2016

Developing genomic models for cancer prevention and treatment stratification

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

DEVELOPING GENOMIC MODELS FOR CANCER PREVENTION AND TREATMENT STRATIFICATION

by

DANIEL GUSENLEITNER

B.S., University of Applied Sciences, Hagenberg, Austria, 2008
M.S., University of Skövde, Sweden, 2009
M.S., University of Applied Sciences, Hagenberg, Austria, 2010

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2016

Approved by

First Reader ____
Stefano Monti, Ph.D.
Associate Professor of Medicine

Second Reader ____
Marc Lenburg, Ph.D.
Associate Professor of Medicine

DEDICATION

To my father Markus,

who set me on the path that

eventually led me here to Boston.

ACKNOWLEDGMENTS

To my advisor – Dr. Stefano Monti: Both on a professional and on a personal level I have only deepest respect and admiration for you. Thank you for all your patience, motivation and the immense knowledge that you taught me over the last 4 years. I could not have wished for a better PhD adviser.

To my family, in particular Markus and Regina: No matter how far I stray from home you give me the feeling that there will always be a place that I can come back to. Thank you, for instilling me with the drive to use my potential, and the restless thirst for gaining knowledge.

To my extended family – Harold, Nacho and Tom: You are the brothers I never had. Each one of you significantly contributed to the person I am today. I admire your curiosity, passion and drive and I consider myself very fortunate that you are part of my life.

To my PhD thesis committee – in particular Björn Chapuy: Thank you for your insightful comments, your encouragement, but also your hard questions which incented me to consider my research from very different angles.

To my fellow labmates – Liye, Francesca, Vinay, Amy and Yuxiang: It has been a pleasure to work with such a great group of brilliant scientists. You made me feel like I was working among friends not only colleagues.

To my friends – Everyone from bubify and in particular Yözen, Joshua and Ania: Thank you for providing a never ending stream of intellectual stimulus. Each of you taught me something different about the world.

To the bioinformatics program – in particular Tom Tullius, Scott Mohr, Gary Benson, Caroline Lyman, Johanna Vasquez, and David King. Thank you for your tireless work; it is the reason why this program is so outstanding.

To my love – Margaret Amelia Ouzts-Gusenleitner: Ever since I have met you, my life has turned around for the better. You are always there for me when I need you and you accept me for the quirky person I am. You keep me grounded and give me the feeling you care, even for the smallest things. I am excited to spend the rest of my life with you. I love you.

DEVELOPING GENOMIC MODELS FOR CANCER PREVENTION AND TREATMENT STRATIFICATION

(Order No.)

DANIEL GUSENLEITNER

Boston University Graduate School of Arts and Sciences,

and

College of Engineering 2016

Major Professor: Stefano Monti, Associate Professor of Medicine

ABSTRACT

Malignant tumors remain one of the leading causes of mortality with over 8.2 million deaths worldwide in 2012. Over the last two decades, high-throughput profiling of the human transcriptome has become an essential tool to investigate molecular processes involved in carcinogenesis. In this thesis I explore how gene expression profiling (GEP) can be used in multiple aspects of cancer research, including prevention, patient stratification and subtype discovery.

The first part details how GEP could be used to supplement or even replace the current gold standard assay for testing the carcinogenic potential of chemicals. This toxicogenomic approach coupled with a Random Forest algorithm allowed me to build models capable of predicting carcinogenicity with an area under the curve of up to 86.8%

and provided valuable insights into the underlying mechanisms that may contribute to cancer development.

The second part describes how GEP could be used to stratify heterogeneous populations of lymphoma patients into therapeutically relevant disease sub-classes, with a particular focus on diffuse large B-cell lymphoma (DLBCL). Here, I successfully translated established biomarkers from the Affymetrix platform to the clinically relevant Nanostring nCounter© assay. This translation allowed us to profile custom sets of transcripts from formalin-fixed samples, transforming these biomarkers into clinically relevant diagnostic tools.

Finally, I describe my effort to discover tumor samples dependent on altered metabolism driven by oxidative phosphorylation (OxPhos) across multiple tissue types. This work was motivated by previous studies that identified a therapeutically relevant OxPhos subtype in DLBCL, and by the hypothesis that this stratification might be applicable to other solid tumor types. To that end, I carried out a transcriptomics-based pan-cancer analysis, derived a generalized PanOxPhos gene signature, and identified mTOR as a potential regulator in primary tumor samples.

High throughput GEP coupled with statistical machine learning methods represent an important toolbox in modern cancer research. It provides a cost effective and promising new approach for predicting cancer risk associated to chemical exposure, it can reduce the cost of the ever increasing drug development process by identifying therapeutically actionable disease subtypes, and it can increase patients' survival by matching them with the most effective drugs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2YRB2	2 year rodent bioassay
AUC	Area under the receiver operating characteristic curve
BCR	B-Cell receptor signaling
BRCA	Breast Cancer
CCC	Comprehensive Consensus Clustering
CCLE	Cancer Cell-line Encyclopedia
cMap	Connectivity Map
CPDB	Carcinogenic Potency Database
DEG	differentially expressed genes
DLBCL	Diffuse Large B-Cell Lymphoma
FC	fold change
FDR	False discovery rate
FISH	Fluorescence In Situ Hybridization
FN	False negative
FP	False positive
GSEA	Gene set enrichment analysis
HNSC	Head and Neck Squamous Carcinoma
IHC	Immunohistochemistry
KS	Kolmogorov-Smirnov
HR	Host response

LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Carcinoma
MSigDB	Molecular Signature Database
OCR	Oxygen Consumption Rate
OxPhos	Oxidative Phosphorylation
PCA	Principal component analysis
RF	Random forest algorithm
ROC	Receiver operating characteristic
ROS	Reactive Oxygen Species
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TG-GATEs	Toxicogenomics project- Genomics Assisted Toxicity Evaluations
TN	True negative
TP	True positive

CHAPTER 1 - INTRODUCTION, BACKGROUND, RATIONALE

1.1 Transcriptional profiling

1.1.1 Rational of quantifying gene expression

The central dogma of molecular biology describes the production of proteins from DNA. In short, a specific sequence of DNA, also known as a protein-coding gene, is transcribed into pre-mRNA by RNA polymerase. This mRNA is then spliced, and a 5’ cap and a 3’ poly-adenine tail are added. Finally, the mature mRNA species is translated into proteins by the ribosome. The set of all proteins is called the proteome, which provides a snapshot of the biological processes within a cell. Unfortunately, it is still not feasible to measure the level of all proteins, so the next best alternative is to quantify the levels of all mRNA species, which make up the transcriptome. Even though there is not always a strong correlation between the levels of mRNA and corresponding proteins due to post transcriptional gene regulatory events (Greenbaum et al. 2003), quantification of mRNA abundance still provides biologically relevant information, such as differences between molecular cancer subtypes or biological processes that are activated as response to treatment with chemical compounds. The following section will present three different methods to measure the transcriptome.

1.1.2 Oligonucleotide Microarrays

There is a variety of different types of gene expression microarrays. This section will focus on the most popular Affymetrix gene arrays since these were used for transcription quantification in rats in Chapter 2 and in humans in Chapter 4 and 5.

Affymetrix GeneChips™ are composed of hundreds of thousands of probes, each of which is complementary to a specific mRNA sequence. Each type of probe contains a unique 25-mer DNA oligonucleotide, which is synthesized onto the microarray using photolithographic synthesis. For a microarray experiment, mRNA from a biological sample is extracted, purified, fluorescently labeled and applied onto a GeneChip. Each mRNA species can only hybridize onto an exact complementary oligonucleotide. After hybridization, mRNA that could not attach to a complementary probe is washed away and the chip is scanned. During this process, probes that contain labelled hybridized mRNA emit light and the intensity of this emission can be used to determine the amount of hybridization of a particular mRNA species (Heinrich Goehlmann & Talloen 2009).

During standard Affymetrix preprocessing, 14 unique probes are summarized into a probe set and each gene is represented by one or more of these probe sets. However, the GeneChips used in the experiments throughout the thesis (Human Gene U133A/B/Plus2.0 and Rat Gene 230.2) were designed with a dated genome annotation, which have issues such as pseudogenes and probes that include SNPs. Thus, in order to ameliorate these issues and also to avoid the one-to-many mapping between probe sets and genes, custom chip definition files from Brainarray (Dai et al. 2005) were used to summarize the probes directly to gene levels throughout this thesis.

1.1.3 RNA Sequencing

Microarrays are heavily dependent on the quality of the genome annotations. There are an approximate 500,000 to 2,000,000 common SNP in the human genome, and probes designed in regions encompassing these SNP will result in misleading estimation of transcript abundance levels (Siu et al. 2011). Next generation sequencing does not suffer from this issue.

While RNA-Seq does make use of reference genomes or transcriptomes, it allows for errors in the alignment between pieces of transcripts known as sequencing reads and the reference. Generally, microarray technology allows only the detection of the specific analytes, which it was designed for, whereas RNA-Sequencing is more flexible and is better suited to answer open ended research questions. Thus, RNASeq has a much broader range of applications, such as the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusions, mutations/SNPs and changes in gene expression (Maher et al. 2009). Furthermore, current RNA-sequencing methods allow not only the quantification of messenger RNA level, they also give insight into total RNA, non-coding RNA, micro RNA, transfer RNA, and ribosomal RNA (Ingolia et al. 2012).

In a typical workflow of RNA sequencing, mRNA is converted into a library of cDNA fragments, by either RNA fragmentation and transcription into cDNA using reverse transcriptase or by using transcriptase first and DNA fragmentation afterwards. Then sequencing adaptors are added to each cDNA fragment and the first N (typically 50-100) base pairs are determined using high throughput sequencing technology. For paired end sequencing the fragment is sequenced from both sides, which allows better identification of splice junctions. The resulting sequence reads are aligned onto a reference genome or transcriptome and the number of aligned reads can be used to quantify the expression level of each gene (Wang et al. 2009).

1.1.4 Nanostring nCounter©

Both oligonucleotide microarrays and RNA-sequencing are well suited as research tools for the discovery of biological mechanisms of actions or the discovery of molecular subtypes. However, for clinical applications they are often too costly, and more importantly they are not reliably reproducible as they suffer from considerable batch effects (Su et al. 2014; Tillinghast 2010; Shi et al. 2010). The Nanostring nCounter© platform does not suffer from the same issues, but it is limited to multiplexing only up to 800 genes (Geiss et al. 2008). A standard workflow for gene expression-based analysis is to use full transcriptome quantification assays to identify a set of genes, a gene signature, that can serve as a clinical diagnostic or prognostic indicator. Typical signatures are usually in the range of dozens to few hundreds transcripts (e.g. (Lamond et al. 2013; Biroschak et al. 2013; Smaglo et al. 2015)), which can then be measured using the nCounter system. nCounter can measure gene expression without amplification or cloning, can detect gene expression from as little as 300 ng mRNA, works well with mRNA extracted from paraffin-embedded formalin-fixed (FFPE) samples, and is therefore ideal in clinical settings, where the amount of RNA is very limited since it is sourced from tissue biopsies (Geiss et al. 2008).

Similar to RNA-Sequencing, nCounter provides a digital readout of the amount of a transcript in a sample. The levels of each transcript can be established by counting the number of molecules of each sequence type and calculating concentration with reference to internal standards (Geiss et al. 2008). To account for batch effects, nCounter analyses are performed in batches of 12 samples, where one of the 12 samples contains standardized spike-in oligonucleotide samples that can be used to normalize across batches.

The nCounter platform relies on probes but gives a direct quantification of mRNA species measured. During that process a target mRNA of interest is hybridized to both capture probe and reporter probe. Both of these two probes have a gene specific sequence that is complimentary to the mRNA. The capture probe is able to bind onto the nCounter cartridge, while the reporter probe contains a barcode that is specific to the target mRNA. After hybridization, the excess probes are removed. The purified probe/target complex is bound onto a streptavidin-coated slide via biotinylated capture probes and electrophoresis is used to elongate and align the molecules. Biotinylated anti-5' oligonucleotides that hybridize to the 5'-repeat sequence are added. The stretched reporters are immobilized by the binding of the anti-5' oligonucleotides to the slide surface via the biotin. Voltage is turned off and the immobilized reporters are prepared for imaging and counting (Geiss et al. 2008).

1.2 Supervised Classification methods

Chapters 2-4 all heavily rely on supervised classification methods, which are introduced in this section.

1.2.1 Random Forest

Random forests were developed by (Breiman 2001) and are based on decision trees. In these trees each leaf represents a specific class label, whereas branches represent conjunctions of features that lead to those class labels. Each node within the tree splits samples based on a feature that is chosen so that entropy of class correspondence within the splits is minimized. Decision trees are popular because they are straightforward to visualize, where the impact of each feature is immediately apparent. However, decision trees tend to overfit training sets and do poorly on independent test sets (Hastie et al. 2009).

Random forests are a class of ensemble learning methods, which are able to avoid this kind of overfitting. See Figure 1.1 for a graphical representation. Specifically, the random forest algorithm generates a multitude of different decision trees, each of which can predict the class label and then uses the most common of all individual classes to derive a final class. In the case of binary classification the mean outcome (0, 1) of all decision trees can be used to calculate class probability. The differences in the individual decision trees are achieved by bagging and by random feature selection. Bagging is a machine learning method that generates new trainings sets of the same size by uniform sampling with replacement (Breiman 1996), whereas the random feature selection is applied to the selection of each node within each decision tree. Instead of choosing the best of all features, it chooses the best feature from a random subsample that is usually the size of the square root of all features.

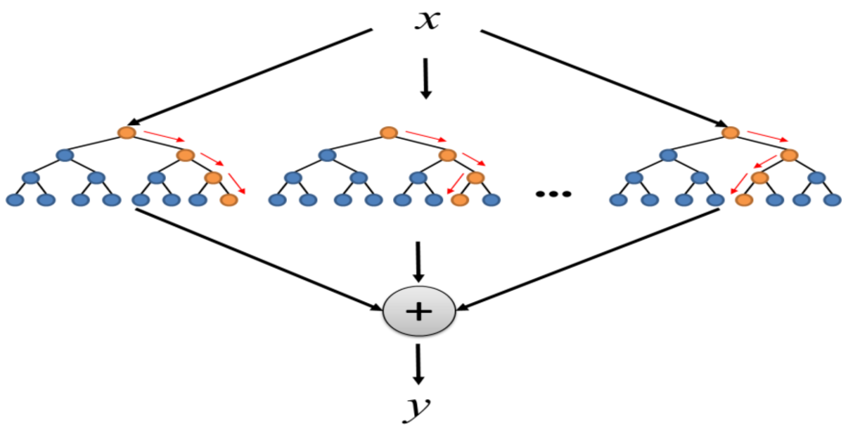


Figure 1.1: Random Forest overview

The figure shows an example of a random forest based on three decision trees. An unknown sample x can be classified by each of the three tree and all results are aggregated to calculate the final class probability y.

1.2.2 Elastic Nets

Elastic nets (Zou & Hastie 2005) or logistic regression with elastic net regularization is another supervised classification method. Logistic regression is a special case of generalized linear model and measures the relationship between a binary dependent variable and one or more independent variables by estimating probabilities using a logistic function. Unlike linear regression, the conditional outcome variable follows a Bernoulli distribution, which means it results in probabilities that are bound between 0 and 1. Since the model is a generalized linear model the least squares method can be used to fit the model onto the data. This method minimizes the sum of squared residuals, where a residual is the difference between an observed value and a value that is provided by the model. This can be expressed in the following way:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2)$$

Where X represents the features, y the dependent outcome variable and β represents the coefficients for the model, where each coefficient corresponds to one feature.

Similar to decision trees, least-squares also suffers from overfitting the training data, which leads to poor performance in independent test sets. This is particularly the case when the number of features greatly outnumbers the number of samples, which is most often the case in genomics data. To overcome this issue regularization methods are employed, which usually take the form of penalty terms that reduce model complexity. One such regularization method is the Tikhonov regularization or ridge regression that adds a L2-norm penalty:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2)$$

where λ_2 controls the weight of the penalty. Ridge regression reduces the magnitude of the regression coefficients β; however, while all parameters are reduced they still remain non-zero. An alternative regularization method is the Lasso (least absolute shrinkage and selection operator) (Tibshirani 1994; Tibshirani 2011), which adds a L1 norm penalty:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_1\|\beta\|_1)$$

which effectively reduces the coefficients of features with low information content to zero, i.e. it reduces the model to the most important features. However, the

Lasso has its own limitations: It can only select as many features p as samples n, which is a problem with genomics data, where $p \gg n$ and additionally it tends to select only one feature out of a group of highly correlated features, which makes it less robust.

Finally, the elastic net is able to overcome the issues of both the ridge regression and the Lasso, by linearly combining both L1 and L2 penalty terms:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1)$$

It both reduces the magnitude of the regression coefficients β and sets their value to 0 for features that are less relevant. As a side note, depending on the parameters λ, the elastic net can be reduced to the Lasso ($\lambda_1 = 0$) or the ridge regression ($\lambda_2 = 0$). The model presented here shows the naïve version of an elastic net. The actual version used in this thesis uses an α parameter [0,1] that controls the tradeoff between the two penalty terms. An elastic net with $\alpha = 0$ is basically a ridge regression, while $\alpha = 1$ amounts to a pure Lasso.

1.2.3 Shrunken centroids

Shrunken centroids or PAM (Hastie et al. 2011) are based on nearest centroid classification, which assigns observations the labels of the class of trainings samples whose centroid is closest to the observation. This means the algorithm compares the features of a new sample to each class centroids, which are the means of the features of all samples in each class. The class label of the centroid closest to new sample based on the Euclidian distance is used as prediction outcome.

Shrunken centroids is an extension of this simple type of classifier, where each value in the centroids is shrunken towards zero by a specified threshold τ, i.e. τ is subtracted from positive values and added to negative values. If a value is reduced to zero it is removed from the classification. This has two advantages: It reduces noise and it provides an in-build feature selection mechanism (Hastie et al. 2011).

1.3 Thesis overview

Malignant tumors remain one of the leading causes of mortality with over 8.2 million deaths worldwide in 2012 (Stewart & Wild 2014). Due to fractured nature of cancer each subtype in each tissue type has its own molecular profile and its own resistances to treatment regimens, thus as of now no overall cure is in sight. Over the last two decades, high-throughput profiling of the human transcriptome has become an increasingly essential tool to investigate molecular processes involved in carcinogenesis. Gene expression profiling can help shedding light on the heterogeneous nature of cancer, but more importantly it can find common and ideally targetable biological processes that cancer relies on.

In this thesis I explore how high throughput transcriptional profiling coupled with statistical machine learning methods can have an impact at every stage of cancer research. In the abstract most of the following projects consists of three steps. 1.) During feature selection, the abundance of gene expression measurements is reduced to a manageable set of clinically/biologically relevant genes, which is often referred to as a gene signature. 2.) Next a prediction model based on this gene signature is derived from a trainings dataset that usually consists of a few dozen to thousands of samples, which we call a biomarker. More specifically, a biomarker is defined as a gene signature and the associated mathematical rule that translates the expression levels of the signature's genes into a probabilistic score of class membership. 3.) And finally, in order to avoid overfitting the dataset each biomarker is validated in an independent dataset. This validation ensures that the prediction does not only reflect the specific nature of the trainings set but captures the underlying molecular processes that lead to differences in phenotypes.

I applied this abstract workflow to cancer prevention, where I modeled the differences between cancer causing and harmless chemicals, patient stratification, where I built models that assign lymphoma patients to actionable subgroups that are treated differently and finally subtype discovery, where I attempted to link an established subtype in lymphoma across tissue boundaries into other cancer types.

1.3.1 Genomic models of environmental and chemical carcinogenicity

Despite an overall decrease in incidence of and mortality from cancer, about 40% of Americans will be diagnosed with the disease in their lifetime, and around 20% will die of it (Stewart & Wild 2014). Only 27% (breast) to 42% (prostate) of cancer risk can be attributed to heritable factors (Lichtenstein et al. 2000), the rest is caused by environmental factors such as chemical carcinogens. Current approaches to test carcinogenic chemicals adopt the 2-year rodent bioassay, which is costly and time-consuming. As a result, fewer than 2% of industrially available chemicals have actually been tested (Fitzpatrick 2008). However, evidence accumulated to date suggests that gene expression profiles from model organisms exposed to chemical compounds reflect underlying mechanisms of action, and that these toxicogenomic models could be used in the prediction of chemical carcinogenicity.

In Chapter 2, I test the applicability of toxicogenomics to model chemical carcinogenicity using the DrugMatrix, a publicly available dataset of primary tissue from rats that were exposed to a large panel of chemical compounds and drugs. I then validate these prediction models for carcinogenicity on an independent dataset, the TG-GATEs and finally, I show my successful attempt to identify pathways that are involved in carcinogenesis using a purely data-driven approach.

1.3.2 Molecular classifiers for aggressive B-Cell lymphomas

The third chapter transitions away from prevention to the treatment of cancer. Instead of predicting chemical carcinogenicity, I built models that predict molecular subtypes, which can be targeted with different treatment regiments. This type of precision medicine maximizes the survival of cancer patients.

Every year, 75,000 patients in the United States are diagnosed with aggressive non-Hodgkin lymphoma (Rummel 2010). Diffuse large B cell lymphoma (DLBCL) — the most common variety — accounts for 31% of new cases. Even though DLBCL tumors are heterogeneous at the molecular level and show different pathogenesis, all patients are treated with the same chemotherapeutic cocktail. This lack of targeted treatments likely helps to explain why more than a third of patients succumb to their disease (Friedberg 2008).

In Chapter 3 and 4, I describe the development of three molecular biomarkers for the classification of aggressive B-cell lymphoma, each of which is able to distinguish between therapeutically relevant subtypes. The first biomarker stratifies tumor samples into Burkitt's and Diffuse Large B-Cell Lymphoma (DLBCL) – two disease subtypes with clearly distinct therapeutic regimens – and can be used as a diagnostic tool to supplement current tools such as IHC staining and FISH tests. Additionally, it also allows the quantification of intermediate cases, which right now are either assigned to BL or DLBCL, without showing the full phenotype of either. The second biomarker stratifies DLBCL patients into MYC high and MYC low classes with the purpose of identifying patients that could benefit from potential alternative treatments targeting the transcription factor cMYC. The third biomarker stratifies DLBCL patients into one of three molecular subtypes (BCR/OxPhos/HR) based on the published consensus clustering classification (CCC) described in Monti et al. 2005. Each of these three subtypes has been shown to be potentially amenable to targeted treatments, some of which are currently in clinical trials.

All three biomarkers are translated from the research settings, i.e. full transcriptome assays and fresh frozen biopsies, into clinical diagnostic tools relying on the Nanostring nCounter platform and formalin-fixed paraffin embedded (FFPE) tissues.

1.3.3 Identifying tumors dependent on oxidative phosphorylation across different cancer types

The samples in one of the therapeutically relevant subtypes in the comprehensive consensus clustering (CCC) classes in DLBCL show a dependence on altered metabolism driven by oxidative phosphorylation (OxPhos)(Monti et al. 2005). A similar subtype was found in melanoma (Vazquez et al. 2013), which supports the hypothesis that OxPhos dependent cancers can be found in a variety of cancers. This is of particular interest since it has been shown that Oxphos dependent cancer cell-lines can be specifically targeted (e.g. with PPARγ inhibitors (Caro et al. 2012))

In Chapter 5, I present a transcriptomics-based pan-cancer analysis that uses a novel method called ASSIGN to stratify samples by their predicted level of OxPhos activity. The generated predictions were then validated in *in-vitro* experiments. Furthermore, I describe the derivation of a generalized PanOxPhos gene signature, and the search for potential transcriptional regulators of the associated molecular phenotype.

2 GENOMIC MODELS OF ENVIRONMENTAL AND CHEMICAL CARCINOGENCITY

The work in this chapter was published:

Gusenleitner D, Auerbach SS, Melia T, Gómez HF, Sherr DH, Monti S 2014. Genomic models of short-term exposure accurately predict long-term chemical carcinogenicity and identify putative mechanisms of action. *PloS one*, 9(7), p.e102579

2.1 **Abstract**

Background**:** Despite an overall decrease in incidence of and mortality from cancer, about 40% of Americans will be diagnosed with the disease in their lifetime, and around 20% will die of it. Current approaches to test carcinogenic chemicals adopt the 2-year rodent bioassay, which is costly and time-consuming. As a result, fewer than 2% of the chemicals on the market have actually been tested. However, evidence accumulated to date suggests that gene expression profiles from model organisms exposed to chemical compounds reflect underlying mechanisms of action, and that these toxicogenomic models could be used in the prediction of chemical carcinogenicity.

Results: In this study, we used a rat-based microarray dataset from the NTP DrugMatrix Database to test the ability of toxicogenomics to model carcinogenicity. We analyzed 1,221 gene-expression profiles obtained from rats treated with 127 well-characterized compounds, including genotoxic and non-genotoxic carcinogens. We built a classifier that predicts a chemical’s carcinogenic potential with an AUC of 0.78, and validated it on an independent dataset from the Japanese Toxicogenomics Project consisting of 2,065 profiles from 72 compounds. Finally, we identified differentially expressed genes associated with chemical carcinogenesis, and developed novel data-driven approaches for the molecular characterization of the response to chemical stressors.

