Stefan and Tobias. I am especially grateful to André, Lisa and Tabea for proof-reading biological parts of my PhD and for their useful feedback; thanks!

Finally, and most importantly, I want to thank my family and my friends. First and foremost, I thank my parents for their confidence and support throughout the course of my studies at all times and for the kind-hearted and loving people they are. Special thanks go to Carina and Torben for proof-reading parts of my work and for their useful feedback. Genuine thanks go to all of you for our shared time outside of science, although it was too scarce from my side, for which I hereby apologize.