Conclusion: Here, we validate a toxicogenomic approach to predict carcinogenicity, and provide strong evidence that, with a larger set of compounds, we should be able to improve the sensitivity and specificity of the predictions. We found that the prediction of carcinogenicity is tissue-dependent and that the results also confirm and expand upon previous studies implicating DNA damage, the peroxisome proliferator-activated receptor, the aryl hydrocarbon receptor, and regenerative pathology in the response to carcinogen exposure.

2.2 Introduction

[T]he development of truly useful, predictive tests of human carcinogens still lies in the future.

– R.A. Weinberg (Weinberg 2013)

Despite an overall decrease in mortality from cancer, about 41% of Americans will be diagnosed with the disease and about 21% will die from it (Howlader et al. 2013). The incidence of certain cancers is increasing for unknown reasons, and there is substantial evidence suggesting that inherited genetic factors make only a minor contribution (Paul Lichtenstein et al. 2000), while the percentage of cancer cases that can be attributed to infectious diseases remains stable at about 16–18% (Danaei 2012). It has thus been widely hypothesized that accumulating environmental chemicals play a significant role in sporadic cancer (Davis et al. 2013; Sorensen et al. 1988; Lee Davis et al. 2007). There is also growing recognition that the role played by environmental pollutants in human cancer is under-studied, and that more formal approaches to the analysis of the biological consequences of prolonged exposure to pollutants are needed (IBCERCC 2013; Leffall & Kripke 2010).

High-throughput genomic approaches have been successfully applied toward the elucidation of the molecular mechanisms of cancer initiation and progression, to the identification of novel therapeutic targets, and to the development of diagnostic and prognostic biomarkers, resulting in thousands of publications. However, their application to the study of the environmental causes of cancer has not received as much attention.

Standard approaches to carcinogen testing have adopted the 2-year rodent bioassay (2YRB) as the de facto “gold-standard”. The 2YRB requires, for each compound, the use of more than 800 rodents and for each rodent a histopathological analysis of more than 40 tissues, with a cost per compound in the $2-4 million range depending on route of administration, number of doses to be examined, and chemical being evaluated. As a result, only approximately ~1,500 of the ~84,000 chemicals in commercial use have been tested (Bucher & Portier 2004; Gold et al. 2005; Huff et al. 2008; Waters et al. 2010). Furthermore, substantial recent literature questions the reliance on animal assays to model the biology of human carcinogenicity for regulatory purposes

(Boobis et al. 2008; Cohen 2010). On the other hand, the evidence accumulated to date suggests that gene expression profiles of model organisms or cells exposed to chemical compounds reflect underlying biological mechanisms of action and can be utilized in higher throughput assays to predict the long-term carcinogenicity (or toxicity) of environmental chemicals (Waters et al. 2010). Multiple mechanisms of action for rodent hepatocarcinogenicity have been implicated by the analysis of toxicogenomics data, including DNA damage, regenerative proliferation, xenobiotic receptor activation, peroxisome proliferation and steroid-hormone mediated carcinogenesis (Waters et al. 2010; Fielden et al. 2007; Nie et al. 2006). Furthermore, several studies have tested the predictability of (genotoxic and non-genotoxic) carcinogenicity of chemical compounds from the expression profiles of animal models' tissues or cell cultures exposed to the chemicals, and provide preliminary evidence that gene expression-based carcinogenicity prediction is indeed feasible (Waters et al. 2010). While offering valuable insights, and significantly informing the analytic approach reported here, most of these studies were limited to a relatively small number of compounds or to a limited set of transcripts, and have not thoroughly explored the effects of time and dose of exposure, or issues of portability of the models across independently generated, genome-wide expression datasets.

In this study, we present the results of our analysis of two large cohorts of rat-based expression profiles from animals exposed to hundreds of well-annotated chemicals with varying carcinogenicity and genotoxicity (DrugMatrix, (Ganter et al. 2005); TG-

GATEs, (Uehara et al. 2010), see Materials). The profiles represent short-term (hours or days) exposure assays, and, when paired with the available long-term (2 years) carcinogenicity labels of the compounds profiled, provide ideal data with which to test the hypothesis that long-term exposure phenotypes can be accurately modeled by short-term gene expression-based assays. To our knowledge, the collection we assembled represents the largest toxicogenomics resource analyzed to date, and allows us to rigorously evaluate issues of batch-to-batch variability, tissue-, time-, and dose-dependency, sample size adequacy, and determination of the optimal number of genes/transcripts necessary to achieve maximum predictive accuracy.

Here, we detail our predictive model building effort based on a discovery set, the DrugMatrix, comprising 1,221 expression profiles in liver corresponding to 127 chemical compounds tested at multiple doses and exposure times. We then present the results of our evaluation on a completely independent validation set, the TG-GATEs, consisting of 2,065 profiles corresponding to 72 compounds, and we show that our classifier does generalize without loss of accuracy. We investigate the impact of tissue type-, dose-, time-dependency, and sample size on carcinogenicity prediction and also introduce a gene set projection method aimed at increasing the biological interpretability of the predictive model while improving the robustness of the classification across independent datasets. Finally, we present the results of our analysis aimed at the characterization of the carcinogenome, defined as the set of genes and pathways that reflect mechanisms of action associated with carcinogenesis, and of our effort at defining data-driven gene modules reflecting complementary mechanisms of action relevant to chemical carcinogenesis. A graphical overview of all analyses is provided in Figure I.1.

2.3 **Results**

2.3.1 Multi-tissue exploratory data analysis

Principal component analysis (PCA) was performed to identify the major sources of variation in the DrugMatrix dataset. A plot of the first two principal components shows that the data are stratified by tissue type (Figure 2.1a), with heart and thigh muscle tissue results clustering tightly on the lower left side, kidney on the upper left side, and liver tissue and cultured hepatocytes on the right side. 46.3% and 26.1% of the overall variance in the data is explained by the first and second principal components, respectively. Hierarchical clustering of the samples yields similar stratification by tissue of origin (data not shown). These results suggest that tissue is a major confounding factor, and for that reason all subsequent analyses were performed within a given tissue type.

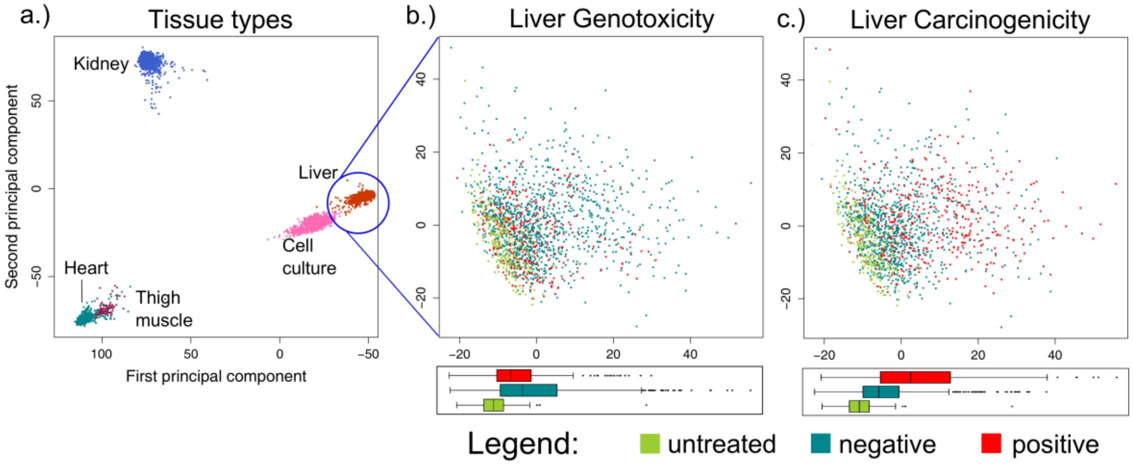


Figure 2.1: Principal component analysis (PCA) of the DrugMatrix.

a) The first two principal components of all samples in the DrugMatrix dataset. b) Liver samples with color coding for controls, samples treated with genotoxic or non-genotoxic samples. c) Liver samples with color coding for carcinogenicity.

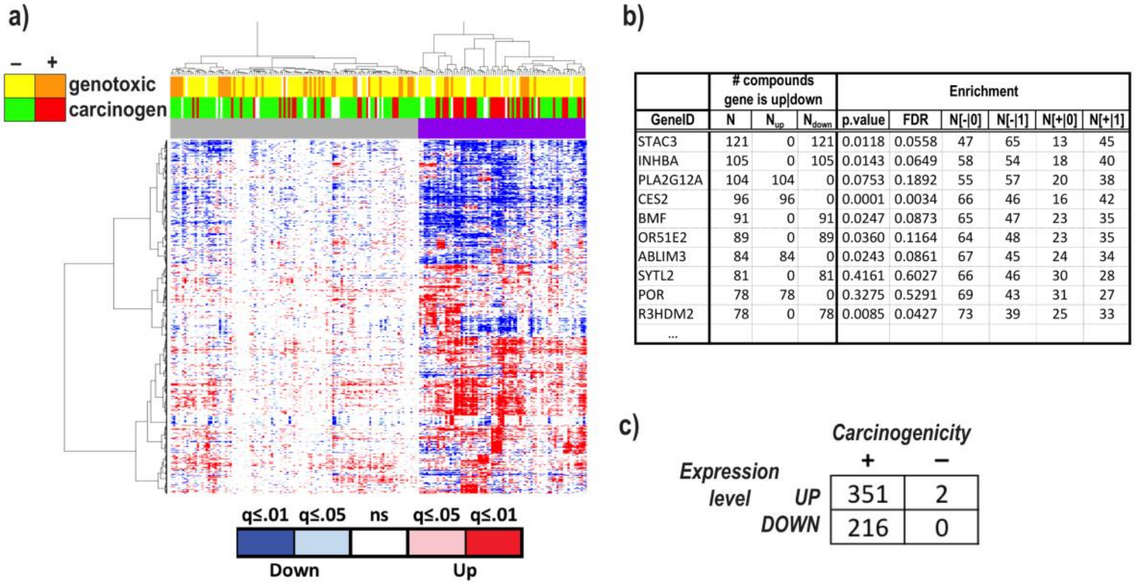
The Carcinogenic Potency Database (CPDB) was used as arbiter of tissue specific carcinogenicity for each compound (Methods and Materials). PCA performed within liver only (Figure 2.1b and Figure 2.1c) shows that the segregation induced by the genotoxicity and carcinogenicity phenotypes is not as marked as the segregation by tissue type, underscoring the need for tissue-specific analyses. Of note, the overall changes in transcript abundance induced by genotoxic compounds are smaller than the changes induced by carcinogenic compounds (1^{st} PC variance of 76.5 versus 182.4, respectively; see boxplots at bottom of Figure 2.1b and Figure 2.1c). This outcome may reflect the fact that genotoxic compounds mediate carcinogenicity through a single mechanism, i.e., DNA damage, while non-genotoxic carcinogens induce malignancy through a variety of pathways including, but not limited to chronic nuclear or growth factor receptor activation, aberrant activation of kinase and calcium channel signaling cascades, increased proliferation, altered apoptosis signaling, and/or altered metabolism, all of which would be expected to yield a broader spectrum of transcriptional changes than those resulting solely from DNA damage, a point to which we will return.

2.3.2 Molecular Characterization of the Transcriptional Response to Chemical Perturbation

Next, we sought to rigorously define the transcriptional response to chemical carcinogens in terms of the genes and signaling pathways significantly associated with chemical perturbations, and differentially expressed between carcinogens and non-carcinogens, as well as between sub-types of carcinogens. To this end, we carried out within- and across-compound differential and pathway enrichment analyses of the DrugMatrix liver samples.

2.3.3 Defining the perturbational transcriptome

We first aimed at characterizing the *perturbational transcriptome* – defined as the set union of the genes that significantly respond to chemical perturbation by *any* compound – and to evaluate whether the perturbation patterns are significantly associated with the carcinogenicity of the compounds. To this end, we identified for each compound the transcripts significantly up- or down-regulated with respect to the matched controls, across multiple durations of exposures. In total, 2,745 (~24%) transcripts showed significant (FDR≤0.01, fold-change≥1.5) up-/down-regulation for at least 5 compounds relative to their matched controls. Of these, 569 had a significant association with the carcinogenicity phenotype at an FDR q-value≤0.05 (see Methods). To obtain a global view of the expression patterns across compounds, a data matrix was generated with each compound represented by the column vector of the 'treatment vs. control' t-scores. Hierarchical clustering of the resulting matrix (Figure 2.2a) yielded a clear segregation of compounds into two clusters, with one highly enriched for carcinogenic compounds (Fisher test p=6.5 x 10^{-6}), and with a significantly higher number of up/down-regulated genes (Kolmogorov-Smirnov test p=0.01, see Methods and Figure I.2). The analysis further showed that: i) genes up-/down-regulated by multiple compounds are either *always* up-regulated or *always* down-regulated, but rarely both (Figure 2.2b); ii) significant up-/down-regulation occurs more often in response to carcinogens than to non-carcinogens, with ~20% of these genes exhibiting a pattern of statistically significant association between up-/down-regulation and carcinogenicity status (Figure 2.2b, 'Enrichment' columns); and iii) the overwhelming majority (567 out of 569) of the transcripts significantly associated with carcinogenicity were enriched in the carcinogenic group, and of these almost two thirds were up-regulated (Figure 2.2c).



	# compounds gene is up|down			Enrichment					
GeneID	N	N_{up}	N_{down}	p.value	FDR	N[-|0]	N[-|1]	N[+|0]	N[+|1]
STAC3	121	0	121	0.0118	0.0558	47	65	13	45
INHBA	105	0	105	0.0143	0.0649	58	54	18	40
PLA2G12A	104	104	0	0.0753	0.1892	55	57	20	38
CES2	96	96	0	0.0001	0.0034	66	46	16	42
BMF	91	0	91	0.0247	0.0873	65	47	23	35
OR51E2	89	0	89	0.0360	0.1164	64	48	23	35
ABLIM3	84	84	0	0.0243	0.0861	67	45	24	34
SYTL2	81	0	81	0.4161	0.6027	66	46	30	28
POR	78	78	0	0.3275	0.5291	69	43	31	27
R3HDM2	78	0	78	0.0085	0.0427	73	39	25	33
...									

		Carcinogenicity	
		+	–
Expression level	*UP*	351	2
	DOWN	216	0

Figure 2.2: Defining the carcinogenome.

a) Hierarchical clustering of 191 profiles/138 compounds (columns) and genes (rows), with each compound represented by the vector of 'treatment *vs*. control' differential expression t-scores. The heatmap is color-coded according to the significance level (q-values) of the corresponding t-scores. Notice the right cluster (top purple color bar) and its enrichment in carcinogenic (red) compounds (Fisher test p=8.5 x 10^{-6}). b) Top 10 genes ranked according to the number of compounds inducing their significant up-/down-regulation (FDR≤0.01 and fold-change≥1.5. See complete list in Table S2.28). Each gene was also tested for its association with carcinogenicity across compounds ('Enrichment' columns) by performing a Fisher test between the gene status (0: not differentially expressed; 1: differentially expressed) and the compounds' status (+ = carcinogenic; - = non-carcinogenic). c) Contingency table detailing the distribution of the genes whose compound-induced up-/down-regulation pattern is significantly associated with carcinogenicity status of the compounds.

In summary, our analysis shows that carcinogenic compounds (irrespective of their mode of toxicological action) induce a more pervasive (more genes) and marked (significant) transcriptional response than non-carcinogens, a response that is consistent across multiple compounds, and that manifests itself more often as an up-regulation of expression than a down-regulation. Furthermore, this heightened response is mainly driven by non-genotoxic mechanisms, since no significant enrichment for genotoxicity is observed in either cluster.

2.3.4 Signatures of carcinogen exposure

Next, we carried out differential analysis aimed at comparing a gene's expression between carcinogens and non-carcinogens (171 vs. 362 liver samples respectively, with replicates of the same condition averaged), irrespective of their level in the controls. The main purpose of this analysis was not the selection of features for predictions, but rather the investigation of the exposure-induced transcriptional changes toward the elucidation of mechanisms of response. Rigorous statistical testing based on a moderated t-test (see Methods), yielded a list of 2,263 differentially expressed genes (DEG) at a false discovery rate (FDR) q-value≤0.01, with 1,232 genes up-regulated and 1,031 genes down-regulated in response to carcinogens. Of note, although the DEGs are highly statistically significant, their fold-change is relatively small, with only 56 genes having FC≥1.35 in either direction (Table 2.1), suggesting that the significance reflects the large sample size, and that it is driven by a relatively small subset of compounds. This is confirmed by a visual inspection of the heatmap displaying the top 250 differentially expressed genes (see web portal (Gusenleitner et al. 2013)), which shows a large heterogeneity within each group. Despite the considerable heterogeneity of the response, a focus on the top markers listed in Table 2.1 confirms that several genes linked to changes in liver swelling, or hepatomegaly (e.g., *ZDHHC2*, *AQP7*, *IL33*), centrilobular hepatic eosinophilia, peroxisome proliferation (e.g., *HDC, ACSL3*), hepatocellular hypertrophy (e.g., *ACOT1, STAC3, CPT1B*) and hepatic lipid accumulation (*HSPB1*, *LRP1*, *NOL3*) were differentially regulated. The identification of pathology-associated biomarkers is consistent with the observation that pathological manifestations in short-term studies are associated with cancer outcomes in rodents, and that pathology such as Cirrhosis in humans is a risk factor for hepatocellular carcinoma (Allen et al.; Simonetti et al. 1991). In addition genes associated with genotoxicity (e.g., *JAM3, BTG2, MDM2, PLN, NHEJ1, CCNG1, MGMT*) appear to be significantly up-regulated in response to carcinogen exposure.

Within the list of carcinogenic compounds, comparison of genotoxic carcinogens vs. non-genotoxic carcinogens yields a list of 191 (126 up, 65 down) DEGs with a FDR≤0.01, but only 86 of these genes have a FC≥1.35 (40 up, 46 down) (Table I.2). This comparison further highlights the significant up-regulation of well-established markers of DNA damage response (CDKN1A/p21, MDM2), liver fibrosis (e.g., AhR), liver hyperplasia (e.g., CYP1A1) and liver inflammation (e.g., BCL6) in response to genotoxic carcinogens, and the up-regulation of markers of liver steatosis, (e.g., CYP4A11, DECR1, EHHADH) and hepatocellular peroxisome proliferation (e.g., ACOX1) in response to non-genotoxic carcinogens.

Several of the genes differentially regulated are associated with tumor initiation (e.g., AhR, CYP1A1, CYP1A2, MDM2, EGR1, and NFKBIZ), further suggesting that genomic outcomes of short-term exposure truly reflect the longer-term process of malignant transformation. A detailed list of all DEGs, including hyper-enrichment analyses of the top genes using DAVID (Huang et al. 2009) is available at [1] (Gusenleitner et al. 2013).

2.3.5 Pathway enrichment analysis

Pathway enrichment analysis by GSEA of the 'carcinogen *vs.* non-carcinogen' signature showed a strong enrichment of DNA damage and repair pathways (e.g., p53, base excision repair, mismatch repair), as well as of regulators of cell proliferation (e.g., *E2F*, *NF-κB*, G_1-S transition), protein turnover (e.g., proteosome, ubiquitin-mediated proteolysis), and enrichment of metabolic pathways (e.g., oxidative phosphorylation and fatty acid oxidation). Further analysis of the 'genotoxic *vs.* non-genotoxic carcinogen' signatures highlighted the major role played by DNA damage and repair pathways in the former, and cell metabolism and oxidative stress in the latter. This is consistent with previously reported studies, which emphasize DNA damage response as a distinctive transcriptional signature of direct DNA modification, and increased cell proliferation,

[1] http://smonti.bumc.bu.edu/~smonti/environcology/rat_carcinogenome/

oxidative stress and metabolism as characteristic of indirect, non-genotoxic modes of action (Waters et al. 2010). Also of notice was the high heterogeneity in the response to non-genotoxic carcinogens when compared to the genotoxic carcinogens, as reflected in the lower number of gene sets significantly enriched in the signature of the former than of the latter. As noted above, this likely reflects the existence of multiple mechanisms of non-genotoxic carcinogenesis, which cannot be adequately captured by a simple dichotomous comparison using anything but a large database.

In summary, our supervised analysis of the DrugMatrix data recapitulates and refines the known *repertoire* of transcripts and associated biological pathways previously implicated in the response to carcinogen exposure, thus confirming the quality of the expression data analyzed and their adequacy for our predictive model building effort, to which we now turn.

2.3.6 Predictive Models of Genotoxicity and Carcinogenicity in the DrugMatrix

The PCA analysis shows that overall expression patterns are mainly driven by tissue type. Furthermore, methods to control for tissue type, such as “subtraction” of the tissue-associated PCA components, or inclusion of tissue type as predictor to build *tissue-agnostic* classifiers, were not fruitful (see Supplement, Table I.3 and Figure I.3). Consequently, we henceforth report our results based on the analysis of the liver samples since this tissue was profiled with the largest number of well-annotated chemicals and its phenotypic annotation was the most thorough.

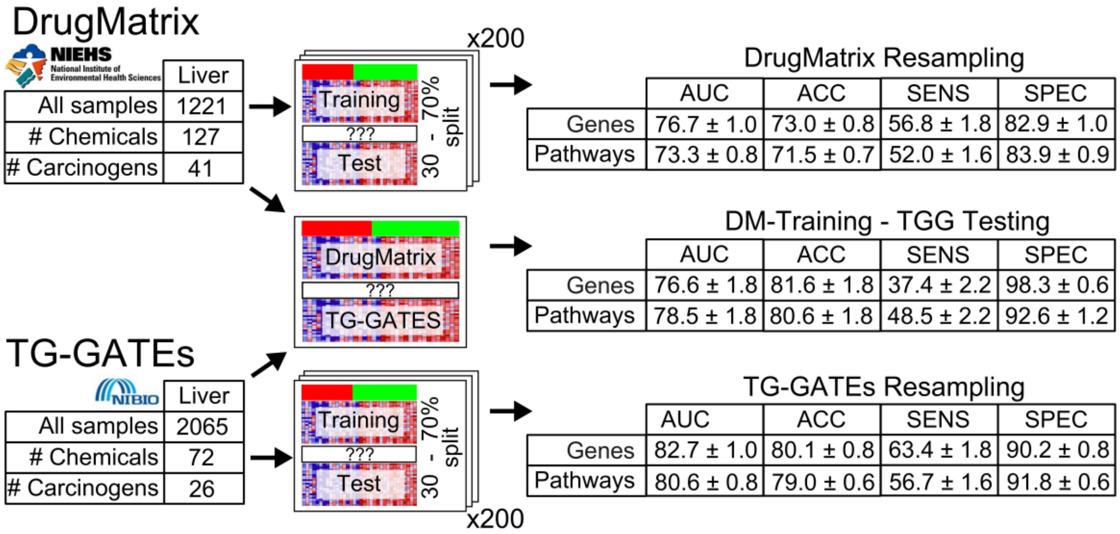
The Random Forest (RF) algorithm (Breiman 2001) was selected as the classifier of choice because of its computational efficiency, flexibility, and ability to model continuous and discrete data simultaneously, as well as to capture complex phenotypes. For each sample, the classifier produces a score between 0 and 1, corresponding to the probability of the compound being carcinogenic (or genotoxic). As the primary evaluation criterion of a classifier's prediction performance, we report the area under the ROC curve (AUC). Additionally, we also report sensitivity, specificity, positive and negative predictive value, and false discovery rate corresponding to the probability threshold that achieves the highest accuracy in the training set (see Methods for further details).

2.3.7 Genotoxicity prediction

Predictive models of genotoxicity based on a 500-gene Random Forest classifier were built from the DrugMatrix liver samples. The random resampling-based estimation of classification performance yielded an AUC (area under the ROC curve) of 75.1%.

2.3.8 Tissue-specific carcinogenicity classifiers

We defined *tissue-specific* labels of carcinogenicity to train a set of predictive models. The resulting carcinogenicity classifier achieved a prediction performance as measured by AUC of 76.7% in liver tissue (Figure 2.3, summary statistics in Table I.4), which represents an improvement of 11.9% with respect to the tissue-agnostic results (Supplement). Using a zero-one loss function to select the optimal classification threshold, corresponding to a zero cost for correct classification (both TP and TN), and a cost of 1 for incorrect classification (both FP and FN), results in a classifier with sensitivity of 56.8% and specificity of 82.91%. However, there is a tradeoff between sensitivity and specificity and, if required, the former can be increased at the cost of the latter. For example, changing the ratio between the penalties of FP and FN to 1:5 increases the sensitivity to 80.4% while the specificity drops to 54.4% (Figure 2.4b). The AUC measures all the possibilities of such tradeoffs.



DrugMatrix

	Liver
All samples	1221
# Chemicals	127
# Carcinogens	41

TG-GATEs

	Liver
All samples	2065
# Chemicals	72
# Carcinogens	26

DrugMatrix Resampling

	AUC	ACC	SENS	SPEC
Genes	76.7 ± 1.0	73.0 ± 0.8	56.8 ± 1.8	82.9 ± 1.0
Pathways	73.3 ± 0.8	71.5 ± 0.7	52.0 ± 1.6	83.9 ± 0.9

DM-Training - TGG Testing

	AUC	ACC	SENS	SPEC
Genes	76.6 ± 1.8	81.6 ± 1.8	37.4 ± 2.2	98.3 ± 0.6
Pathways	78.5 ± 1.8	80.6 ± 1.8	48.5 ± 2.2	92.6 ± 1.2

TG-GATEs Resampling

	AUC	ACC	SENS	SPEC
Genes	82.7 ± 1.0	80.1 ± 0.8	63.4 ± 1.8	90.2 ± 0.8
Pathways	80.6 ± 0.8	79.0 ± 0.6	56.7 ± 1.6	91.8 ± 0.6

Figure 2.3: Classification results overview.

Random resampling classification results on the DrugMatrix (top) as well as the TG-GATEs (bottom) datasets using 200 iterations. In addition, the results of a model trained on all DrugMatrix samples and tested on TG-GATEs (middle) are shown. Results based on the regular gene expression data and on the data projected onto pathway space (canonical pathways of MSigDB – C2:CP, see Methods) are reported. For each testing scheme,

area under the receiver operating characteristic (ROC) curve (AUC), as well as accuracy, sensitivity and specificity of a classifier trained with a zero-one loss function (FP:FN = 1:1), and 95% confidence intervals are reported.

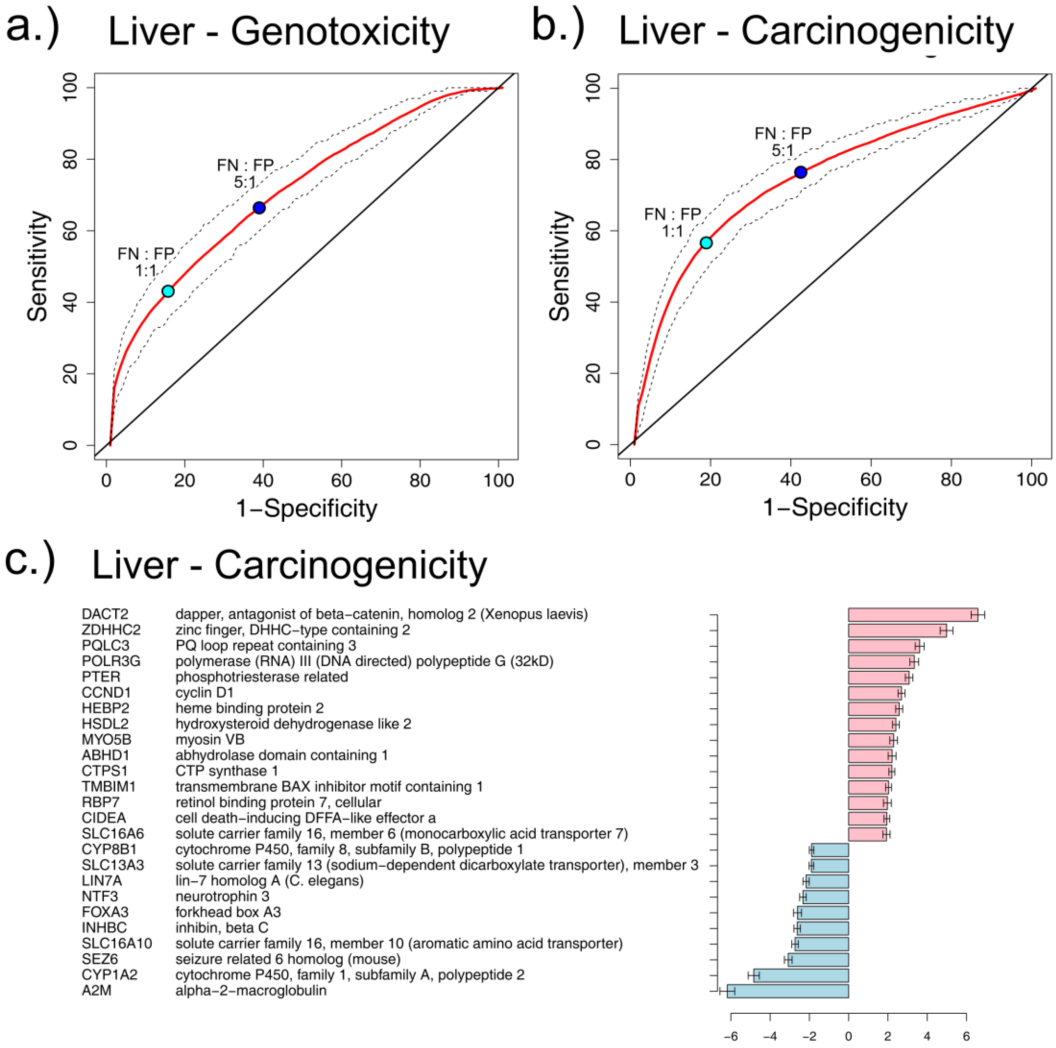


Figure 2.4: ROC curve and variable importance for carcinogenicity prediction.

ROC curve of random forest classification in liver of: a) genotoxicity and b) carcinogenicity. For carcinogenicity, tissue specific class labels from the carcinogenicity potency data base (CPDB) were used. The red curves show the mean of the 200 reruns, whereas the dashed curves indicate the first and third quartile respectively. The teal dot indicates a classifier assigning equal costs to false positives (FP) and false negatives

(FN) (zero-one loss), whereas the blue dot indicates a classifier assigning a cost of 5 for FN and 1 for FP. c) Variable Importance of the random forest model. Blue denotes genes that are down-regulated in the carcinogenic group, whereas red denotes up-regulation.

2.3.9 Inclusion of compounds' structural features as predictors

The availability of structural features characterizing the 3-dimensional chemical structure of the profiled compounds allowed us to evaluate their predictive power (see Materials). To this end, we performed Random Forest classification of all compounds in the DrugMatrix using the structural features, instead of gene expression, as predictors. Evaluation by random resampling yielded an AUC of 70.9% when predicting genotoxicity, and 59.9% when predicting hepato-carcinogenicity (see Table I.7 and Figure I.4), results significantly worse than those obtained based on gene expression. To assess their complementarity, we also evaluated the performance of a Random Forest classifier integrating both gene expression *and* structural features. The resulting model yielded an AUC of 77.7% for hepato-carcinogenicity and 80.1% for genotoxicity (Table I.8 and Figure I.5), suggesting that the information encoded in the structural features is indeed marginally complementary to gene expression.

2.3.10 Comparison to other classifiers

The Random Forest classifier was *a-priori* chosen because of its computational efficiency and its ability to model variable interactions, to handle models incorporating both continuous and discrete variables, and to model complex phenotypes. For completeness, its performance was compared with that of two additional state-of-the-art classification methods: Shrunken Centroids (PAMR) (Tibshirani et al. 2002) and Support Vector machine (SVM) (Chang & Lin 2011), using the same random resampling evaluation scheme. The results in Table I.5 (SVM) and Table I.6 (Shrunken Centroids) show that the Random Forest significantly outperforms both the SVM and the Shrunken Centroids classifiers, providing support for our modeling choice.

2.3.11 Effect of compound sample size on prediction

While the number of well-annotated liver samples in the DrugMatrix was very large (n=1,221), the number of *distinct* compounds tested was comparatively small (127 compounds, 41 of which were labeled as carcinogenic according to the 2YRB). To assess whether we had reached the maximally achievable predictive accuracy, we analyzed learning curves for both carcinogenicity and genotoxicity based on *down-sampling*, whereby AUCs were estimated for classifiers built on training sets of progressively larger size (see Methods). As shown in Figure 2.5, the learning curves (in red) and the corresponding trend lines (blue) manifest a clear upward orientation, and their shape shows no “plateauing,” suggesting that an increased and attainable number of compounds will indeed significantly improve predictive accuracy.

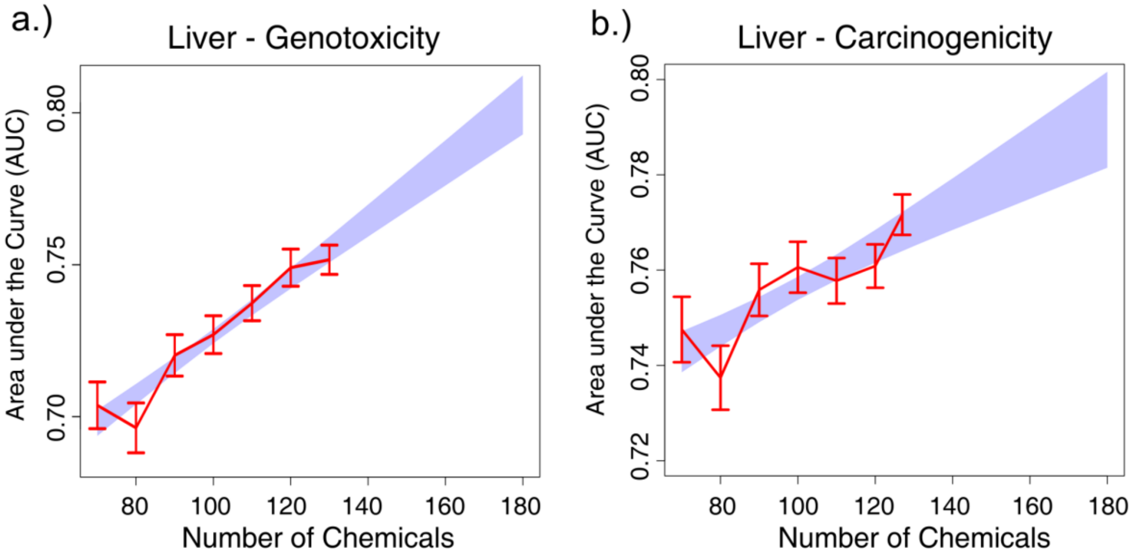


Figure 2.5: Classification learning curves

As a function of the number of chemicals for: a) genotoxicity and b) carcinogenicity in liver. The actual AUC values are in red and include the 95% confidence interval for each value. The predicted values of a fitted linear regression model are shown in blue.

In summary, our Random Forest-based classifier trained on the gene expression data from the DrugMatrix was capable of predicting carcinogenicity with a random resampling AUC of 77.6%, and significantly outperformed other state-of-the-art classifiers (SVM, shrunken centroids, and others), thus making us confident that our modeling approach would generalize well to new untested chemicals.

2.3.12 Validation of the predictive models on an independent dataset: TG-GATEs

The performance of our classification model was next evaluated on an independent *validation set*, the TG-GATEs (see Materials). To this end, a final 500-gene random forest classifier of liver carcinogenicity was trained on *all* of the available compounds in the DrugMatrix (n=127) using the tissue-specific carcinogenicity labels. The top 50 markers as ranked by variable importance are shown in Figure 2.4c. The resulting classifier was then applied to the TG-GATEs. To achieve a truly independent validation set, 25 compounds that were tested in both datasets were excluded, leaving 47 chemicals for validation, corresponding to 1,333 expression profiles (each compound was tested at multiple doses, times, and in triplicates). The Random Forest classifier was then applied to the subset of primary liver samples from the repeat experiments in the TG-GATEs, yielding an AUC of 76.6% (Figure 2.3, summary statistics in Table I.9, ROC curves in Figure I.6 and Figure I.7). Of interest, the prediction of the 25 compounds present in both datasets, yielded a higher AUC of 80.8%, even though those compounds were tested at different doses in the two datasets (data not shown).

2.3.13 Prediction of dose-dependent carcinogenicity

The prediction performance of the models trained on DrugMatrix and tested on TG-GATEs provides supporting evidence of the validity of our approach since significant classification accuracy was achieved across datasets despite the difference in experimental conditions (dose and time) of the two datasets, and the known dataset-to-dataset bias inherent in the Affymetrix microarray platform (Shi et al. 2010; Fielden et al. 2011). To further evaluate the best achievable classification performance, we next applied our random resampling scheme within the TG-GATEs. Besides the differing dose and exposure times profiled in the two datasets, an additional difference between the DrugMatrix and TG-GATEs lies in the more precise compound annotation of the latter, where carcinogenicity labels reflect a compound's actual carcinogenicity at the administered dose. The DrugMatrix doses, on the other hand, are all at or above the standard administered doses reported in the Carcinogenic Potency Data Base (CPDB). This raises the possibility that some of the compounds labeled as non-carcinogen by the CPDB at the standard dose might be carcinogenic at the higher doses tested in the DrugMatrix, and consequently be given a false negative labeling for training and testing purposes. Confirming this possibility, evaluation by random resampling within the TG-GATEs, where all the doses were within the CPDB range, showed an overall increase in classification performance with an AUC of 82.7% (summary statistics in Table I.10, ROC curves in Figure I.8). To further evaluate the dependency of these results on the dose-specific labeling, we also measured classification performance based on a dose-independent annotation of TG-GATEs, by using the minimum dose labeling for all the profiles at any dose (thus reproducing the compound-labeling criteria used in the DrugMatrix). This led to a significant reduction in the prediction performance, with an AUC of 69.3% (summary statistics in Table I.11, ROC curves in Figure I.8), results similar to those achieved in the DrugMatrix.

2.3.14 Effect of time and dose on prediction

With the predictive model established and validated on an independent dataset, we next tested the impact of exposure time and dose on the effect of a chemical compound. The repeat samples (see Materials) from TG-GATEs correspond to systematic tests of chemical compounds at four different exposure times between 4 and 29 days and at three doses, with three replicates for each condition. Predictive accuracy for each time-dose combination was assessed based on the random resampling scheme, and the corresponding AUCs and 95% confidence intervals are shown in Table 2.1. The results range from an AUC of 58.6% with the lowest dose and shortest time to an AUC of 86.8% for the highest dose at the longest time of exposure. Prediction performance is more dependent on the dose level and less on the duration of exposure. This is evident when considering only the highest dose, where the AUC varies only by 4.7% between 4 and 29 days.

Table 2.1: AUC for different time points and doses in TG-GATEs

comparing the prediction results based on differing a times and doses in the repeat subset of TG-GATEs. Each classification was performed 200 times. The table reports the mean AUC as well as the 95% confidence intervals.

		Dose		
		low	**middle**	**high**
Exposure time	**4 days**	58.6 ± 2.0	73.8 ± 1.6	82.1 ± 1.6
	8 days	70.7 ± 1.8	81.7 ± 1.0	84.2 ± 1.4
	15 days	73.6 ± 1.8	82.2 ± 1.2	82.8 ± 1.6
	29 days	73.9 ± 2.0	79.2 ± 1.2	86.8 ± 1.2

In summary, validation of our carcinogenicity classifier on an independent dataset confirmed the predictive accuracy obtained in the discovery set, thus proving the robustness and generalization capability of our modeling approach. Furthermore, the increased accuracy we achieved by training and testing within the same validation dataset, while taking advantage of dose-dependent labels, further emphasizes the critical role played by across-dataset bias, and the importance of using accurate (dose-dependent) phenotypic labels.

2.3.15 Carcinogenicity prediction of un-annotated compounds

The availability of the short-term histopathology reviews for the samples profiled in TG-GATEs allowed us to preliminarily assess our ability to predict the carcinogenicity of chemicals not included in the CPDB, and thus begin to address our ultimate goal of predicting the carcinogenicity of as-yet untested chemicals. To this end, we derived two binary scores from the histopathology findings included in the TG-GATEs, a fully data driven score, *H-score$_d$*, and a manually derived score, *H-score$_m$* (see Materials), and used these scores as gold-standard proxies of the carcinogenic potential of a given compound-time-dose instance against which to test our classifier's accuracy.

Since this evaluation required the time-consuming manual review of histopathology findings, the analysis was limited to a subset of the available samples. In particular, repeat samples from rats exposed at maximum dose and maximum time (29 days) were selected. Next, a 500-gene Random Forest classifier was trained on the samples with the same exposure time and dose level for which hepatocarcinogenicity status was available (n=108). This classifier was applied to the prediction of all unknown compounds (n=252), and only samples with prediction probability above 0.66 (carcinogenic) or below 0.33 (non-carcinogenic) were selected, yielding a final set of 124 samples for which manual (and blind) review of the histopathology findings was available. The comparison of the classifier's predictions with the pathology-derived scores is summarized in Table 2.2. The classifier's sensitivity with respect to both scores is very high, with only the three replicates of mexiletine showing discordance between the classifier's prediction (non-carcinogen) and the histopathology scores (carcinogen). The specificity is comparatively lower with respect to both scores, and in particular with respect to the manually derived *H-score*$_m$; however, the false positive instances mostly correspond to compounds whose multiple replicates disagree with respect to their *H-score*$_m$, that is, the false positive instance was predicted as positive by our classifier, but was *H-score*$_m$ negative, while the additional replicates of the same compound were both predicted and *H-score*$_m$ positive (bucetin, doxorubicin, sulindac, trimethadione). We expect that with a longer time of exposure the pathology report would also show evidence for carcinogenicity.

Table 2.2: Validation of prediction using pathological items

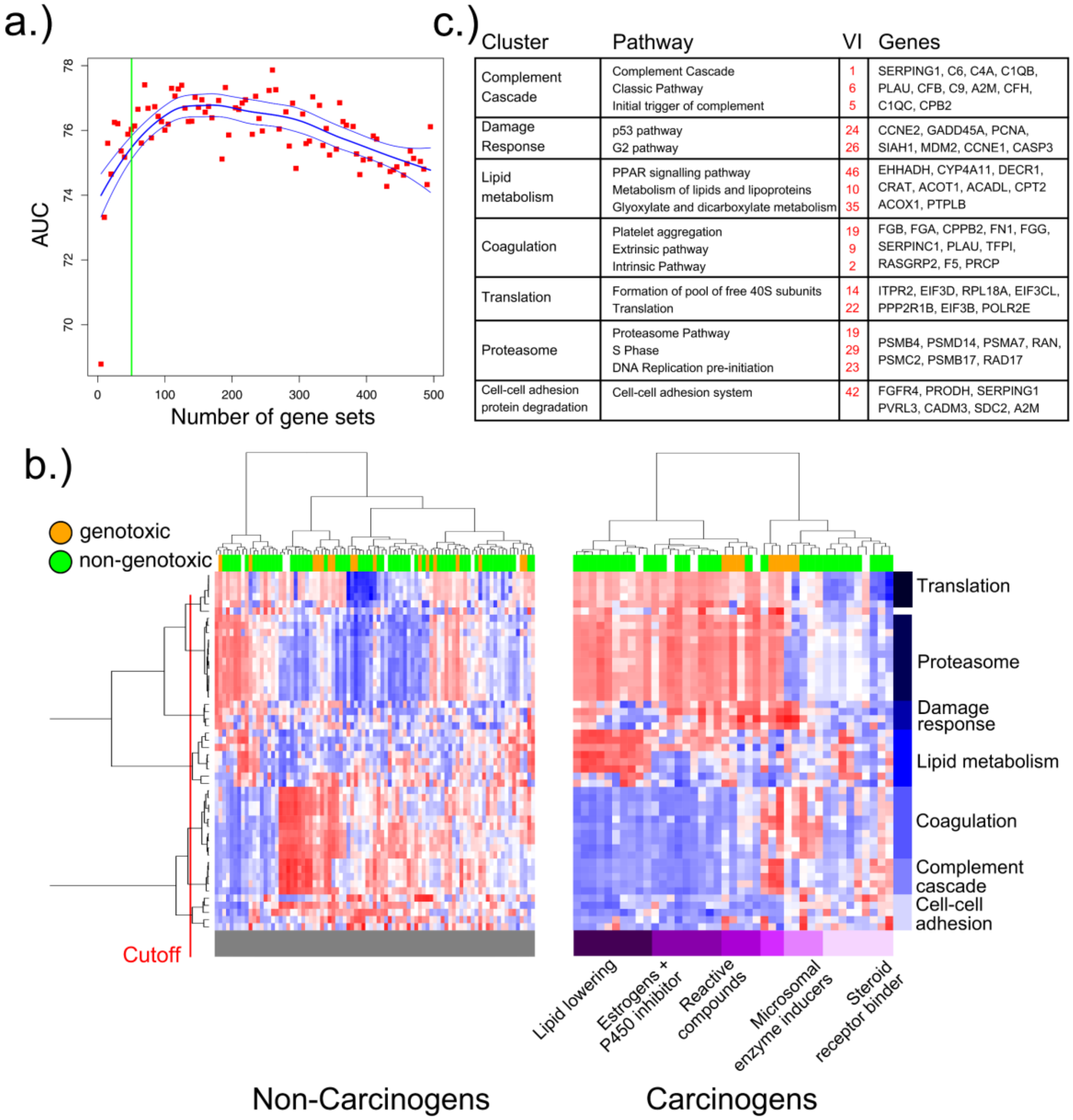
The first column shows the concordance between the high confidence predicted liver samples that were treated for 29 days at the highest dose level and fully data-driven histopathological score (H-score$_d$), whereas the second column indicates the concordance with the manually derived score (H-score$_m$).

	H-score$_d$	***H-score$_m$***
#Samples	124	124
Accuracy	89.5 ± 5.5	79.8± 7.1
Sensitivity	94.3 ± 4.1	95.8 ± 3.5
Specificity	77.8 ± 7.3	57.7 ± 8.6
PPV	91.2 ± 4.9	75.8 ± 7.4
NPV	84.8 ± 6.3	90.9 ± 4.9
FDR	8.8 ± 4.9	24.2 ±7.4

2.3.16 Toward biologically interpretable predictive models: Gene Set Projection

Our next effort was aimed at increasing the interpretability and cross-platform robustness of the classifier. To this end, we adopted a *gene set projection* approach, whereby the data are mapped from single genes to gene sets representing well-annotated biological pathways and processes (Figure I.10). Gene sets are then used in place of single genes as the input variables to the classifier, with a gene set value reflecting the activation/inactivation of that gene set in response to a given compound (see Methods). The 733 canonical pathways included in the MSigDB c2.cp compendium (Liberzon et al. 2011) were used as our candidate gene sets, thus yielding a 733-by-1173 gene set-based matrix from the original 10371-by-1173 gene-based matrix. The classification performance of gene set-based random forest classifiers was evaluated by random resampling (Figure 2.3) both within the DrugMatrix (Table I.12) and the TG-GATEs (Table I.10), yielding a liver carcinogenicity AUC of 73.3% and 80.6%, respectively. These results are slightly worse than those attained based on the original gene-based data. However, training on the gene set-projected DrugMatrix and testing on the TG-GATEs resulted in an increased predictive performance as shown in (Table I.9) (AUC of 78.5%). This is likely due to the normalization implicit in the gene set projection, which involves the scaling of each compound's profile against the matching controls, and thus contributes to removing potential sources of across-dataset bias.

To determine the minimum number of gene sets necessary to reach maximum prediction performance, classifiers with an increasing number of gene sets were built and evaluated. First, gene sets were ranked by their *variable importance* (see methods) as measured by a Random Forest classifier built on all gene sets. Next, RF classifiers using an increasing number of gene sets selected from the variable importance-ranked list were built and evaluated based on the same 70%-30% train-test split previously described. The results (Figure 2.6a) show that 50 gene sets are sufficient to reach an AUC of 76%, and approximately 150 are necessary to reach the maximum predictive performance of 76.8%.



Cluster	Pathway	VI	Genes
Complement Cascade	Complement Cascade Classic Pathway Initial trigger of complement	1 6 5	SERPING1, C6, C4A, C1QB, PLAU, CFB, C9, A2M, CFH, C1QC, CPB2
Damage Response	p53 pathway G2 pathway	24 26	CCNE2, GADD45A, PCNA, SIAH1, MDM2, CCNE1, CASP3
Lipid metabolism	PPAR signalling pathway Metabolism of lipids and lipoproteins Glyoxylate and dicarboxylate metabolism	46 10 35	EHHADH, CYP4A11, DECR1, CRAT, ACOT1, ACADL, CPT2 ACOX1, PTPLB
Coagulation	Platelet aggregation Extrinsic pathway Intrinsic Pathway	19 9 2	FGB, FGA, CPPB2, FN1, FGG, SERPINC1, PLAU, TFPI, RASGRP2, F5, PRCP
Translation	Formation of pool of free 40S subunits Translation	14 22	ITPR2, EIF3D, RPL18A, EIF3CL, PPP2R1B, EIF3B, POLR2E
Proteasome	Proteasome Pathway S Phase DNA Replication pre-initiation	19 29 23	PSMB4, PSMD14, PSMA7, RAN, PSMC2, PSMB17, RAD17
Cell-cell adhesion protein degradation	Cell-cell adhesion system	42	FGFR4, PRODH, SERPING1 PVRL3, CADM3, SDC2, A2M

Figure 2.6: Putative Modes of Action of carcinogenic chemical compounds

a) Classification performance (AUC, averaged over 100 iterations of random resampling) of a random forest classifier as a function of the number of gene sets used as predictors. 150 gene sets are needed to reach maximum AUC, while 50 are sufficient to get 99% of the expected maximum AUC. b) Heatmaps of the top 50 pathways as ranked by their variable importance derived from a random forest classifier of hepato-carcinogenicity. Rows correspond to pathways, clustered into biological processes; columns correspond to

chemical compounds. The left and right heatmaps show all non-carcinogenic and carcinogenic compounds, respectively. Only profiles corresponding to maximum duration and dose treatments, with replicates averaged, are displayed. A detailed version of the right heatmap with all pathways and compounds labeled is available in Figure S11. c) Details of the biological processes associated with the clustering, showing the single differentially regulated pathways and their variable importance ranking, as well as the driving genes.

2.3.17 From predictive models to mechanisms of action

The list of gene sets as ranked by their variable importance provide a set of complementary and potentially interacting biological pathways shown to be statistically associated with chemical carcinogenesis. This is markedly different from the GSEA ranking, which evaluates each gene set individually and does not take into account its possible interaction with other gene sets.

We exploited these properties of the variable importance ranking toward a data-driven identification of the likely mechanisms of action relevant to chemical carcinogenesis. To this end, we projected the DrugMatrix data corresponding to the max-dose and max-duration exposures (to maximize signal) onto the top 50 gene sets as ranked by variable importance. We then performed hierarchical clustering to identify modules of coordinated gene sets likely to reflect distinct mechanisms of action. The resulting heatmap is shown in Figure 2.6b. Multiple gene sets are clustered in distinct modules each reflecting a different biological process that likely contributes to a compound's mechanism of action (MoA). These include a suppressed normal liver function module (complement cascade, platelet aggregation plug formation as well as classic, common and extrinsic pathway), a metabolism of lipids and lipoproteins module, as well as the PPARα signaling pathway, damage response (p53 pathway) and proliferation (DNA Replication pre initiation) modules.

Even though there are only 41 distinct carcinogenic compounds tested in the dataset, the gene set projection-based clustering results highlight the considerable heterogeneity in the response to carcinogen exposure, likely reflecting distinct mechanisms of cancer induction, and point to a promising approach to their data-driven categorization. A notable example is represented by the seven genotoxic compounds clustered under the orange color bar on top of the heatmap (Figure 2.6b). Genotoxic compounds induce direct DNA modifications and cells respond by up-regulation of components of the damage response machinery, such as the p53 pathway and the G2 pathway. A second example is the down-regulation of regular non-metabolic liver function (complement cascade, platelet aggregation and classic pathway) in almost all carcinogenic compounds. We suspect this loss of function is due to elevated stress on the cells and possibly even a first sign of field effects necessary to support transformation. This clearly suggests that the various classes of carcinogens can not only be defined by the mechanisms that eventually lead to carcinogenesis, but also by the loss of specific normal functions within a tissue type, emphasizing the need to consider each tissue type separately.

A third cluster of compounds exclusively captures lipid lowering compounds (Simvastatin, Clofibrate, Gemofibrozil, etc.), which all show a significant up-regulation of lipid metabolism pathways (metabolism of lipids and lipoproteins, glyoxylate and dicarboxylate metabolism). Lipid-lowering drugs have been under suspicion as potential carcinogens for more than a decade (Newman 1996), and aberrant lipid metabolism has been shown to be an essential feature in Hepatocellular Carcinomas (Patterson et al. 2011) as well as cancers in other tissue types (e.g., ovarian cancer (Pyragius et al. 2013)).

Finally, more than two thirds of the carcinogenic compounds show an up-regulation of the proteasome pathway. This is interesting since a large body of scientific literature (e.g. (Fielden et al. 2011; Crawford et al. 2011) identifies the ubiquitin-proteasome pathway as an important component for maintaining a balance between cell growth and apoptosis, thereby controlling tumor propagation and survival.

Taken together, these results suggest that gene set projection is a helpful approach for controlling for batch-to-batch and cross-dataset variability, while increasing a classifier’s interpretability by making explicit the biological pathways that contribute to prediction.

2.4 Materials and methods

2.4.1 Data Resources

The Carcinogenic Potency Database (CPDB[2]) (Fitzpatrick 2008) was used as the primary source to determine a compound’s long-term carcinogenicity and genotoxicity. The CPDB records the results of 6,540 chronic, long-term animal cancer tests on 1,547 chemicals. For this study we used the outcomes of the 2-year male rat-based bioassay to annotate the carcinogenicity of our chemical compounds, while the outcome of a corresponding salmonella auxotroph-based Ames test was used as proxy for genotoxicity. Carcinogenicity information was summarized in a *tissue-agnostic* carcinogenicity label, set to be positive if the compound was found to cause cancer in *any* tissue type, negative otherwise. Additionally, *tissue-specific* carcinogenicity labels were also defined for liver.

The *discovery set* is based on the DrugMatrix[3] (Ganter et al. 2005), a major toxicogenomic resource made public by the National Toxicology Program (NTP) and is available through the Gene Expression Omnibus (GEO) with the accession number GSE57822. The DrugMatrix contains 5,587 gene expression profiles from male rat primary tissues (liver, kidney, heart and thigh muscle) and cultured rat hepatocytes, corresponding to treatments with 376 chemicals, and including 994 control samples from

[2] http://toxnet.nlm.nih.gov/cpdb/

[3] https://ntp.niehs.nih.gov/drugmatrix/index.html

rats kept in matched conditions. Each compound was administered at multiple doses and durations (6 hours - 7 days), and each combination of tissue, compound, time and dose was profiled in triplicates. Of the 376 chemicals tested, 255 are annotated with either carcinogenicity or genotoxicity information in the CPDB, corresponding to 3,448 profiles (a detailed description is provided in Table S2.20). Not all tissues were profiled for each compound tested. In particular, a total of 127 compounds with both hepatocarcinogenicity *and* genotoxicity annotation were profiled in liver, yielding a set of 1,221 profiles available for model building.

The *validation set* is based on the *Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system* (TG-GATEs[4]), a product of a collaboration between the Japanese government and Japanese pharmaceutical companies (Takashima et al. 2006; Uehara et al. 2010), and is available through ArrayExpress (E-MTAB-800). The TG-GATEs includes 21,385 samples of male rat primary liver and kidney tissues, and cultured hepatocytes all profiled on the Affymetrix Rat 230.2 platform. TG-GATEs tested 131 chemical compounds, for 72 of which information on liver carcinogenicity is available (Table I.18). The profiles from primary tissues correspond to two experimental groups: in the *single* group, rats were exposed at a single time point, and mRNA was extracted after 3 to 24 hours, in the *repeat* group, rats were exposed daily for 4 to 29

[4] http://thedatahub.org/dataset/open-tggates

days, and mRNA was extracted at each of four end points (4, 8, 15, and 29 days), and at each of three doses (low, medium, high). For this study we used only the *repeat* group of TG-GATEs. Of the 72 compounds tested in TG-GATEs, 25 were also tested in the DrugMatrix, leaving 47 unique compounds for validation. Comparison of the overlapping chemicals shows that the doses used in the TG-GATEs are lower than those used in the DrugMatrix (Table I.19). Annotation for liver carcinogenicity was performed by a board certified toxicologist through review of existing literature resources from carcinogenicity bioassays. A treatment (chemical-dose combination) was annotated as hepatocarcinogenic if it was determined that it would produce a statistically significant increase in liver cancer (any type) in a 2-year rat cancer bioassay. All dose levels used to generate the TG-GATES data were presumed to be acceptable for use in 2-year bioassay (i.e., animals would survive to the extent that they would be at risk for the development of cancer).

2.4.2 Computational Tools

Analyses were performed based on custom scripts developed using the statistical programming language R (R Core Team 2012) and several Bioconductor packages (Gentleman et al. 2004).

2.4.3 Data Processing

Both Affymetrix datasets were normalized using the R Bioconductor package frma and frmaTools (McCall et al. 2010). Probe specific effects and variances for the Affymetrix Rat 230.2 platform were pre-computed using 2000 samples randomly drawn from the DrugMatrix dataset and then used to normalize both the DrugMatrix and TG-GATEs datasets.

2.4.4 Defining the perturbational transcriptome

The list of genes that significantly respond to chemical perturbation was identified by carrying out a series of two-group t-tests between the control samples and the corresponding treatment samples for each compound separately, while correcting for the confounding effect of time. A gene-by-compound matrix was then constructed, with each column representing the vector of "control *vs*. treatment" t-scores for the corresponding compound. A total of 191 profiles, corresponding to 138 compounds (some at multiple doses) for which either carcinogenicity or genotoxicity information was available, were considered for this analysis. Only the genes with FDR-corrected q-value≤0.01 and fold-change≥1.5 in at least five compounds were included. Hierarchical clustering of both the compounds and the genes based on the t-scores' matrix was performed, and the results visualized in a heatmap with the color-coding based on the t-test's q-values (Figure 2.2a). Association between cluster membership and carcinogenicity (genotoxicity) status of the compound was assessed by Fisher test.

Each gene was tested for its association with carcinogenicity by performing a Fisher test between the gene status (0: not differentially expressed; 1: differentially expressed) and the compound status (+: carcinogenic; –: non-carcinogenic) across compounds, and the nominal p-values were corrected for multiple hypothesis testing by the FDR procedure (Figure 2.2b, columns grouped under 'Enrichment').

2.4.5 Differential Analysis and Pathway Enrichment Analysis

We derived standard differential gene expression signatures using the R/Bioconductor package Limma (Smyth 2005), which is based on linear modeling and a moderated t-test. Since labels for genotoxicity (GT) as well as carcinogenicity (CG) were available in the DrugMatrix, we used multiple binary phenotypes: GT vs. Non-GT, CG vs. Non-CG, GT-CG vs. Non-GT-CG, and Non-GT-CG vs. Non-GT-Non-CG. For TG-GATEs we only tested CG vs. Non-CG. Expression profiles from multiple replicates of the same condition were averaged so as to avoid inflating statistical significance. We also performed a hyper-enrichment analysis of the top 200 differentially expressed genes (up-regulated) of each scheme using DAVID - EASE (Huang et al. 2009) and plotted heatmaps of top differentially expressed genes with a false discovery rate (FDR) corrected q-value≤0.05 and a fold change≥1.2. Finally, we used the same binary phenotypes to run gene set enrichment analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005) using collections C2 (canonical pathways), C3 (transcription factor targets) and C6 (cancer pathways) from MSigDB (Liberzon et al. 2011) version 3.0.

2.4.6 Classification Methods

The Random Forest algorithm (as available through the R package `randomForest`) implements an ensemble classification approach combined with *bagging*, whereby multiple decision trees are inferred from random subsets of the training data, and the class predictions of the component trees are combined by majority voting. After evaluation of multiple sizes, a Random Forest based on 500 trees (the package's default) was selected as the size that yielded the best trade-off between accuracy and computational efficiency. In addition to the performance measurements, we also report the *variable importance* for each gene. This measurement reflects the increase of the error rate across all trees, if the value of the tested gene is randomly permuted when testing.

For comparative purposes, the *shrunken centroid* and the *support vector machine* classifiers, as implemented in the R packages `pamr` (Hastie et al. 2011) and `e1071` (Meyer et al. 2012), respectively, were also evaluated.

2.4.7 Performance evaluation criteria

To assess classification performance, we used a *random resampling* or *bagging* scheme (Breiman 1998) whereby the dataset was randomly split into properly stratified training- and test-set pairs multiple times, a predictive model was inferred from each training set, and tested on the corresponding test set (see Figure I.9). A 70%-30% train/test split was adopted, and was repeated 200 times to obtain robust accuracy estimates and their corresponding 95% confidence intervals. Importantly, since multiple instances of the same compound are included in the dataset, the train/test split was carried out so that all instances of the same compound were only present in the train- *or* the test-set. The prediction for each sample consisted of a value between 0 and 1, to be interpreted as the probability of the corresponding compound of being carcinogenic (genotoxic). The area under the ROC curve (AUC) was chosen as our primary evaluation criterion since this measure is independent of the threshold chosen to call a compound carcinogenic (genotoxic). The choice of the appropriate threshold depends on the relative costs assigned to false negatives and false positives, and these in turn depend on the primary purpose for which the classifier is used, an assessment that is beyond the scope of this study. For completeness, accuracy, sensitivity, specificity, positive and negative predictive values, and false discovery rate are also reported for every classification task (Table I.20) with the positive classification threshold optimized to maximize accuracy (i.e., minimize a zero-one loss (Berger 1985)) within the training set.

2.4.8 Comparison with published signatures

We compared our random forest prediction model to two published gene signatures: A 141 gene carcinogenicity signature (Ellinger-Ziegelbauer et al. 2008) (Ellinger-Ziegelbauer 2008) and a 23 gene non-genotoxic carcinogen signature (Fielden et al. 2011). Both signatures were mapped to Rat Ensembl gene identifiers using Biomart (Kasprzyk 2011) and subsequently tested by training on the DrugMatrix and testing on the compounds within TG-GATEs that did not overlap with the DrugMatrix. Since the Fielden et al 2011 signature was specifically derived from non-carcinogenic compounds; we used cross-validation in the DrugMatrix using all annotated liver samples and on the non-genotoxic subset only. For both signatures we used a Support Vector Machine as classification algorithm (R-package `e1071` (Meyer et al. 2012)) since it was also used in the original publications.

2.4.9 Gene set Projection

Gene set projection was used to map the original data from gene space to gene set space. In particular, each treated sample was compared with the set of corresponding control samples, and a weighted Kolmogorov-Smirnov *enrichment* score was calculated for each gene set (Subramanian et al. 2005) (Figure I.10). This enrichment score reflects the up or down-regulation of *a-priori* defined pathways or gene sets following treatment with the profiled compound. The projection transforms the data from the original gene-by-sample matrix representation to a *gene set*-by-sample matrix, with the entry in row *i*, column *j* reporting the enrichment score for the *i*-th gene set in the *j*-th sample. The set of canonical pathways included in the c2_cp collection of the MSigDB repository was used for the projection (MSigDB version 3) (Liberzon et al. 2011). The resulting projection is different from the one that would be obtained by “single-sample GSEA” (Barbie et al. 2009), since each compound-time instance is normalized against the matched controls, thus yielding a gene ranking that reflects the true differential expression between treatment and control. The projected data thus obtained were then used to train classification models with gene sets in place of genes as the predictive features.

2.4.10 Learning curves for sample size estimation

Learning curves relating classification AUC to compound’s sample size were built based on a variation of the standard random resampling scheme. Starting from a training set consisting of 70 compounds, up to the total number of compounds in increments of 10, AUC means and standard deviations were estimated based on 200 random resampling iterations. The estimated AUCs and their corresponding number of compounds are shown in Figure 2.5, together with linear regression lines fitted on the [sample size; AUC] pairs.

2.4.11 Histopathology annotation of TG-GATEs

TG-GATEs provides the results of histopathology exams of tissues from the profiled animals, including high-resolution whole slide digital images of their liver and kidney on the TG-GATEs portal[5]. The histopathology findings are coded into 133 categorical covariates, each taking values in the range 0-4 (0: pathology not observed; 4: pathology was severe) and includes items such as *liver microgranuloma* and *liver hypertrophy centrilobular*. To summarize these findings and relate them to carcinogenicity, we defined two binary scores (negative/positive) to label each of the compound-dose-time instances. The first score ($H\text{-}score_d$) is data-driven and represents the logic OR of all the covariates, denoting an instance as positive if *any* of the covariates for that instance has a value greater than zero (i.e., if there is *any* type of positive histopathology evidence).

The second score ($H\text{-}score_m$) results from the manual review of a compound-dose-time instance by a board certified toxicologist with experience designing and interpreting subchronic and chronic toxicity/carcinogenicity studies. Factors taken into account when scoring the samples included the degree of adversity associated with specific pathologies (e.g. necrosis is typically considered the most adverse of pathologies), the historical association between the pathological manifestation and subsequent liver cancer outcomes

[5] http://toxico.nibio.go.jp/

in a 2-year bioassay, the severity of the pathology observed, and the multiplicity of pathology types. Due to the time-consuming nature of the manual review, only a subset of compound-dose-time instances were annotated, corresponding to the repeat samples from rats exposed at the maximum dose for 29 days (maximum time). The manual review and annotation of the instances was blinded, that is, the carcinogenicity status predicted by the classifier was withheld at the time of the instances' annotation. The resulting scores were used as a proxy measure of carcinogenicity to evaluate the prediction performance of our classifiers on compounds for which no CPDB annotation was available.

2.5 Discussion

Through our computational analysis of two large rat-based gene expression datasets, we conclusively validated the hypothesis that expression profiles of short-term exposure are highly predictive of the long-term carcinogenicity of (exposure to) chemicals as measured by the 2-year rodent bioassay. Additionally, we extensively evaluated the capability of gene expression profiling to model the transcriptional effects of exposure to chemical perturbations, and showed that the integration of data-driven analysis and pathway-centered annotation best captures the biological processes and pathways that this exposure affects.

2.5.1.1 Building carcinogenicity biomarkers

Analysis of expression data from multiple tissues (liver, kidney, heart and thigh muscle) showed that the most effective approach to carcinogenicity prediction necessitates the definition of tissue-specific classifiers. Consequently, we focused our classification effort on data from liver, since this tissue had the largest number of profiles and compounds evaluated, as well as the most thorough compound annotation.

Classification performance.

Our classifiers based on the Random Forest, and on as few as 500 genes as predictors (selected by variance filtering), yielded predictive accuracy as measured by AUC ranging from 76.7 (DrugMatrix) to 82.7 (TG-GATEs), with the sensitivity/specificity trade-off depending on the cost function adopted (Figure 2.3). The predictive accuracy of the classifier trained on the discovery set (DrugMatrix) and tested on the validation set (TG-GATEs) yielded an AUC of 76.6%, which increased to 78.5% when using gene set projection, proving that our random resampling approach provided an accurate and unbiased estimation of prediction performance. Of notice, the classification performance within the TG-GATEs (by random resampling) exceeded the performance across datasets (AUC: 82.7% vs. 76.7% - Figure 2.3). This is likely due to the dose-specific carcinogenicity annotation in the TG-GATEs, a hypothesis that is confirmed by direct comparison of cross-validation results with and without dose-specific labeling in the dataset (AUC: 82.7% vs. 69.3%). It also suggests that the carcinogenicity classifiers trained on the DrugMatrix are underperforming due to mislabeling and could be improved by the use of dose-specific carcinogenicity labels.

2.5.1.2 Comparison to published models.

We were also interested in comparing our predictive model to two published gene signatures: the Ellinger-Ziegelbauer et al. 2008 - 512-gene carcinogenicity signature (Ellinger-Ziegelbauer et al. 2008) and the Fielden et al. 2011 - 23-gene non-genotoxic carcinogen signature (Fielden et al. 2011). To this end, the two published signatures and associated predictive models were trained on the DrugMatrix and tested on TG-GATEs (Table I.13, see Methods). Our model performed considerably better in predicting all carcinogenic compounds (AUC: 76.64 vs. 61.75 and 69.56). For the Fielden et al 23-gene signature, we also performed a cross-validation within the DrugMatrix using only non-genotoxic compounds, which resulted in an AUC of 62.59.

2.5.1.3 Carcinogenicity is a complex phenotype.

Supervised analysis of the DrugMatrix (differential analysis and GSEA) shows that the "exposure to carcinogens" phenotype is not adequately modeled as a simple dichotomy, especially when we consider the non-genotoxic carcinogens. This is reflected in the results of the differential analysis (see online portal web portal (Gusenleitner et al. 2013)), where, due to the very large sample size, a considerable number of genes are identified as significantly differentially expressed (554). However, inspection of their fold-changes (i.e., the ratio of their within-class mean expressions), as well as of the heatmap of the top markers, suggest that the differential signal is driven by relatively small subsets of compounds where the exposure induces a very marked up- or down-regulation. GSEA also supports this conclusion, as shown by the lower number of gene sets significantly enriched in the signature of non-genotoxic carcinogens as compared with the genotoxic carcinogens. As previously noted, this likely reflects the existence of multiple mechanisms of non-genotoxic carcinogenesis, all of which cannot be adequately captured by a simple dichotomous categorization. The heterogeneity of the phenotype also helps explain the superior performance of the Random Forest, a classifier based on an ensemble of decision trees. The decision tree formalism naturally lends itself to address classification problems that can be partitioned into sub-problems each governed by a possibly distinct classification rule. This formalism fits well the nature of our phenotype, since we can expect different classification rules to apply to different compound groups governed by distinct mechanisms of action.

2.5.1.4 Adequacy of compound sample size.

Although the gene expression datasets analyzed are comparatively large, the number of chemicals tested is still relatively limited, representing only ~9% of the compounds for which carcinogenicity annotation is available (and less than 0.16% of the compounds on the market). Additionally, a disproportionate number of compounds analyzed in the DrugMatrix act through the peroxisome-proliferating receptor (PPAR) pathway, hence compounds acting through other mechanisms of action might not be adequately represented. Our down-sampling simulation analysis aimed at evaluating sample size adequacy shows that the classification learning curve (see Figure 2.4) does not reach a plateau, thus suggesting that inclusion of additional compounds spanning a wider range of mechanisms of actions will enable the training of more precise classifiers, as well as the identification of a more extensive taxonomy of pathways relevant to carcinogenesis.

2.5.1.5 Gene set projection and interpretability vs. accuracy tradeoff.

Projection of the expression data matrix into gene set space, and subsequent classification using the gene sets as predictors, had the dual advantage of increasing the interpretability of the model (by identifying pathways and processes relevant to cancer induction) and of making it more robust across datasets (by correcting for batch-to-batch bias). However, it adversely impacted the predictive accuracy modestly within datasets (see Figure 2.3). Consequently, the choice of whether or not to adopt gene set projection will depend on the expected difference between the training set and the new profiles to be classified. We hypothesize that an increased sample size (number of compounds) will reduce the difference in predictive accuracy between gene-based and gene set-based prediction, and thus make the interpretability of the latter approach the major determinant of its choice. In this study, we relied on pre-defined gene sets as defined in the MSigDB repository. However, we recognize that an alternative, fully data-driven approach is possible, where unsupervised clustering methods can be applied toward the identification of sets of tightly co-regulated genes and the corresponding groups of samples (compounds) defining their “co-regulation context”. Combined with techniques of pathway annotation, this approach might lead to the definition of gene sets more relevant to the task of predicting carcinogenicity while maintaining their biological interpretability.

2.5.1.6 Optimal number of genes to assay.

The availability of data from a whole-transcriptome array allowed us to evaluate the dependency of a classifier's performance on the number and identity of the genes used as predictors, and to determine what would be a sufficient number of gene markers to include in a custom array designed to model chemical carcinogenicity. As noted, the selection of the top 500 genes as ranked by variance (rigorously carried out within the training set of each training-/test-set split) was sufficient to train a Random Forest classifier with highest predictive accuracy. Increasing the number of genes to 1000 or more, or replacing the variance ranking with a t-score ranking (with respect to the phenotype to be predicted) did not measurably affect the predictive accuracy (see Table I.14 and Table I.15). Similarly, by selecting the 2nd set of top 500 genes (i.e., from the 501st to 1000th genes ranked by variance), the 3rd set, etc., predictive accuracy decreased only marginally (see Table I.16). These results confirm the often-made observation that the effective dimensionality of gene expression data is well below the nominal number of genes profiled in the array, and that considerable redundancy among genes exists. Since predictive accuracy alone does not provide a high enough resolution to fully drive gene selection, interpretability and biological relevance will need to be used as additional criteria to guide inclusion.

2.5.1.7 *From predictive models to mechanisms of action.*

Using the pathway projection, we were able to identify modules of coordinated gene sets, each reflecting a different biological process that likely contributes to a compound's MoA. These tentative modules are in concordance with findings in published literature (Holsapple et al. 2006) and include a metabolism of lipids and lipoproteins module in parallel with the PPARα signaling pathway, damage response (p53 pathway) and proliferation (DNA replication pre-initiation) modules. A notable example for the power of this approach is represented by the group of seven genotoxic compounds. Genotoxic compounds induce direct DNA modifications and cells respond by up-regulation of components of the damage response machinery, such as the p53 pathway and the G2 pathway, outcomes captured in one of the modules.

Novel findings include the identification of a suppressed normal liver function module (complement cascade, platelet aggregation plug formation, as well as classic, common and extrinsic pathways). This is particularly intriguing since it emphasizes the potential role played by loss of normal tissue function in carcinogenesis. Equally of notice was the identification of a module reflecting up-regulation of the proteasome in response to carcinogens. The proteasome is closely tied to ribosome function, which is in turn linked to cell proliferation.

Even though we have a large number of profiles at our disposal (2195 liver samples in the Drugmatrix), there are only 127 well-annotated tested compounds and only 41 of these are known hepatocarcinogens in rodents. Furthermore, there are various

(>5) mechanisms of action, as shown in Figure 2.6, through which carcinogens can act. The Random Forest, coupled with variable importance ranking is successful in disentangling these mechanisms and provides a data-driven definition of their biological meaning; however, a larger number of compounds will be necessary to exhaustively define the carcinogenome.

2.6 Conclusions and future outlook

Toxicogenomic short-term exposure studies based on in-vivo (rat) models remain expensive and time consuming and therefore limit the number of chemical compounds that can be tested. Furthermore, as noted, animal models make for an imperfect proxy to test human carcinogenicity. To address both these shortcomings, the next generation of toxicogenomics tests is poised to rely on *in vitro* human models amenable to high-throughput screening (Interagency Breast Cancer and Environmental Research Coordinating Committee (IBCERCC) 2013; Leffall & Kripke 2010). This transition will introduce new challenges, including the accurate translation of in-vitro chemical doses to in-vivo relevance, as well as the need for adoption of organotypic culture models capable of capturing the cross-talk between multiple cell types. Further development of computational methods that accurately map the chemical response to activation/inactivation of relevant pathways of carcinogenicity will become essential to provide the essential link between the exposure and the adverse phenotype.

3 MOLECULAR CLASSIFIERS FOR AGGRESSIVE B-CELL LYMPHOMAS

The work in this chapter was published:

Carey CD, Gusenleitner D, Chapuy B, Kovach AE, Kluk MJ, Sun HH, Crossland RE, Bacon CM, Rand V, Dal Cin P, Le LP, Neuberg D, Sohani AR, Shipp MA, Monti S, Rodig SJ. 2015. Molecular Classification of MYC-Driven B-Cell Lymphomas by Targeted Gene Expression Profiling of Fixed Biopsy Specimens. *The Journal of molecular diagnostics : JMD*, 17(1), pp.19–30.

3.1 Abstract

Burkitt lymphoma (BL) and diffuse large B-cell lymphoma (DLBCL) are aggressive tumors of mature B-cells that are distinguished by a combination of histomorphologic, phenotypic, and genetic features. A subset of B-cell lymphomas, however, has one or more characteristics that overlap BL and DLBCL, and are categorized as B-cell lymphoma unclassifiable, with features intermediate between BL and DLBCL (BCL-U). Molecular analyses support the concept that there is a biological continuum between BL and DLBCL that includes variable activity of MYC, an oncoprotein once thought to be only associated with BL, but now recognized as a major predictor of survival among patients with DLBCL treated with R-CHOP. We tested whether a targeted expression profiling panel could be used to categorize tumors as BL and DLBCL, resolve the molecular heterogeneity of BCL-U, and capture MYC activity using RNA from formalin-fixed paraffin embedded biopsies. A diagnostic molecular classifier accurately predicted pathological diagnoses of BL and DLBCL, and provided more objective sub-classification for a subset of BCL-U and genetic “double-hit” lymphomas as molecular BL or DLBCL. A molecular classifier of MYC activity

correlated with MYC IHC and stratified patients with primary DLBCL treated with R-CHOP into high- and low-risk groups. These results establish a framework for classifying and stratifying MYC-driven, aggressive B-cell lymphomas based upon quantitative molecular profiling that is applicable to fixed biopsy specimens.

3.2 Introduction

The World Health Organization (WHO) classification of tumors defines neoplastic diseases according to unique clinical and biological characteristics (Campo et al. 2011). Burkitt lymphoma (BL) and diffuse large B-cell lymphoma (DLBCL) are aggressive tumors of mature B-cells categorized as individual tumor types. The reliable differentiation of BL from DBLCL is important, as these tumors are treated with distinct chemotherapeutic regimens (Magrath et al. 1996; Habermann et al. 2006).

BL is a neoplasm composed of monomorphic, intermediate-sized lymphocytes that are positive for markers of mature, germinal-center B-cells and negative for the anti-apoptotic protein BCL2. The vast majority of cells (>95%) are positive for the proliferation marker Ki67/MIB1. The genetic hallmark of BL is a balanced translocation involving the *MYC* oncogene and, most commonly, the immunoglobulin heavy chain locus (*IGH*) (Campo et al. 2011; Hecht & Aster 2000). Mutations in *TCF3* and *ID3* are also common (Schmitz et al. 2012; Love et al. 2012). In contrast, DLBCL is composed of pleomorphic, large lymphoid cells and, in general, less apoptosis and a lower proliferation index than BL. DLBCLs express markers of mature B-cells, with or without evidence of germinal center cell derivation, and a majority express BCL2. Genetically, only a small subset of DLBCLs have a *MYC* translocation and mutations in *TCF3* or *ID3* are rare. However, mutations in genes encoding the components of the NF-kB and B-cell receptor signaling pathways are common (Campo et al. 2011; Zhang et al. 2013; Morin et al. 2013; Lohr et al. 2012; Savage et al. 2009; Barrans et al. 2010).

Most cases of BL and DLBCL are diagnosed with high confidence using traditional histopathologic, immunophenotypic, and targeted genetic analyses. However, it is not uncommon to encounter tumors with one or more features overlapping BL and DLBCL. The 2008 WHO Classification of Lymphoid Tumors recognized these cases with the novel diagnostic category, “B-cell lymphoma unclassifiable, with features intermediate between DLBCL and BL” (BCL-U) (Campo et al. 2011). BCL-U is, by definition, a heterogeneous group, and its diagnosis requires that pathologists make subtle distinctions in histomorphology, immunophenotype, and genetics that may not be highly reproducible.

Molecular classification of aggressive B-cell lymphomas using comprehensive gene-expression profiles (GEPs) of RNA isolated from frozen tumor samples accurately differentiates BL from DLBCL and confirms that a subset of cases has transcriptional signatures intermediate between BL and DLBCL (Dave et al. 2006; Hummel et al. 2006). However, the pathological diagnoses corresponding to these ‘biologically intermediate’ tumors have been inconsistent (Hummel et al. 2006).

Complicating the evaluation of aggressive lymphomas is the recognition that high MYC expression and biological activity, once thought to be only associated with BL, are major, independent predictors of poor clinical outcome among patients with primary DLBCL treated with R-CHOP (Johnson et al. 2012; Kluk et al. 2012; Zhou et al. 2014; Cook et al. 2014; Perry et al. 2014). In some series, the prognostic value of MYC is enhanced among tumors that co-express BCL2 (Johnson et al. 2012; Green et al. 2012; Horn et al. 2013; Hu et al. 2013). Indeed recent evidence suggests that high co-expression of MYC and BCL2 in tumor cells provides a biological basis for the inferior outcome among patients with the activated B-cell (ABC) type DLBCL when treated with standard chemotherapy (Hu et al. 2013).

DLBCL with high MYC activity cannot be identified with certainty by morphologic or genetic studies alone (Kluk et al. 2012). The detection of MYC in fixed tumor biopsy specimens by immunohistochemistry (IHC) has the potential to identify DLBCLs with high MYC protein that corresponds to high MYC biological activity (Kluk et al. 2012). However, IHC methods are difficult to standardize between institutions and the interpretation of IHC staining is subjective (de Jong et al. 2007).

These data highlight a need for quantitative methods that capture the phenotypic, genetic and molecular heterogeneity of aggressive B-cell lymphomas in clinical practice. Molecular classification based upon the unique gene expression profiles of BL, DLBCL, and MYC-driven B-cell lymphomas has the potential to satisfy this need, but, until recently, gene expression profiling (GEP) has not been amenable to FFPE tissues

(Rimsza et al. 2011; Scott et al. 2014; Masqué-Soler et al. 2013; Linton et al.). Here we report a method of targeted expression profiling followed by a 2-stage molecular classifier of aggressive mature B-cell lymphomas that is applicable to FFPE biopsy specimens.

3.3 **Materials and Methods**

3.3.1 Tumor and Patient Cohorts

This study was performed with approval from the institutional review boards of Brigham and Women's Hospital (BWH) and Massachusetts General Hospital (MGH). For each case, one or both of the corresponding pathologists of this study (SJR and AS) reviewed hematoxylin and eosin (H&E) stained slides and the original diagnostic reports to ensure that the final diagnosis fulfilled 2008 WHO diagnostic criteria.

The training set (n=41) comprises 12 BLs and 29 DLBCLs (one additional DLBCL later failed analytical quality control). The BLs were selected based on the quality of available tissue and include all BL subtypes, as well as pediatric and adult patients (median age of diagnosis 30.5 years, range 3-62 years, *Supplementary Table 1*). The DLBCLs were selected from a previously published larger series of adult patients (Kluk et al. 2012) who had all been diagnosed as ‘DLBCL-NOS’ (DLBCL not otherwise specified). Previously, MYC IHC-High was defined as >50% expression in tumor cells, and MYC IHC-Low was defined as ≤50% (Kluk et al. 2012). For training, cases were deliberately selected in order to represent the ‘extremes’ of MYC IHC-High (median

70%; n=13) and MYC IHC-Low (median 20-30%; n=16) in order to assist development of the MYC activity classifier. We did not select DLBCLs for either the training or test sets on the basis of 'cell of origin' (COO) subtype (Alizadeh et al. 2000).

The test set (n=55) is composed of 9 BLs (all adult patients, 8 sporadic and 1 immunodeficiency-associated), 41 DLBCLs and 5 BCL-Us (Table II.1). Four additional cases failed analytical quality control. The DLBCLs included 1 'single hit lymphoma' (SHL), characterized by a *MYC*-rearrangement in isolation and 3 'genetic double hit lymphomas' (DHL), characterized by a combination of *MYC* and *BCL2*-rearrangements. The DLBCLs were chosen on the basis of the quality of available biopsy material and in order to represent a full range of MYC IHC expression. The BCL-Us were selected on the basis of available cases and were all DHLs. Four of 5 BCL-Us were characterized by a combination of *MYC* and *BCL2*-rearrangements and the remaining case had concurrent *MYC* and *BCL6*-rearrangements.

Patients included in an outcome cohort ('*Outcome series'*; n=40, 22 patients from the training set and 18 from the test set) were derived from a single institution (BWH). All had confirmed primary DLBCL and received standard immuno-chemotherapy (R-CHOP: rituximab, cyclophosphamide, doxorubicin, vincristine, prednisone) as previously reported (Kluk et al. 2012). All clinical data were collected prior to, and independent of, the reference and index tests reported in this study

3.3.2 Immunohistochemistry and Cytogenetic Analyses

MYC IHC was performed on 96 tumors using a rabbit monoclonal antibody (clone Y69, Epitomics/Abcam, cat. #ab32072) as described (Kluk et al. 2012). The status of the *MYC* locus was determined by fluorescence *in situ* hybridization (FISH) analysis for 96 tumors using Vysis LSI MYC "break-apart" probe set (cat. #05-J91-001), as described (Kluk et al. 2012). FISH analyses were performed on indicated cases using the *BCL2-IgH* dual fusion (cat. #05-J71-001), and *BCL6-IgH* "break-apart" (cat. #01N23-020; Abbott Laboratories, Abbott Park, IL) probe sets respectively, following manufacturer's recommendations. For a minority of cases, a karyotype was obtained as part of the original diagnostic evaluation (Kluk et al. 2012).

3.3.3 RNA Extraction and Profiling

FFPE tissue blocks were sectioned immediately prior to the RNA extraction. For each block, the initial 10µm section was discarded and 3x 10 µm subsequent sections were taken for analysis. If the estimated surface area of lesional tissue was < 5mm^2 an extra 10µm section was taken. Total RNA was isolated using the Qiagen RNeasy kit (catalog # 73504, Qiagen, Hilden, Germany) and quantified using Nanodrop spectrophotometry (Nanodrop Products, Thermo Science, DE). RNA was diluted to 150-200ng / 5µL, aliquoted and stored at -80°C until use.

For the multiplexed, digital gene expression analysis, 150-200ng of RNA for each sample was hybridized with 20µl of reporter probes / reaction buffer and 5 µl of capture probes at 65°C for 20 hours. The hybridized samples were then processed on the NanoString nCounter preparation station for 2.5 hours and expression data were subsequently generated on the NanoString nCounter digital analyzer (NanoString Technologies, Seattle, WA) using the 600 fields of view setting over 4 hours (Geiss et al. 2008). In total tumors from 96 patients were profiled, with a further 5 tumors (5%) failing analytical quality control.

3.3.4 Target Selection for the Initial and Final Profiling Panels

Candidate gene targets were initially selected from published GEPs of BL and DLBCL (Dave et al. 2006; Hummel et al. 2006) with preference given to genes within the TCF3/ ID3 signaling pathway (Schmitz et al. 2012), published MYC targets (Zeller et al. 2003; Mori et al. 2008; Schuhmacher et al. 2001; Kim et al. 2006; Chapuy et al. 2013; Schlosser et al. 2005; Yu et al. 2005), and GEPs of frozen tissue corresponding to DLBCL samples in the training set (Monti et al. 2012). These were supplemented by additional targets of interest including housekeeping genes (Figure II.1*).*

The initial panel of 200 probes included 37 unique transcripts distilled from a previously published "TCF3 signature" (Schmitz et al. 2012). These were subsequently validated, by *in silico* differential analysis (DA), as best distinguishing BL from DLBCL in two independent series of B-cell non-Hodgkin lymphomas (Dave et al. 2006; Hummel et al. 2006) (Figure II.1). The panel also included transcripts from 7 published datasets of MYC targets (101 targets) (Zeller et al. 2003; Mori et al. 2008; Schuhmacher et al. 2001;

Kim et al. 2006; Chapuy et al. 2013; Schlosser et al. 2005; Yu et al. 2005) that were validated (False Discovery rate (FDR) < 0.25; fold change (FC) > 1.3) by DA against Affymetrix U133 microarray GEPs of frozen DLBCLs with corresponding MYC IHC scores from matched FFPE tissue in the training cohort(Kluk et al. 2012; Monti et al. 2012) and differentially expressed genes suggested by DA of the GEPs of frozen DLBCLs with corresponding MYC IHC scores (FDR < 0.25; FC >2.0). Finally they were supplemented with BCL2 and related family members (5 targets), "house-keeping" control transcripts (15 targets), and select markers of specific cell lineages (CD3e, CD68, CD19, CD79a, CD20; Table II.2)

The final profiling panel, targeting 80 transcripts was derived by analyzing data from the training set, by both ranking the importance of each included gene and estimating how many could be excluded without compromising the predicted accuracy (Table II.3). The predicted accuracy of each classifier was assessed on the training set using leave-one-out cross-validation (LOO-CV), as well as on an independent test dataset.

3.3.5 Housekeeping gene transcripts

Six housekeeping (HK) genes were selected based on the following criteria: i) low variation across samples; ii) even coverage along the expression range; iii) exclusion of the most highly expressed HK genes, since at very high levels, the variation level of the HK genes is comparable to the variation of the other genes, and iv) exclusion of genes within regions of known recurrent copy number alteration in lymphoma (Monti et al. 2012). Based on these criteria, we selected the following 6 gene targets: AAMP, HMBS, KARS, PSMB3, TUBB, and H3F3A.

3.3.6 Data normalization

Data from the preliminary targeted profiling panel (200 genes) and the final profiling panel (80 genes) were cross-normalized using expression data from 6 cases tested with both panels. Normalization of the NanoString data was performed using the R package NanoStringNorm (http://cran.rproject.org/web/packages/NanoStringNorm). We used the sum of the expression values to estimate the technical assay variation, the mean to estimate background count levels and the sum of the six housekeeping genes to normalize for the RNA sample content. Additionally, the data were log2 transformed.

3.3.7 Classification models

Classification models were selected based on the training cohort using a bootstrapping scheme, where 75% of the samples were drawn to train a classification model, which was then tested on the remaining 25% of the samples, with the train/test split repeated 100 times. Elastic nets (Hui Zou 2005), linear and polynomial support vector machines (SMV), shrunken centroids (Tibshirani et al. 2002) and a random forest algorithm (Breiman 2001) were evaluated as candidate prediction models. For the MYC activity classifier, samples with MYC IHC >50% and ≤50% were classified as MYC IHC-High and IHC-Low, respectively, and these labels were used in the training of the MYC activity classifier (Kluk et al. 2012).

Features were selected based on differential expression, and their number determined based on LOO-CV performed on the training cohort; based on this procedure, 21 genes were used for the diagnostic classifier and 61 for the MYC classifier (Table II.4). Based on their performance on the training dataset we selected the elastic net with an alpha parameter of 0.1 and a lambda of 0.1 as the classifier of choice for both stages.

Classification accuracy of the final elastic net models was assessed on the training cohort using LOO-CV and comparing the predictions with the outcome of the IHC staining. Unbiased validation was then performed by training elastic net models on the entire training dataset and applying them to the classification of cases in the test cohort.

3.3.8 Creation and Validation of the Molecular Classifiers

An elastic net (Hui Zou 2005) prediction model was selected for both classifiers, based on a bootstrapping evaluation scheme on the training set. The Diagnostic classifier comprised 21 genes, the MYC activity classifier 61 genes, and 8 genes were common to both classifiers (Table II.4). In addition, 6 housekeeping genes complete the final profiling panel.

Elastic net models output class probabilities between 0 and 1 for each class (probability of class BL in the diagnostic classifier, and of class MYC IHC-High in the

MYC Transcriptional Activity classifier), reflecting the confidence of a sample prediction. Prior to analysis, and in order to reflect the concept of a biological ‘intermediate’ between BL and DLBCL, we defined Diagnostic scores of >0.75 as representing ‘molecular BL’ (mBL), <0.25 ‘molecular DLBCL’ (mDLBCL), and 0.25-0.75 ‘molecularly intermediate’, respectively. MYC activity scores of "1" and “0” correspond to tumors with high MYC and low MYC (as modeled on IHC expression (Kluk et al. 2012)) with greatest probability, respectively (detailed in *Supplementary Methods*). During development of the MYC activity classifier 0.5 was optimized as the cut-off with the highest estimated accuracy to classify tumors with high and low MYC activity. Therefore 0.5 is used for statistics regarding the efficacy of the classifier and for correlation to clinical outcome.

3.3.9 Reproducibility of the Assay

The test set and outcome series were profiled using 2 ‘builds’ of the 80-gene profiling panel. The binding efficiency of probes varies between builds and therefore the final dataset was compiled by normalizing to both housekeepers and then between builds, using on the expression profiles of tumor RNA that were profiled on both. RNA from a subset of cases was profiled multiple times over the course of the study to determine the reproducibility of the assay (Figure II.2).

3.4 Results

RNA was isolated from FFPE tissue corresponding to 41 aggressive B-cell lymphomas (training set) and was profiled using an initial panel targeting 200 unique transcripts (Figure 3.1). The resulting data were used to derive a pair of molecular classifiers, firstly to distinguish BL from DLBCL and secondly to distinguish high and low MYC activity in DLBCL using a parsimonious, 80-gene signature (Figure 3.1 and Table II.3; Table II.4).

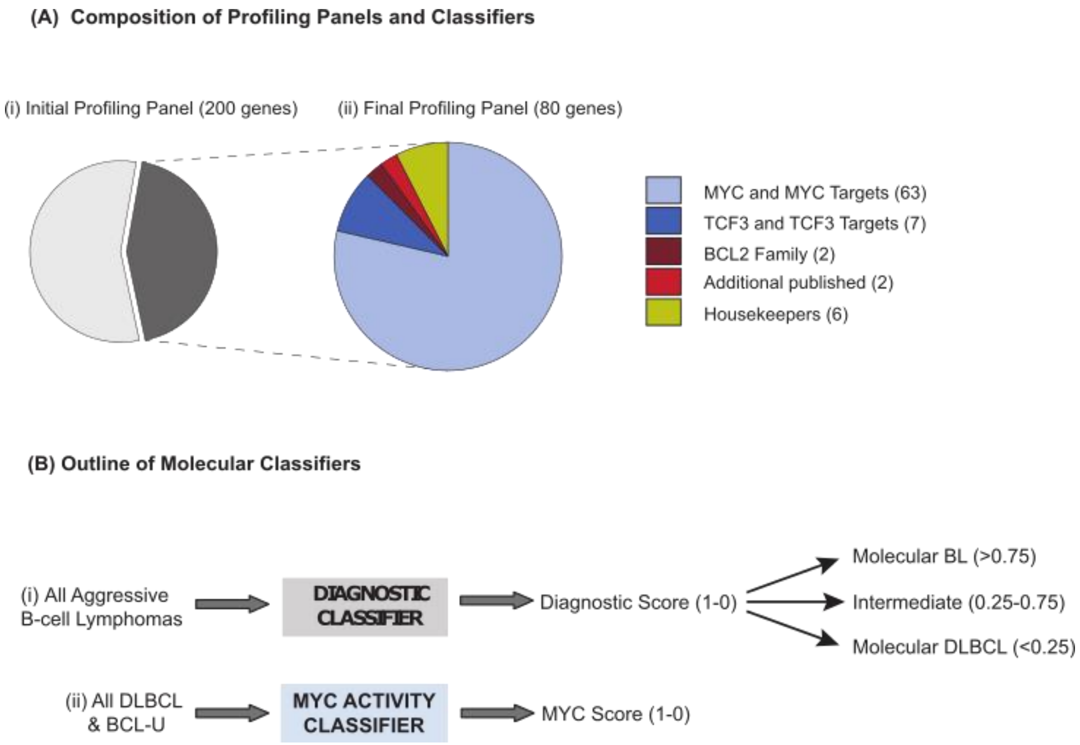


Figure 3.1 Target gene selection and the creation of molecular classifiers.

(A) Schematic showing the distribution of gene transcripts that were assayed in the initial and final profiling panels. (B) Schematic outlining the protocols for the molecular classification of (i) all aggressive B-cell lymphomas, and (ii) cases with the pathological diagnosis of DLBCL and BCL-U. Twenty-one genes were used

for the Diagnostic classifier and 61 genes were used for the MYC Activity classifier. In addition 8 genes were common to both classifiers and there were 6 housekeeping genes.

3.4.1 Unsupervised Clustering of Targeted Expression Profiles of Select Lymphomas

Unsupervised clustering of the normalized expression data from the 200-gene signature segregated the 42 tumors into distinct groups that showed a close correlation (39/42, 92.9%) with the original pathological diagnoses of BL, DLBCL MYC IHC-High, and DLBCL MYC IHC-Low (Figure 3.2). Figure 3.2 was constructed using 'one minus Pearson' hierarchical clustering in Gene-e[6] for the training dataset. This dataset was first normalized using Nanostring's 'nSolver' software package and then transformed to Log2 using Gene-e prior to hierarchical clustering. One case, diagnosed as BL, clustered with DLBCL MYC IHC-High cases. Central review of this case confirmed that the tumor was originally diagnosed correctly. These data support the *in silico* methods used to develop the initial profiling panel and demonstrate the technical feasibility of the approach to broadly group aggressive lymphomas into clinically relevant categories.

[6] http://www.broadinstitute.org/cancer/software/GENE-E/

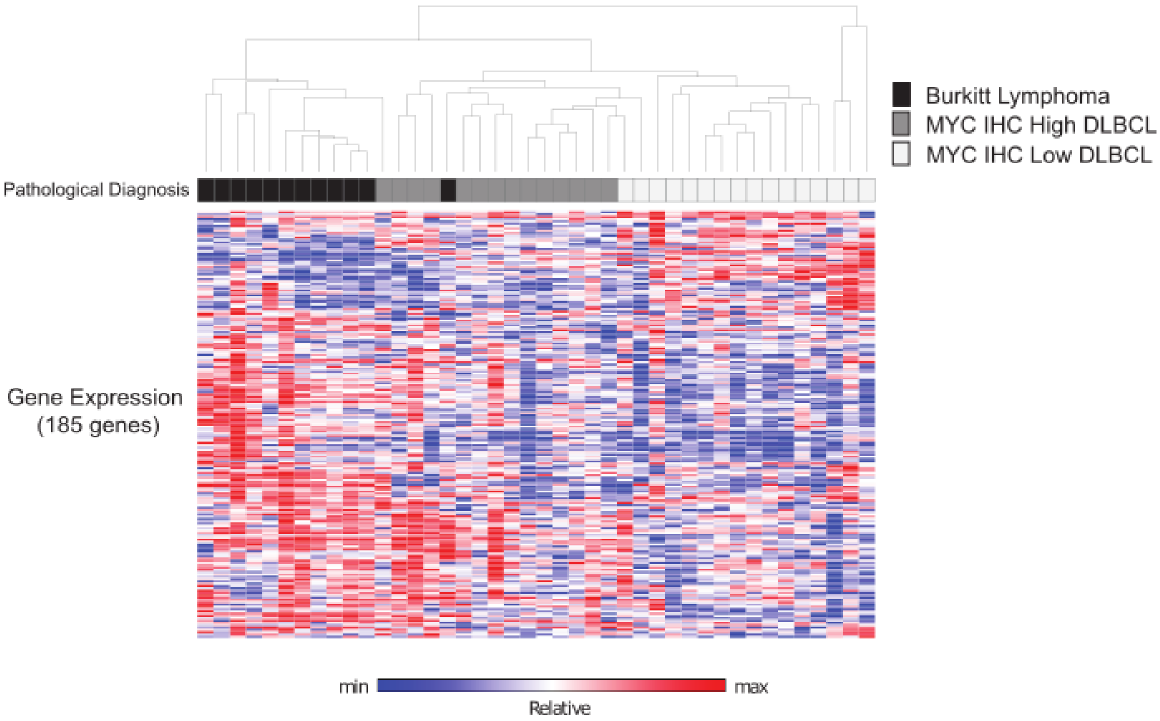


Figure 3.2 Heatmap of trainings cohort

Unsupervised clustering of the normalized transcript values from 42 tumors comprising the training cohort and including all probes in the initial profiling panel (1 case, DLBCL30 later failed quality control during classification). The original pathological diagnosis (first line) and relative gene expression for the 185 genes comprising the initial profiling panel (heatmap) are shown (housekeeping genes excluded).

3.4.2 Performance of the Diagnostic Molecular Classifier on the Training and Test sets

We tested the diagnostic molecular classifier against data derived from the training set in a LOO-CV, Figure 3.3). When ranked by the diagnostic classifier scores, these data recapitulated the results obtained from the original unsupervised clustering analysis using the 200-gene panel. Thirty-five of 41 (85%) cases classified as mBL or mDLBCL with high confidence and correctly matched the pathological diagnoses of BL or DLBCL, respectively (Figure 3.3, Table 3.1). Six cases had diagnostic scores of >0.25 and <0.75 and thus were not assigned to the categories of mBL or mDLBCL. Nevertheless, 3 of the "molecularly intermediate" cases had a pathological diagnosis of BL and 2 of these had a diagnostic score >0.5; three "molecularly intermediate" cases had a pathological diagnosis of DLBCL and 2 had a diagnostic score < 0.5. We conclude that in our training cohort a 21-gene classifier can be used to distinguish the majority (85%) of pathological BL from DLBCL.

Table 3.1: Performance Statistics of Molecular Classifiers

Diagnostic Classifier: Only cases classified with high confidence (as mBL or mDLBCL) are included. The sensitivity refers to the ability of the test to identify pathological BL as molecular BL ('mBL').

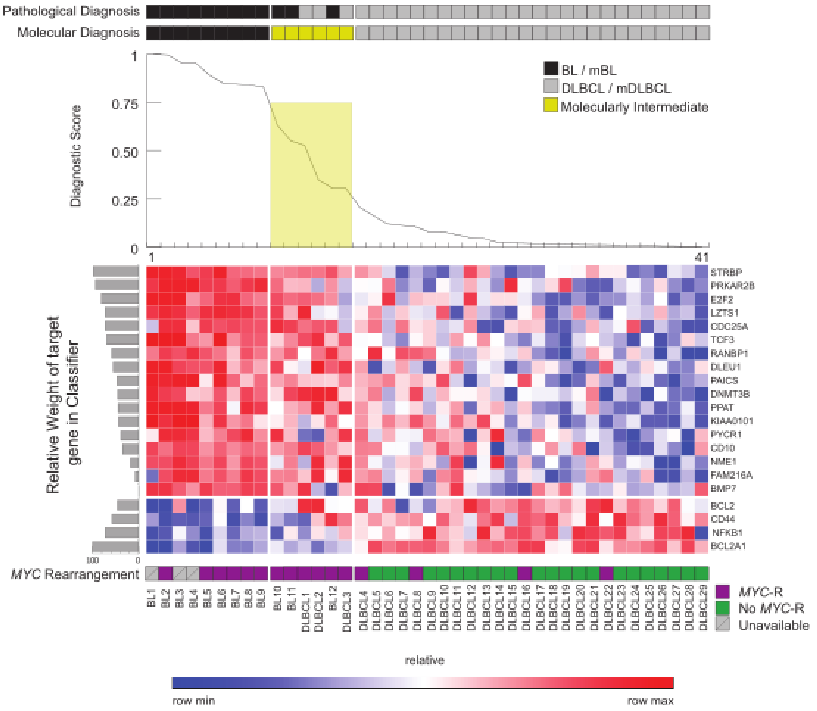
MYC Activity Classifier: Only cases with matched MYC IHC and MYC Activity scores are included. The sensitivity refers to the ability of the test to identify tumors with high MYC IHC expression (>50%) as having MYC Activity score >0.5.

	Diagnostic Classifier*		**MYC Activity Classifier¶**				
	Training	Test	Training (All)	Training (non-BL)	Test (All)	Test (non-BL)	Outcome Series
Percentage of cases classified	85%	92%	100%	100%	100%	100%	100%
Accuracy	1	1	0.925	0.897	0.804	0.795	0.868
Sensitivity (95% CI)	1 (0.66-	1 (0.59-	0.917 (0.73-	0.846 (0.55-0.98)	0.773 (0.55-	0.688 (0.41-	0.750 (0.35-0.96)

	1.0)	1.0)	0.99)		0.92)	0.89)	
Specificity (95% CI)	1 (0.87-1.0)	1 (0.91-1.0)	0.938 (0.70-0.90)	0.938 (0.70-0.90)	0.828 (0.64-0.94)	0.857 (0.67-0.96)	0.900 (0.73-0.98)
Positive Predictive Value (PPV)	1	1	0.957	0.917	0.773	0.733	0.667
Negative Predictive Value (NPV)	1	1	0.882	0.882	0.828	0.828	0.931

We next profiled and classified an independent test set of 55 cases that included 9 BL, 41 DLBCL, and 5 cases with the pathological diagnosis of BCL-U (Figure 3.3*B*). Among the non-BLs were one genetic “single-hit” lymphoma (genetic SHL, with isolated *MYC*-translocation, tDLBCL1) and 8 genetic “double-hit” lymphomas (DHLs), all with *MYC*-translocations. 7 DHLs had coexistent *BCL2*-translocation and one had a coexistent *BCL6*-translocation, tDHL1-8, (Figure 3.3*B*).

A.)


Pathological Diagnosis
Molecular Diagnosis
Diagnostic Score
1
0.75
0.50
0.25
0
1
41
BL / mBL
DLBCL / mDLBCL
Molecularly Intermediate
Relative Weight of target gene in Classifier
100
0
STRBP
PRKAR2B
E2F2
LZTS1
CDC25A
TCF3
RANBP1
DLEU1
PAICS
DNMT3B
PPAT
KIAA0101
PYCR1
CD10
NME1
FAM216A
BMP7
BCL2
CD44
NFKB1
BCL2A1
MYC Rearrangement
MYC-R
No MYC-R
Unavailable
BL1
BL2
BL3
BL4
BL5
BL6
BL7
BL8
BL9
BL10
BL11
DLBCL1
DLBCL2
BL12
DLBCL3
DLBCL4
DLBCL5
DLBCL6
DLBCL7
DLBCL8
DLBCL9
DLBCL10
DLBCL11
DLBCL12
DLBCL13
DLBCL14
DLBCL15
DLBCL16
DLBCL17
DLBCL18
DLBCL19
DLBCL20
DLBCL21
DLBCL22
DLBCL23
DLBCL24
DLBCL25
DLBCL26
DLBCL27
DLBCL28
DLBCL29
relative
row min
row max

B.)

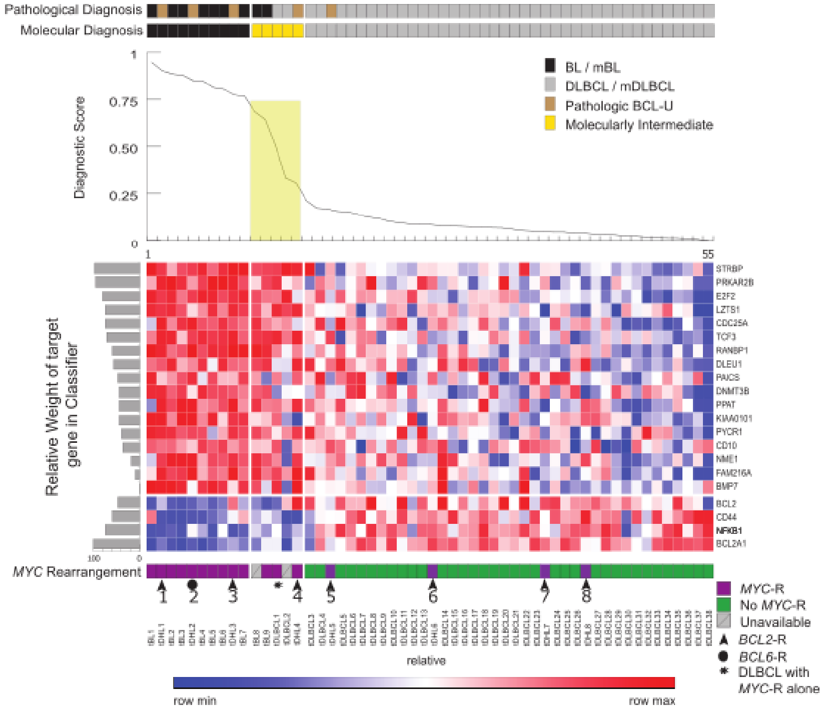

Pathological Diagnosis
Molecular Diagnosis
Diagnostic Score
1
0.75
0.50
0.25
0
1
55
BL / mBL
DLBCL / mDLBCL
Pathologic BCL-U
Molecularly Intermediate
Relative Weight of target gene in Classifier
100
0
STRBP
PRKAR2B
E2F2
LZTS1
CDC25A
TCF3
RANBP1
DLEU1
PAICS
DNMT3B
PPAT
KIAA0101
PYCR1
CD10
NME1
FAM216A
BMP7
BCL2
CD44
NFKB1
BCL2A1
MYC Rearrangement
1
2
3
*
4
5
6
7
8
tBL1
tDHL1
tBL2
tBL3
tDHL2
tBL4
tBL5
tBL6
tDHL3
tBL7
tBL8
tBL9
tDLBCL1
tDLBCL2
tDHL4
tDLBCL3
tDLBCL4
tDHL5
tDLBCL5
tDLBCL6
tDLBCL7
tDLBCL8
tDLBCL9
tDLBCL10
tDLBCL11
tDLBCL12
tDLBCL13
tDHL6
tDLBCL14
tDLBCL15
tDLBCL16
tDLBCL17
tDLBCL18
tDLBCL19
tDLBCL20
tDLBCL21
tDLBCL22
tDLBCL23
tDHL7
tDLBCL24
tDLBCL25
tDLBCL26
tDHL8
tDLBCL27
tDLBCL28
tDLBCL29
tDLBCL30
tDLBCL31
tDLBCL32
tDLBCL33
tDLBCL34
tDLBCL35
tDLBCL36
tDLBCL37
tDLBCL38
relative
row min
row max
MYC-R
No MYC-R
Unavailable
BCL2-R
BCL6-R
DLBCL with MYC-R alone

Figure 3.3 Heatmaps of Diagnostic Classifier

(A) Leave-one-out cross-validation (LOO-CV) of the final profiling panel and Diagnostic Classifier for the training cohort: BL and DLBCL cases categorized according to the original pathological diagnosis (first line), the assigned molecular diagnosis (second line, diagnostic scores of 0.25-0.75 categorized as 'molecularly intermediate'), diagnostic score (line graph, third line, intermediate values shaded), the relative expression of the indicated transcripts (heatmap) including the relative contribution of each to the classifier (horizontal shaded bar graphs, left side), and MYC-rearrangement status (bottom line). (B) Results of the Diagnostic Classifier for the test cohort: BL, BCL-U and DLBCL cases categorized according to the original pathological diagnosis (first line), the assigned molecular diagnosis (second line, diagnostic scores of 0.25-0.75 categorized as 'intermediate'), diagnostic score (line graph, third line, intermediate values shaded), the relative expression of the indicated transcripts (heatmap) including the relative contribution of each to the classifier (horizontal shaded bar graphs, left side), and MYC-rearrangement status (bottom line). The cases of genetic DHL are numbered and additional gene rearrangements are indicated by arrowheads (BCL2-) or a dot (BCL6-). The 'single hit' DLBCL, with MYC-rearrangement only, is indicated by an asterisk.

The diagnostic classifier successfully segregated all pathological BL from all DLBCL (Figure 3.3, Table 3.1). Forty-six of 50 (92%) BL and DLBCL were classified with high confidence. Two BL and two DLBCL had intermediate diagnostic scores, but among these, the diagnostic scores for the BL were >0.5 and for the DLBCL ≤0.5. The DLBCL with the highest diagnostic score (case tDLBCL1, score=0.5) was the genetic SHL. The diagnostic classifier demonstrated a sensitivity of 1.0 (95% CI 0.66 - 1.0) and specificity of 1.0 (95% CI 0.87 - 1.0) in the test set, for all tumors classified as mBL or mDLBCL (Table 3.1).

Molecular classification segregated subsets of non-BLs with the pathological diagnosis of BCL-U and/or genetic evidence for *MYC*-rearrangements into all three diagnostic categories (Figure 3.3*B*). Three BCL-U/ DHLs (tDHL1, tDHL2, tDHL3) had high diagnostic scores (0.90, 0.85, and 0.77, respectively) and classified as mBL. One DLBCL/ SHL (tDLBCL1) and one BCL-U/ DHL (tDHL4) had lower diagnostic scores (0.50 and 0.31, respectively) and classified as 'molecularly intermediate'. Finally, one BCL-U/ DHL (tDHL5) and three DLBCL/ DHLs (tDHL6, tDHL7, tDHL8) had low diagnostic scores (0.12, 0.05, 0.02, and 0.015, respectively) and classified as mDLBCL. We conclude that the diagnostic molecular classifier reveals molecular heterogeneity among BCL-Us and DLBCLs with *MYC*-translocations.

3.4.3 Molecular and Histopathological Features of BCL-U/ DHL

We next examined the molecular signatures and histopathology of the BCL-Us and DLBCLs with *MYC*-translocations in more detail (Figure 3.4). BCL-U/ DHLs classified as mBL expressed both TCF3-associated transcripts and MYC-associated transcripts at levels that were comparable to BL (Figure 3.4*A*). DLBCL/ DHLs classified as mDLBCL expressed TCF3-associated transcripts at low levels and MYC-associated transcripts at intermediate levels that were comparable to many DLBCLs lacking a *MYC*-translocation (Figure 3.4*A*). Additional transcripts (BCL2, CD44, NFKB1 and BCL2A1) differentially expressed between BL and DLBCL, also showed differential expression among the DHLs, and with the TCF3 and MYC signatures, resulted in the final classification indicated in Figure 3.3B.

Further review of the histopathology of the DHLs revealed distinct features between those that classify with high confidence as mBL and those that classified with high confidence as DLBCL (Figure 3.4*B*). Cases classified as mBL were composed of sheets of tightly packed, intermediate to large-sized cells with homogenous, round nuclei, and scant cytoplasm that resembles the morphological features of classic BL. In contrast, cases classified as mDLBCL were composed of large-sized lymphoid cells with marked pleomorphism and nuclear irregularity typical of DLBCL. We conclude that the molecular classifications of DHLs are supported by multiple molecular signatures, and correlate with distinct histopathological characteristics.

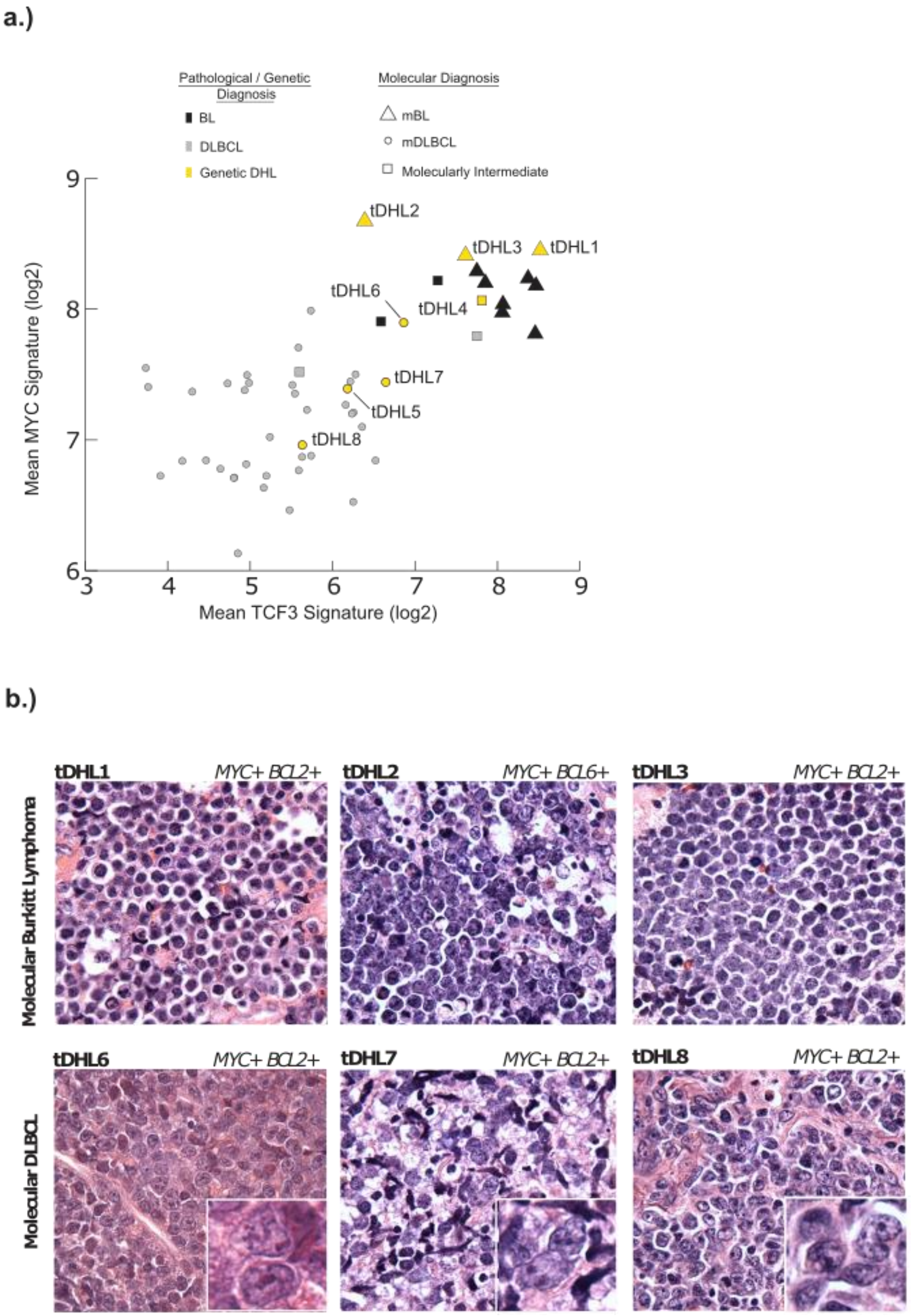

a.)
Pathological / Genetic Diagnosis
■ BL
■ DLBCL
■ Genetic DHL
Molecular Diagnosis
△ mBL
○ mDLBCL
□ Molecularly Intermediate
tDHL2
tDHL3
tDHL1
tDHL6
tDHL4
tDHL7
tDHL5
tDHL8
Mean MYC Signature (log2)
9
8
7
6
3
4
5
6
7
8
9
Mean TCF3 Signature (log2)
b.)
Molecular Burkitt Lymphoma
Molecular DLBCL
tDHL1
MYC+ BCL2+
tDHL2
MYC+ BCL6+
tDHL3
MYC+ BCL2+
tDHL6
MYC+ BCL2+
tDHL7
MYC+ BCL2+
tDHL8
MYC+ BCL2+

Figure 3.4 MYC signature overview

(A) Scatterplot showing the mean TCF3 signature (7 genes, x axis) and mean MYC signature (10 genes, y axis) for each tumor from the test cohort. The mean values for each signature are derived from transcript counts from these genes, as originally used in the diagnostic classifier. Colors indicate the pathological / genetic diagnoses (black for BL, gray for DLBCL, yellow for genetic DHL). Shapes indicate the molecular classification assigned by the diagnostic classifier (triangle for mBL, circle for mDLBCL, square for molecularly intermediate). (B) Histomorphological features of lymphomas with a MYC-rearrangement and either a BCL2- or BCL6- rearrangement (genetic DHL). Hematoxylin and eosin stained sections of DHL classified as molecular BL (top row), and molecular DLBCL (bottom row). Unique identifiers and details of relevant translocations are shown. Cases were photographed at x1000 original magnification. The tumors classified as mDLBCL have inserts highlighting nuclear morphology.

3.4.4 Performance of the MYC Activity Classifier on the Training and Test sets

The MYC activity classifier was tested in the training cohort by LOO-CV. BLs were not used to build the classifier but, as expected, had very high MYC activity scores (Figure 3.5*A*). In addition, all non-BLs with *MYC*-translocation had MYC activity scores >0.5. The sensitivity and specificity of the molecular classifier for identifying MYC IHC-High among all cases in the training set were 0.917 (95% CI 0.73-0.997) and 0.938 (95% CI 0.70-0.99), respectively (Table 3.1). Overall, the correlation between the optimized, molecular MYC activity score and MYC IHC score among non-BLs in the training set was high (Spearman *r*= 0.80, p <0.0001, 95% CI 0.6-0.9, Figure 3.5*A*).

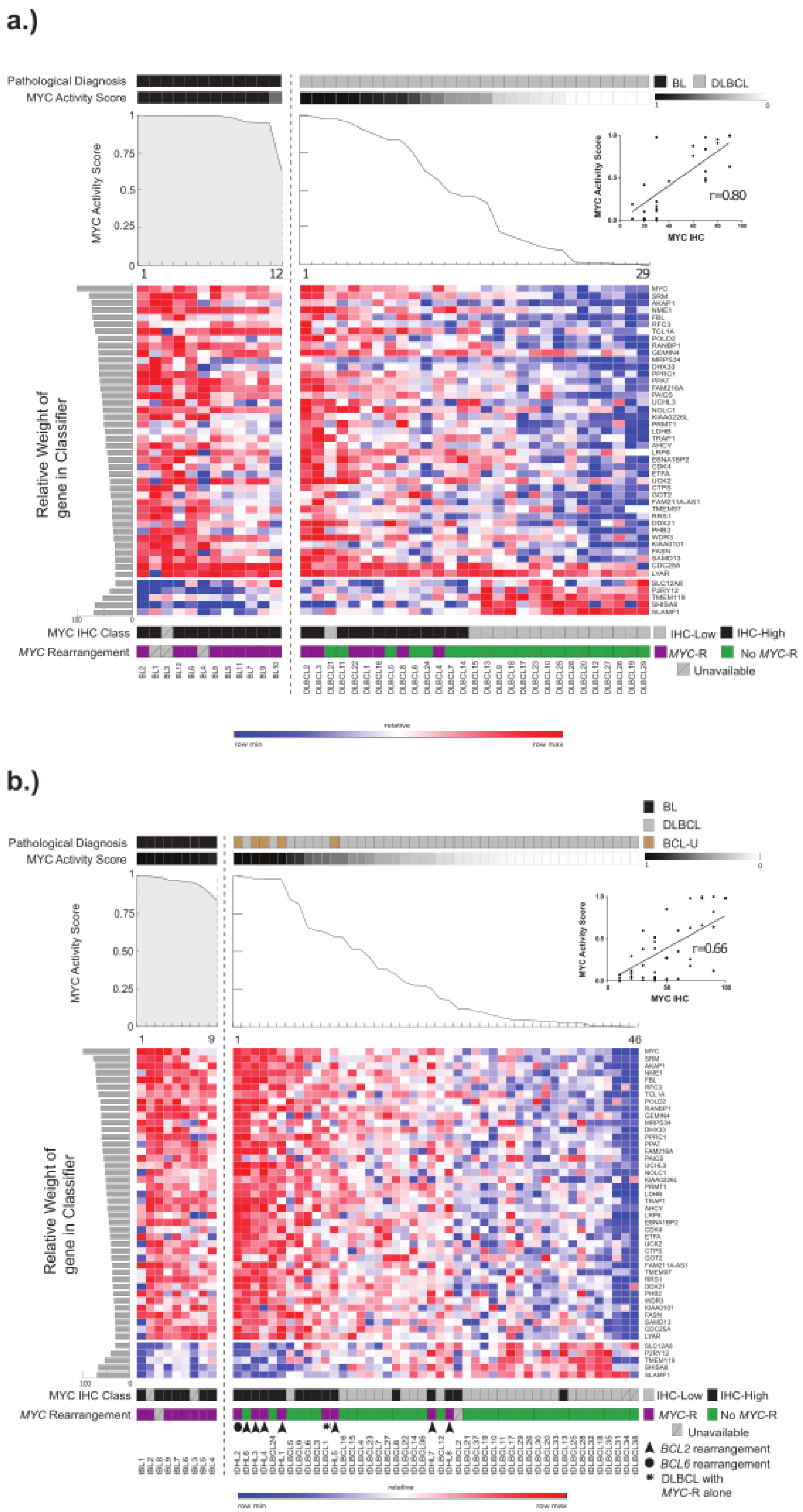

a.)
Pathological Diagnosis
MYC Activity Score
BL
DLBCL
1
0
MYC Activity Score
1
0.75
0.5
0.25
0
1
12
1
29
MYC Activity Score
MYC IHC
r=0.80
Relative Weight of gene in Classifier
100
0
MYC
SRM
AKAP1
NME1
FBL
RFC3
TCL1A
POLD2
RANBP1
GEMIN4
MRPS34
DHX33
PPRC1
PPAT
FAM216A
PAICS
UCHL3
NOLC1
KIAA0226L
PRMT1
LDHB
TRAP1
AHCY
LRP8
EBNA1BP2
CDK4
ETFA
UCK2
CTPS
GOT2
FAM211A-AS1
TMEM97
RRS1
DDX21
PHB2
WDR3
KIAA0101
FASN
SAMD13
CDC25A
LYAR
SLC12A8
P2RY12
TMEM119
SHISA8
SLAMF1
MYC IHC Class
IHC-Low
IHC-High
MYC Rearrangement
MYC-R
No MYC-R
Unavailable
BL2
BL1
BL3
BL12
BL6
BL4
BL8
BL5
BL11
BL7
BL9
BL10
DLBCL2
DLBCL3
DLBCL21
DLBCL11
DLBCL22
DLBCL1
DLBCL16
DLBCL5
DLBCL8
DLBCL6
DLBCL24
DLBCL4
DLBCL7
DLBCL14
DLBCL15
DLBCL13
DLBCL9
DLBCL18
DLBCL17
DLBCL23
DLBCL10
DLBCL25
DLBCL28
DLBCL20
DLBCL12
DLBCL27
DLBCL26
DLBCL19
DLBCL29
relative
row min
row max
b.)
BL
DLBCL
BCL-U
Pathological Diagnosis
MYC Activity Score
1
0
MYC Activity Score
1
0.75
0.50
0.25
0
1
9
1
46
MYC Activity Score
MYC IHC
r=0.66
Relative Weight of gene in Classifier
100
0
MYC
SRM
AKAP1
NME1
FBL
RFC3
TCL1A
POLD2
RANBP1
GEMIN4
MRPS34
DHX33
PPRC1
PPAT
FAM216A
PAICS
UCHL3
NOLC1
KIAA0226L
PRMT1
LDHB
TRAP1
AHCY
LRP8
EBNA1BP2
CDK4
ETFA
UCK2
CTPS
GOT2
FAM211A-AS1
TMEM97
RRS1
DDX21
PHB2
WDR3
KIAA0101
FASN
SAMD13
CDC25A
LYAR
SLC12A8
P2RY12
TMEM119
SHISA8
SLAMF1
MYC IHC Class
IHC-Low
IHC-High
MYC Rearrangement
MYC-R
No MYC-R
Unavailable
▲ BCL2 rearrangement
● BCL6 rearrangement
* DLBCL with MYC-R alone
tBL1
tBL2
tBL8
tBL9
tBL7
tBL6
tBL3
tBL5
tBL4
tDHL2
tDHL6
tDHL3
tDHL4
tDLBCL24
tDHL1
tDLBCL5
tDLBCL9
tDLBCL6
tDLBCL3
tDLBCL1
tDHL5
tDLBCL16
tDLBCL15
tDLBCL4
tDLBCL23
tDLBCL7
tDLBCL27
tDLBCL8
tDLBCL22
tDLBCL14
tDLBCL36
tDHL7
tDLBCL12
tDHL8
tDLBCL2
tDLBCL21
tDLBCL37
tDLBCL19
tDLBCL10
tDLBCL11
tDLBCL17
tDLBCL29
tDLBCL26
tDLBCL30
tDLBCL20
tDLBCL33
tDLBCL13
tDLBCL35
tDLBCL28
tDLBCL32
tDLBCL18
tDLBCL25
tDLBCL31
tDLBCL34
tDLBCL38
relative
row min
row max

Figure 3.5 Heatmaps of MYC classifier

(A) Leave-one-out cross-validation (LOO-CV) of the final profiling panel and MYC Activity Classifier for the training cohort. BL (left side) and DLBCL (right side) are segregated by pathological diagnosis (first line), MYC activity score (second line and line graph), the relative expression of the indicated transcripts (heatmap) including the relative contribution of each to the classifier (horizontal, shaded bar graphs, left side), MYC IHC class (MYC IHC-Low ≤50%, IHC-High >50%; penultimate line) and MYC rearrangement status (bottom line). Inset is the correlation between MYC IHC and MYC activity score for DLBCL only (Spearman r = 0.80; 95% CI 0.6-0.9). (B) Results of the final profiling panel and MYC Transcriptional Activity Classifier for the test cohort: BL (left side), DLBCL and BCL-U (right side) are segregated by pathological diagnosis (first line), MYC activity score (second line and line graph), the relative expression of the indicated transcripts (heatmap) including the relative contribution of each to the classifier (horizontal, shaded bar graphs, left side), MYC IHC class (MYC IHC-Low ≤50%, IHC-High >50%; penultimate line) and MYC rearrangement status (bottom line). Genetic DHLs are indicated as in Fig 3(B). Inset is the correlation between MYC IHC and MYC activity score for non-BL only (Spearman r = 0.66; 95% CI 0.44-0.8).

We next applied the MYC activity classifier to expression data from the independent test set. Again, BL cases showed very high MYC activity scores (Figure 3.5*B*). The sensitivity and specificity of the molecular classifier identifying MYC IHC-High among all cases were 0.773 (95% CI 0.55-0.92) and 0.828 (95% CI 0.64-0.94), respectively (Table 3.1). The correlation between the molecular MYC score and the MYC IHC score for the test set (non-BLs) was lower than for the LOO-CV of the training set, but with overlapping confidence intervals (Spearman *r*= 0.66, p <0.0001, 95% CI 0.44-0.8).

Non-BLs with a *MYC*-translocation were expected to have upregulated MYC activity, and for 5 of 9 cases, tDHL1-4 and tDHL6, the MYC activity scores were high and comparable to those seen for BL (ranging from 0.98-1.00). There was a range of values among the remaining cases. For tDLBCL1 (genetic SHL) and tDHL5, the MYC activity scores were 0.63 and 0.60 and for tDHL7 and tDHL8, the scores were lower at 0.26 and 0.18 respectively. Non-BLs with *MYC*-translocations and high MYC activity scores had a pathological diagnosis of BCL-U whereas those with other MYC activity scores had a pathological diagnosis of DLBCL. We conclude that the MYC activity classifier captures a spectrum of MYC biological activity in BCL-U and DLBCL that shows good correlation with MYC IHC and reveals heterogeneity in MYC biological activity among non-BL with *MYC* translocations.

3.4.5 Clinical Significance of the MYC Activity Score in DLBCL

The MYC activity classifier was constructed in order to categorize aggressive B-cell lymphomas according to MYC biological activity, rather than to predict clinical outcome. The MYC activity scores showed good, but not perfect, correlation with MYC IHC scores in the training and test sets. Therefore, we wished to determine whether the results of the MYC classifier were sufficient to predict clinical outcome in a series for which MYC IHC has prognostic value (Kluk et al. 2012). The overall accuracy of the classifier in the outcome series, compared to a class defined by MYC IHC, was high (0.868, Table 3.1), but the correlation between the MYC activity and MYC IHC scores was lower (*r*= 0.64, Figure 3.6*A*) than observed for non-BLs in the training and test sets (*r*= 0.80 and 0.66, respectively). DLBCLs with MYC activity scores in excess of the optimized classifier cut-point of 0.5 identified a patient population with inferior overall survival that was highly significant (nominal p = 0.0009, log-rank test; hazard ratio= 6.73, Figure 3.6*B*). We conclude that the MYC activity classifier, built upon MYC IHC data, is capable of dividing patients into high-risk and low-risk categories.

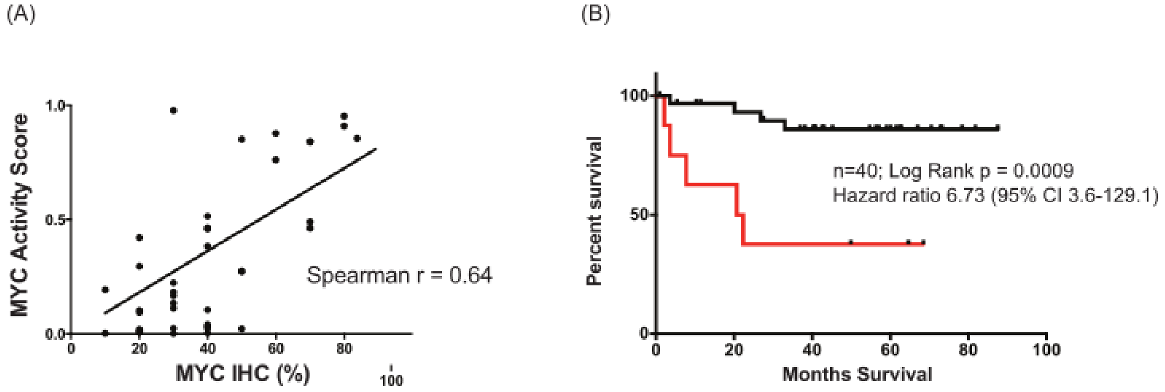


Figure 3.6 Results of the MYC classifier and overall survival (OS) among patients with primary DLBCL treated with R-CHOP-based chemotherapy. (A) The correlation between MYC score and MYC IHC for this outcome series (Spearman r = 0.64). (B) Kaplan-Meier (KM) curve showing Overall Survival (OS) for the outcome series with a MYC score >0.5 (red line) and a MYC score <0.5 (black line).

3.5 Discussion

The WHO currently considers histomorphologic, immunophenotypic and genomic data in order to categorize aggressive B-cell lymphomas (Campo et al. 2011). However, the interpretation of histomorphology and IHC remains subjective and requires expert review. Molecular profiling has the potential to aid diagnostic categorization by providing objective data from normalized gene expression signatures but until recently the degradation of RNA due to formalin compromised the ability to utilize fixed biopsy specimens (Rimsza et al. 2011; Scott et al. 2014; Masqué-Soler et al. 2013; Linton et al.).

We have described a framework for the molecular classification of MYC-driven B-cell lymphomas using targeted expression profiling of RNA isolated from FFPE tissue. The approach described has several features that make it appealing. First, the assay requires only small amounts of FFPE tissue. We and others find that RNA isolated from the equivalent of 2 to 6 X 5-µm FFPE tissue sections is sufficient for analysis (Scott et al. 2014). Second, the assay is robust. We successfully profiled 96 FFPE tumor biopsy samples ranging from 0.5 to 13 years old, with only an additional 5 (5%) failing analytical quality control, and repeat testing of the same samples yielded nearly identical results. Third, the step-wise application of the diagnostic and MYC activity classifiers mimics the diagnostic approach used to evaluate aggressive B-cell lymphomas in clinical practice. Finally, the molecular scores provide quantitative outputs that can be interpreted objectively.

We framed our definition of BL in terms of high MYC and TCF3 transcriptional activity, as these are known major determinants of tumor behavior (Hecht & Aster 2000; Schmitz et al. 2012; Love et al. 2012). DLBCL was defined by variable MYC activity, low TCF3 activity, and high BCL2 and targets of NFKB (Dave et al. 2006). This limited signature was sufficient to categorize >90% of BL and DLBCL in the test set with high confidence and with perfect accuracy (Table 3.1). The results are comparable to those reported in a prior, exploratory study comparing categorization of BL and non-BL using targeted GEP against a ‘gold standard’ global GEP (Masqué-Soler et al. 2013), and validate a molecular, diagnostic classification for cases of well-defined BL and DLBCL.

BCL-U are ‘intermediate’ tumors that share features with BL and DLBCL according to traditional diagnostic evaluation, but 'intermediate' tumors are also identified by molecular analyses (Campo et al. 2011; Dave et al. 2006; Hummel et al. 2006). It is important to note that ‘histomorphologically intermediate’ and ‘molecularly intermediate’ are non-synonymous terms and will categorize mature, aggressive B-cell lymphomas in different ways (Salaverria & Siebert 2011). For example, in our test cohort, 3 BCL-Us classified as mBL. This must be considered inaccurate in the context of WHO classification but is consistent with prior molecular characterization of B-cell lymphomas in which most 'atypical BLs' and a proportion of 'unclassifiable aggressive B-cell lymphomas' classified as mBL (Hummel *et al.*, Figure 2 [2006] (Hummel et al. 2006)). Similarly, small numbers of BL, BCL-U, and DLBCL in our series had diagnostic molecular scores 'intermediate' between mBL and mDLBCL. This result is also consistent with prior analyses in which subsets of atypical BL, 'unclassifiable aggressive B-cell lymphoma', and DLBCL classified as 'molecularly intermediate' (Hummel et al. 2006). These results support the concept that BCL-U is not a discrete diagnostic category, but includes tumors with molecular profiles of mBL, mDLBCL, and intermediate between mBL and mDLBCL.

Non-BL with *MYC*-rearrangement is also a heterogeneous group that includes tumors with the pathological diagnoses of BCL-U and DLBCL by WHO criteria (Campo et al. 2011; Salaverria & Siebert 2011; Aukema et al. 2014; Gebauer et al. 2013; Gebauer et al. 2015). We found DHLs that classified as mBL, 'molecularly intermediate', and mDLBCL. This result also has precedence. A comprehensive GEP analysis of aggressive B-cell lymphomas highlighted groups of DHLs that classified as mBL and *MYC*-rearranged DLBCLs that classified as mDLBCL (Dave *et al.*, Figure 2 [2006]).

Our results were further supported by the examination of the molecular sub-signatures and the histomorphology of the DHLs. We found that a subset of DHLs have a TCF3 signature that is comparable to, or exceeding that of BL. This result was surprising, given recent reports that the TCF3 signature is specific for BL (Schmitz et al. 2012; Love et al. 2012), although a recent study found that *ID3* mutations can occur in DHL (Gebauer et al. 2013). It will be of interest to correlate *TCF3/ ID3* mutation status with molecular diagnosis in future studies.

DHLs that classified as mBL were histomorphologically typical of BL and cases that classified as mDLBCL were histomorphologically typical of DLBCL. Morphological heterogeneity among DHLs is recognized and may have clinical significance (Johnson et al. 2009). It will be important to determine, using larger cohorts, whether the molecular classifier reliably identifies subsets of DHL with distinct histomorphological characteristics, and to relate these data to clinical outcomes.

The prognostic role of MYC in DLBCL is well established, especially in the context of BCL2 expression, and an assessment of MYC activity has been proposed to be an important part of the diagnostic work-up (Johnson et al. 2012; Kluk et al. 2012; Zhou et al. 2014; Green et al. 2012; Horn et al. 2013; Hu et al. 2013; Friedberg 2012). MYC IHC is a single biomarker that serves as a surrogate for MYC activity. The threshold for MYC IHC that separates low from high-risk disease varies between studies from 10-50%, with most suggesting 40% (Valera et al. 2013; Johnson et al. 2012; Kluk et al. 2012; Zhou et al. 2014; Green et al. 2012; Horn et al. 2013; Hu et al. 2013). However IHC is difficult to standardize between centers, even if an automated platform is used (de Jong et al. 2007). An advantage of expression profiling is that the analysis of a large number of gene-transcripts provides redundancy to the assay and captures a transcriptional signature of MYC activity that IHC for MYC alone cannot offer.

The MYC activity classifier was trained using the gene expression profiles of DLBCLs alone, excluding BLs. Its subsequent application to BLs in the training and test sets revealed high MYC activity scores for all cases, which supports the validity of the classifier. We anticipated that the MYC activity scores would show good, but not perfect, correlation with MYC IHC scores, which we observed. This imperfect correlation is likely to reflect the comparison between a single data point for each tumor (MYC protein expression by IHC) and the combination of a more broad set of data as derived from the normalized expression of MYC-target transcripts (MYC activity score). It is also possible that additional MYC targets, not included in our final profiling panel would improve the validity of the MYC activity score. Finally, there are a number of pre-analytical and analytical variables that we must consider when reviewing MYC IHC data, such as time to tissue fixation and intra-observer variability.

Importantly, 5 of the 6 non-BLs with the highest MYC activity scores in the test set had *MYC*-translocations. Yet, we also observed tumors with *MYC*-translocations and intermediate/low scores; indicating variable MYC activity among SHLs and DHLs (Hummel et al. 2006; Aukema et al. 2014).

To evaluate the clinical relevance of these data, we correlated the MYC activity scores to clinical outcome in a small series of R-CHOP-treated patients with primary, *de novo* DLBCL. Segregating tumors into those with high (>0.5) and low (<0.5) MYC activity scores identified patient populations that differed significantly with respect to overall survival (nominal p = 0.0009). The results provide evidence that the MYC activity score, while showing imperfect correlation with IHC and genetics, captures a biological signature of clinical significance. The limited number of primary DLBCLs with documented treatment and outcome required that we include cases from the training and test sets, therefore a more formal validation of the MYC classifier using an independent case series is needed. Ideally, such a study would compare the inter-institutional reproducibility and the prognostic value of the MYC classifier with MYC IHC in a large, multi-institutional cohort.

In summary we have developed a quantitative method for classifying and stratifying aggressive B-cell lymphomas that is applicable to FFPE tissue samples. The molecular classifiers are robust, but likely to improve with further testing and with the inclusion of additional, select gene signatures (Scott et al. 2014). In addition to distinguishing BL from DLBCL, the diagnostic classifier provides unique data regarding the further classification of BCL-Us and DHLs that inform the standard diagnostic methods and warrant further investigation. This platform will allow for the standardized analysis of an expanded cohort of BCL-U and DHL, from which correlations between GEP and traditional pathology, genetics, and somatic mutational analysis can be further examined. The MYC activity classifier captures a key biological and prognostic hallmark of DLBCL and also has the potential to standardize assessment across institutions. The ability of this classifier to predict outcome requires further validation, initially in a large independent cohort where MYC IHC expression is known to be predictive of outcome, and then in the context of a clinical trial.

4 Comprehensive Consensus Clustering classification of diffuse large B-cell lymphoma on Nanostring

4.1 Introduction

Diffuse large B-Cell lymphoma (DLBCL) is the most common type of non-Hodgkin's lymphoma in adults. Even though DLBCL is a very heterogeneous disease both in terms of underlying molecular mechanisms as well as morphological features, virtually all patients are treated with the standard rituximab(R)/CHOP regimen. While this standard chemotherapeutic treatment is able to cure around two thirds of patients, the rest dies from the disease (Friedberg & Fisher 2008), highlighting the need of a further stratification of the patient population into treatable subgroups.

The heterogeneity of DLBCL has been captured in multiple complementary taxonomies based on transcriptional profiling. One well established approach distinguishes DLBCLs based on their transcriptional similarity to normal B-cell subtypes related to cell-of-origin (COO), which makes a distinction between activated B-cells (ABC) and germinal center B-cells (GCB) (Lenz & Staudt 2010) (Basso & Dalla-Favera 2015). This classification was recently translated onto the clinically deployable Nanostring platform using formalin-fixed paraffin-embedded tissue samples (Geiss et al. 2008) (Scott et al. 2014). However, while the COO classification is significantly associated with disease prognosis (with ABC DLBCLs predicted to have significant worse outcome), it has of yet no direct impact on the treatment course.

Another purely transcription-based molecular stratification of DLBCL is the comprehensive consensus clustering (CCC) classification described in (Monti et al. 2005), which identifies three distinct subtypes: B-cell receptor (BCR), Host Response

(HR) and Oxidative Phosphorylation (OxPhos) tumors. The BCR-type DLBCLs have increased expression of proximal components of the B-cell receptor (BCR) pathway and increased reliance upon proximal BCR signaling and survival pathways (Monti et al. 2005; Chen et al. 2013; Chen et al. 2008). The HR-type have a characteristic inflammatory/immune cell infiltrate and include the morphologically defined subset of T-cell/histiocyte-rich B-cell lymphomas (Monti et al. 2005). Finally, the OxPhos-type DLBCL exhibit enhanced mitochondrial energy transduction and selective reliance on fatty acid oxidation (Caro et al. 2012).

In Polo et al. 2007, an ensemble classifier was introduced for the robust prediction of the CCC subtypes from Affymetrix-based DLBCL datasets. The ensemble classifier was further used to predict the CCC phenotye in DBLCL cell-lines, and the predictions were functionally validated *in vitro*. Further validation of the ensemble classifier predictions was carried out in (Caro et al. 2012; Chen et al. 2013), thus providing strong support to the reliability of the approach. However, the ensemble classifier in its current form is limited, since it relies on the Affymetrix platform, which is not reproducible enough for clinical applications and cannot be used to classify single samples. Furthermore, its reliance on a complex combination of multiple classification rules, each using a distinct set of transcripts, makes its predictions difficult to interpret, hence not easily applicable in clinical settings. Finally, its reliance on the Affymetrix platform makes it not applicable to the analysis of FFPE samples, thus further reducing its clinical relevance.

Here, we describe our successful effort at developing a Nanostring-based (Geiss et al. 2008) parsimonious classifier for the accurate and robust CCC classification of patient's expression profiles from frozen tissue samples amenable to the classification of single samples in preclinical and clinical settings. Furthermore, we show a parsimonious two-way classifier that is able to distinguish the host response subtype from the other two classes in formalin-fixed paraffin-embedded (FFPE) tissue.

4.2 **Materials and Methods**

4.2.1 Affymetrix data

Our *Discovery Set I* (Monti et al. 2005) consists of 141 DLBCL tumor samples profiled for gene expression on the Affymetrix U133A/B chip pair. This dataset was used to derive the comprehensive consensus clustering (CCC) labels (Monti et al. 2005). A second DLBCL dataset (Monti et al. 2012) contains 116 DLBCL samples profiled on the Affymetrix U133Plus2.0 chip, with its CCC labels derived based on the ensemble classifier described in (Polo et al. 2007). For 44 of these 116 samples, expression profiles from formalin-fixed material were also available. We refer to the 44-sample dataset as the *validation set*, and the remaining 72-sample dataset as *Discovery Set II*. Two additional datasets profiled on the Affymetrix U133Plus2.0 chip were used for validation: the Lohr dataset an unpublished in-house dataset, consisting of 57 primary DLBCL samples; and the Lenz dataset (E-GEOD-10846), consisting of 414 primary DLBCL samples. The Lenz dataset consists of two distinct cohorts: a 181-sample cohort corresponding to patients treated with the older CHOP (cyclophosphamide, doxorubicin, vincristine, and prednisone) regimen; and a 233-sample cohort corresponding to patients treated with the current Rituximab-CHOP or R-CHOP regimen. Although all samples were collected pre-treatment, initial exploratory analysis showed a considerable batch effect between the CHOP and R-CHOP cohorts; hence we handled them separately throughout this publication.

4.2.2 Selection of markers

We used linear models for microarrays (Smyth 2005) as implemented in the R/Bioconductor package `limma` to identify differentially expressed genes, and gene set enrichment analysis (GSEA) (Subramanian et al. 2005) to look for differentially regulated pathways.

4.2.3 Housekeeping genes

Since the Nanostring platform relies on designated housekeeping genes for cross-sample normalization, we selected a list of genes based on the following criteria evaluated in Discovery Set I: i) minimum variance across samples; ii) even coverage of the range of measured gene expression in the data, by partitioning the expression range into eight tiers, from 4 to 12 (in $\log_2$ space), and by selecting two genes from each tier;

and iii) lack of differential expression with respect to the CCC and COO classifications. The resulting 16 genes are listed in Table III.3.

4.2.4 Nanostring Profiling on Validation Cohort.

For 44 samples from the recently published DLBCL dataset (Monti et al. 2012) with known CCC annotations (ensemble classifier(Polo, et al. 2007) Best10/13) we had frozen and paired formalin-fixed, paraffin-embedded (FFPE) tissue available which we used as our validation cohort. Of note, the 44 samples were excluded from the feature selection process to ensure unbiased classification performance testing. RNA extraction from frozen tissue was performed using Trizol as previously described (Monti et al. 2012). For RNA extraction from FFPE tissue we followed standard protocols using the Qiagen FFPE-RNA extraction kit. The Nanostring assay was performed in the Dana-Faber Cancer Institute Microarray core following standard protocols. Briefly, RNA were assessed for quality and concentration using Agilent Bioanalyzer RNA Nano or Pico chips and a smear analysis was performed using Agilent 2100 Expert software to quantify the percentage of RNA fragments greater than 300nt in each sample. Thereafter, 100ng of RNA with a fragment size of greater than 300nt was profiled using the custom probe set on the Nanostring (Geiss et al. 2008). The custom probe set was composed of 275 probes ordered directly from Nanostring (55 BCR, 104 OxPhos, 80 HR, 16 housekeeping, 20 COO genes). Capture and Reporter Code sets were added to the samples following manufacturer’s protocol and allowed to hybridize at 65°C for 16 hrs. Samples were washed and loaded onto a cartridge using the nCounter Analysis System Prep Station per manufacturer's recommendations. The cartridge was scanned using the nCounter Digital Analyzer at the maximum resolution of 1150 FOV.

4.2.5 Data preprocessing

All Affymetrix microarray data were normalized based on the *Robust Multi-Array Average* (RMA) procedure (Irizarry 2003) implemented in the R/Bioconductor package `affy`. Probes' annotation by Ensembl gene identifiers was based on custom Brainarray CDFs version 18 (Dai et al. 2005). The Nanostring data was normalized using the R package `NanoStringNorm` (Waggott et al. 2012). Mapping from Ensembl gene identifiers to Gene Symbols was performed using the R/Bioconductor package `biomaRt` (Kasprzyk 2011).

To minimize potential batch effects among different datasets, we performed gene-specific normalization, whereby the expression level y_{ij} of gene *i* in sample *j* in the test dataset is transformed as follows:

$$y_{ij} = \frac{y_{ij} - \overline{y_\imath}}{\sigma_{yi}} \sigma_{xi} + \overline{x_\imath}$$

where $\overline{x_\imath}$ and $\overline{y_\imath}$ are gene i's means within the training and test dataset, respectively, and σ_{xi} and σ_{yi} are the corresponding standard deviations. This transformation is based on the assumption that samples in both datasets are drawn from the same population and corrects for systematic measuring biases.

4.2.6 Classification Models

Most prediction models were inferred based on the Elastic net algorithm (Hui Zou 2005) as implemented in the R package `glmnet`. We selected Elastic net because of its superior predictive performance within the Discovery Sets and because of its interpretability, since the resulting classifier outputs gene-specific coefficients that can be directly mapped to the genes' importance in driving the classification. For comparison purposes, we also tested Random forest (Breiman 2001) and Shrunken Centroid (Hastie et al. 2011) classifiers as implemented in the R packages `randomForest` and `pamr`, respectively. For the assessment of each classifier's prediction performance, we measured accuracy, sensitivity and specificity for the three way classification models. At this stage it is not clear whether sensitivity or specificity are clinically more relevant, thus all the models were optimized for overall training set accuracy. For the two class prediction models we also report the area under the *receiver operating characteristic* (ROC) curve (AUC) as implemented in the R/Bioconductor package `ROC`. The prediction performance *within* a dataset was assessed by 10-fold cross-validation (10-CV) in the larger Affymetrix datasets, and leave-one-out cross-validation (LOO-CV) in the Nanostring datasets.

4.3 **Results**

An overview of the experimental design is presented in Figure 4.1. First, a *parsimonious classifier* based on a carefully selected set of CCC genes (CCC signature) was inferred. To this end, we used cross-validation within Discovery Set I to compare competing classification methods; we then applied the best performing classifier to multiple publicly available Affymetrix DLBCL datasets and compared its predictions to those of the original ensemble classifier (Polo et al. 2007). The parsimonious classifier was then validated on the fresh frozen validation dataset, and finally on the FFPE validation dataset, both profiled on the Nanostring platform.

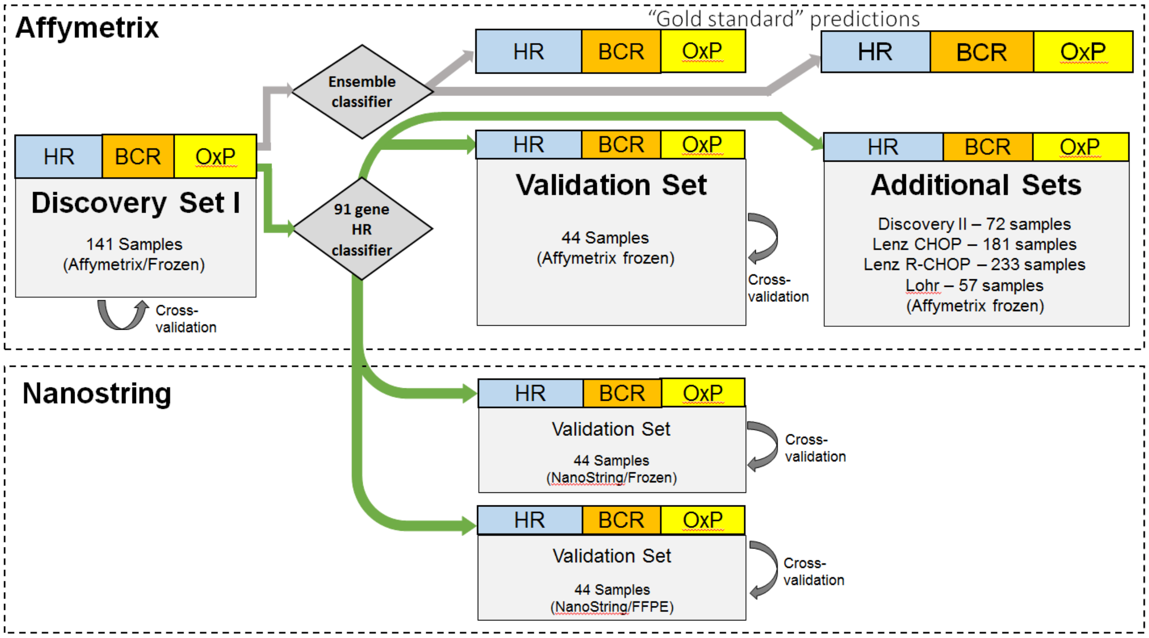


Figure 4.1: Overview of the workflow.

In grey we show the different cohorts used, on top are all the Affymetrix based datasets, on the bottom are the Nanostring data. As gold standard labels we used the ensemble classifier CCC prediction described in (Polo et al. 2007; Monti et al. 2012). Based on the discovery set I we trained an multinominal elastic net classification model that is able to predict CCC classes based on 142 genes. This parsimonious classifier was applied to all datasets and the results were compared to the ensemble classifier results. Of note is that the three validation sets (one Affymetrix and both Nanostring sets) are based on the same 44 tumor biopsy sample, which have been

processed and assayed in a different manner; hence the gold standard labels were derived on the Affymetrix set and then used in the Nanostring sets.

4.3.1 CCC signature selection

An initial set of candidate genes was identified based on their significant enrichment in KEGG, Biocarta and Reactome pathways as tested by GSEA with respect to the CCC phenotype. We used the union of the *leading edge* genes of the top 20 gene sets in each class. We then filtered this initial list based on signal robustness and significance within Discovery Sets I and II, by selecting only genes with fold-change higher than 2.5, false discovery rate (FDR) less than 0.05, and average microarray intensity value greater than 64 (2^6). The significance of the differential expression was assessed by moderated t-test as implemented in limma (Smyth 2005). Finally, we built an Elastic Net model (Hui Zou 2005) from Discovery Set I, and used the estimated genes' coefficients to further prune the candidate list, since the Elastic net-based estimation shrinks to zero the weights of those genes that do not contribute to the classification. The final signature consists of 141 genes (Table III.1), which we used with all classification models in all subsequent evaluations.

4.3.2 Selection of classification model

We ran 10-fold cross-validation on both discovery sets and compared three classifiers: Elastic Net, Random Forest and Shrunken Centroids. The results are summarized in Table III.5. In both datasets, Elastic Net outperformed the Random Forest model, and in one set the Shrunken Centroids (with accuracies of 96.5 vs. 92.2 and 97.1% in Discovery Set I, and 91.8 vs. 91.8% and 90.4% in Discovery Set II). Based on these results, Elastic Net was our classifier of choice, to be evaluated on our validation datasets.

The optimal parameters for classification were selected by maximization of accuracy as estimated by 10-fold cross-validation. Both the parameter alpha, which determines the trade-off between LASSO (Tibshirani 1994) and Tikhonov regularization, and the shrinkage parameter lambda were set to 0.1. A final parsimonious Elastic Net model was then trained based on the entire Discovery Set I. The weights of the signature genes are listed in Table III.4.

4.3.3 Composition of the CCC signature

The elastic net weights provide a data-driven measure of each gene's importance in distinguishing between the three different subtypes. In this section we will describe the most relevant of these genes. The genes that are up-regulated in the BCR subtype include TRMU, CKAP5, PLCG2, FUS, WEE1, ITPR3, SNRPA, PKMYT1, SUPT5H. Those up-regulated in the host response subtype include PD-L1, CTLA4, IL15RA, GNS, PTPRM, AMICA, CFH, CD2, ITGAL, ACTN1, A2M and IL2RB, most of which are related to the adaptive T cell response of the immune system. And finally the genes that have a high weight in the OxPhos subtype include: SPCS3, SUCLG1, NDUFAB1, FADD, MRPS16, ATP6V1D, NDUFB1, NDUFB3, SEC11A, PARK7, many of which are associated with oxidative phosphorylation or the electron transport chain.

4.3.4 CCC predictor in Affymetrix

We applied the parsimonious Elastic net model based on the CCC signature on each of the available Affymetrix datasets, and compared its predictions to those of the Ensemble classifier (Polo et al. 2007). The results, shown in Table 4.1, indicate an accuracies ranging from 81.8 to 93.2%. For comparison, the 10-fold cross-validation on both discovery sets and the validation set (Affymetrix) yielded accuracies between 0.969 and 0.997, while the accuracies were between 81.8 and 96.5%

Table 4.1: CCC prediction results across all datasets.

All predictions in this table are derived by building an elastic net model on the discovery set I, which we then used to predict the class labels across all sets after using gene specific normalization to reduce the batch effect. (ACC: accuracy, SENS: sensitivity, SPEC: specificity)

	Discovery II	Lenz CHOP	Lenz R-CHOP	Lenz Lohr	Validation Affymetrix	Validation Nanostring frozen	Validation Nanostring FFPE
Technology	Affymetrix					Nanostring	
Samples	72	181	233	57	44	44	44
ACC	0.932	0.818	0.906	0.860	0.931	0.886	0.591
SENS – BCR	0.920	0.873	0.932	0.909	0.949	0.929	0.929
SPEC – BCR	0.979	0.833	0.917	0.886	0.961	0.967	0.6
SENS – HR	0.955	0.847	0.897	0.789	0.919	0.938	0.688

SPEC – HR	0.941	0.963	0.952	1	0.962	0.893	0.893
SENS – OxP	0.923	0.600	0.883	0.875	0.925	0.786	0.143
SPEC – OxP	0.979	0.921	0.987	0.902	0.974	0.967	0.9

4.3.5 CCC predictor in Nanostring

With the parsimonious classifier established in the Affymetrix Discovery set, we next tested its performance on Nanostring. Thanks to the availability of paired samples profiled on Affymetrix for each of the patients in the Nanostring validation set, whole-transcriptome CCC predictions based on the ensemble classifier (Polo et al. 2007) were used as the gold standard.

We tested the classification performance on the Nanostring data using our parsimonious model trained on Discovery set I. As shown in Table 4.1, we achieved classification accuracy of 88.6% in the frozen set and 59.1% in the FFPE data. The heatmaps in Figure 4.2 show the actual gene expression profiles in the Nanostring frozen dataset, with the samples ranked by their class probabilities and grouped by subtypes. For comparison, we show the corresponding heatmap for the 44 samples profiled on Affymetrix in Figure III.1 and the ones processed in Nanostring FFPE in Figure III.2 in the Appendix.

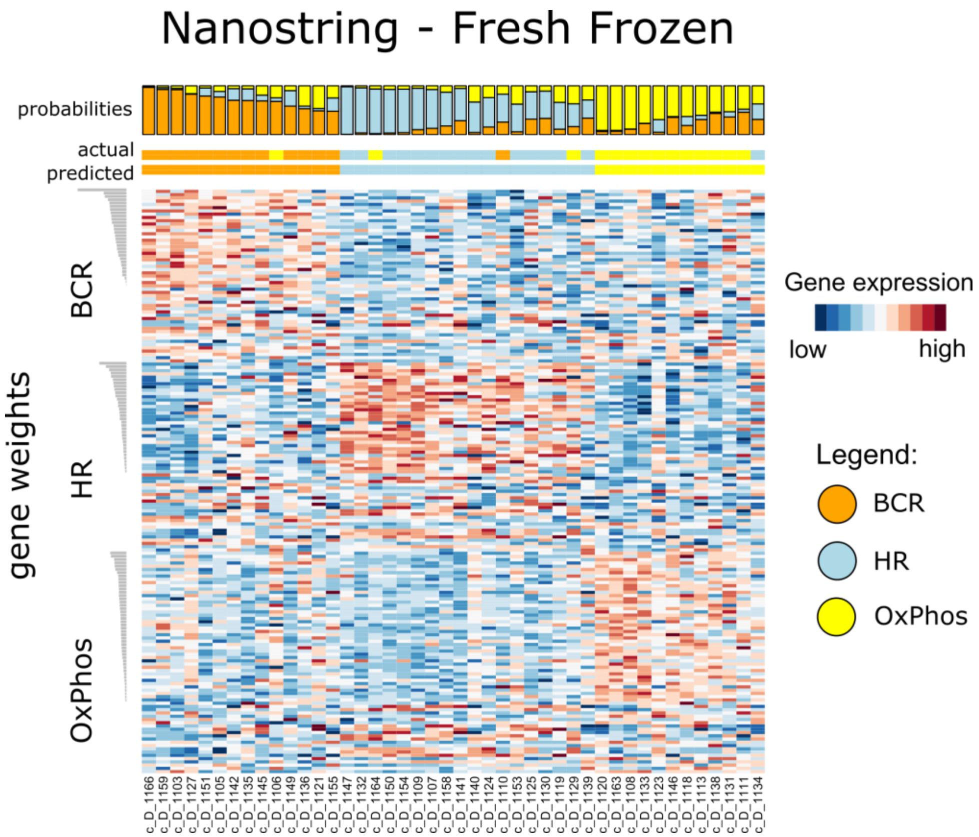


Figure 4.2: CCC Heatmap for Nanostring fresh frozen dataset.

The samples are ordered by class probabilities, based on the predictions of an elastic net model trained on the discovery set I. The top barplot shows the single class probabilities of the classifier, the color-bars below shows the gold standard and predicted CCC subgroups. Each row corresponds to a gene in our CCC signature, which are grouped by class and weights within the elastic net model.

In addition to the classification based on the Affymetrix model, we also used leave-one-out cross-validation (LOOCV) within the Nanostring datasets. In the LOOCV scheme train and test set were both profiled on the same platform, thus potentially eliminating sources of difference between platforms. On the other hand, the sample size was considerably reduced, since each CV fold defined a 43-sample training set, rather than the 141-sample training set available in the Affymetrix platform. Interestingly, elimination of cross-platform differences were not sufficient to compensate for the smaller sample size, and the LOOCV prediction performance was reduced to 81.8% and 59.9% accuracy in the frozen and fixed datasets, respectively (Table III.6). The heatmaps are shown in Figure III.3 and Figure III.4.

We investigated the reason for the poor performance of the FFPE dataset. In Figure 4.3 we show within vs. across histograms, where we look at the correlations of the same gene across platforms (in blue) versus different genes (in red) across platforms. When looking at the Affymetrix frozen samples vs. Nanostring frozen samples there is a very clear separation between the two groups. The vast majority of within correlations are above 0.6, only a few genes did not work. The same is not true when looking at Affymetrix versus Nanostring FFPE. While there is still a separation between the two groups, the overlap is much larger and most correlations are below 0.6.

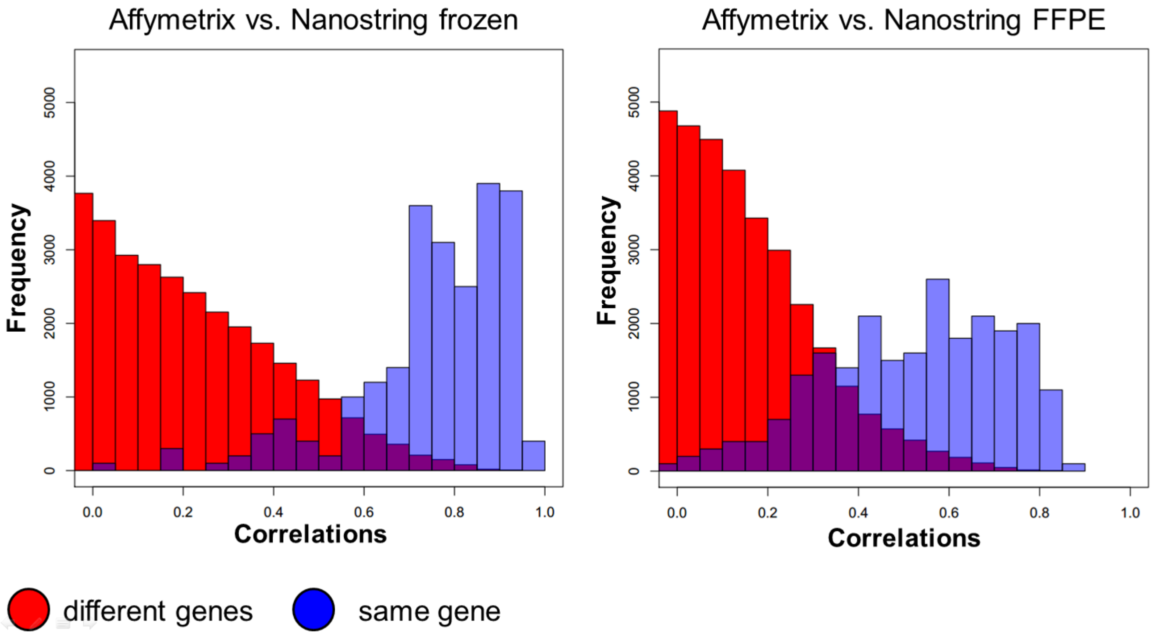


Figure 4.3 Within vs. across correlation between Affymetrix and Nanostring Frozen/FFPE

The two plots show histograms of the across correlations (between different genes on different platforms) in red and within correlations (between the same genes on different platforms) in blue. The across correlations are centered at 0. On the left we show the correlations between Affymetrix and Nanostring frozen, whereas on the right side we show the correlations between Affymetrix and Nanostring FFPE.

4.3.6 Learning Curves for sample size estimation

To determine whether the available sample size was sufficient to achieve maximum prediction accuracy, we carried out down-sampling experiments to estimate learning curves relating classification accuracy to sample size. In particular, starting from a training set consisting of 13 samples, up to the total number of samples (n=44) in increments of 3, properly stratified datasets were randomly sampled 1000 times for each sample size, and accuracy means and standard deviations were estimated based on leave-one-out cross-validation within each of the sampled datasets. The estimated AUCs and their corresponding number of compounds for the frozen set is shown in Figure 4.4, the one for the FFPE set in Figure III.5, together with linear regression lines fitted on the [sample size; accuracy] pairs.

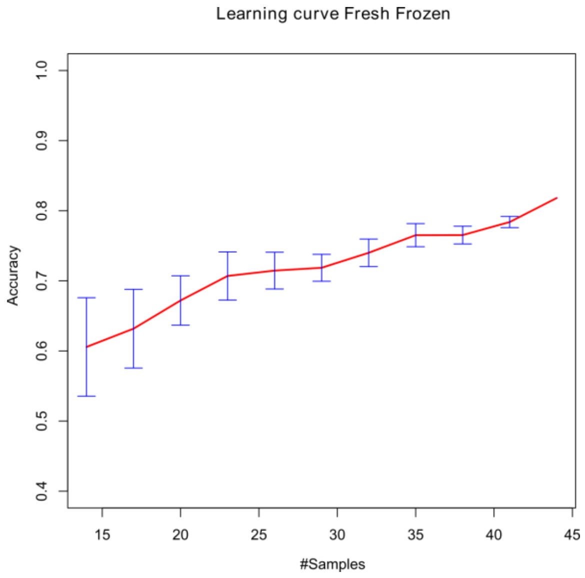


Figure 4.4: CCC Learning curves for the 44 sample Nanostring frozen dataset.

Here we show how well classification works depending on differing sample sizes. For each increment we reran classification 50 times based on random sampled subsets. The red line shows the trend of classification performance, while the blue lines show the 95% confidence intervals based on the 50 reruns. There is a significant upward trend indicating that a larger sample size would result in a better classification performance.

4.3.7 Introducing a host response (HR) classifier

The 3-way classification in NanoString FFPE underperforms in comparison to the classification on the frozen samples, but this is not the case for all classes equally. Table 4.1 shows that the sensitivity for BCR is 0.591 and 0.143 for OxPhos, while it is 0.688 for HR. This can easily be seen in Figure III.2, where most OxPhos samples are classified as BCR samples. As discussed later, FFPE has significant advantages over a freezing with liquid nitrogen, mostly relating to its clinical application. Given that the stratification of the HR samples performs much better than the other two classes we decided to build a two class prediction model that is able to separate host response samples from both OxPhos and BCR samples. Figure 4.5 shows the updated workflow that we used for this two-way classifier.

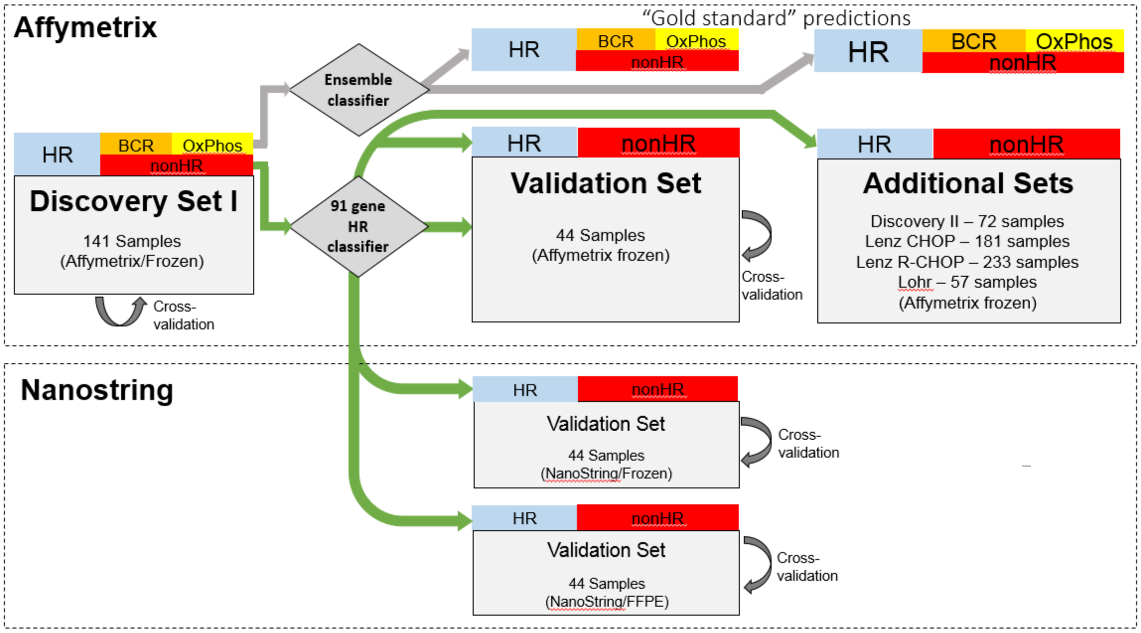


Figure 4.5: Overview of the HR workflow.

In grey we show the different cohorts used, on top are all the Affymetrix based datasets, on the bottom are the Nanostring data. As gold standard labels we used the ensemble classifier CCC prediction described in (Polo et al. 2007; Monti, et al. 2012). Based on the discovery set I we trained an elastic net classification model that stratifies host response (HR) samples from the rest, instead of using three classes. This parsimonious classifier was applied to all datasets and the results were compared to the ensemble classifier results.

4.3.8 HR signature and model selection

We used the same preprocessing as described in 4.3.1 to acquire a raw list of genes. Based on these genes we again trained an Elastic Net model on Discovery Set I, and used the estimated genes' coefficients to further prune the candidate list. The final signature consists of 91 genes (Table III.8), which we used with all HR classification models in all subsequent evaluations.

We also reran a 10-fold cross-validation on both discovery sets and compared our three classifiers; the results are shown in Table III.7. In both datasets, Elastic Net outperformed the other two, with accuracies of 97.9 vs. 95.8 and 93.6% in Discovery Set I, and 90.4 vs. 87.7% and 89.0% in Discovery Set II. Based on these results, Elastic Net was again chosen as the classification model. Both the parameter alpha, and the shrinkage parameter lambda were set to 0.1. A final parsimonious host response model was then trained based on the entire Discovery Set I. The weights of the signature genes are listed in Table III.8.

4.3.9 Composition of the HR signature

The elastic net weights provide a data-driven measure of each gene's importance in distinguishing between the HR subtype and other DLBCL samples. In this section we will describe the most relevant of these genes. The genes that are up-regulated in the HR subtype include PD-L1, CTLA4, IL15RA, GNS, PTPRM, AMICA, CFH, CD2, ITGAL, ACTN1, A2M and IL2RB, most of which are related to the adaptive T cell response of the immune system. The down-regulated part of the HR signature is more heterogeneous, but includes many genes related to the electron transport chain and oxidative phosphorylation: SLC19A1, PKMYT1, CDK1, SHMT2, WEE1, MRPS25/16/9, TRMU, CCNB2, SPCS3, CASP3, UQCR10, RPL35A and MTHFD2.

4.3.10 Host response predictor in Affymetrix and Nanostring

We applied the parsimonious Elastic net model based on the HR signature on each of the available Affymetrix datasets, and compared its predictions to those of the Ensemble classifier (Polo et al. 2007). The results, shown in Table 4.2, indicate an area under the ROC curves (AUC) ranging from 0.967 to 0.989, and accuracies between 90.4 and 96.5%. For comparison, the 10-fold cross-validation on both discovery sets and the validation set (Affymetrix) yielded AUCs between 0.969 and 0.997, while the accuracies were between 90.91 and 97.87%.

As for the Nanostring data, we achieved classification accuracies of 90.9% in the frozen set and of 86.4% in the FFPE data, with AUCs of 0.938 and 0.877, respectively. Figure III.6 shows the ROC curves for both classifications. The heatmaps in Figure 4.6 show the actual gene expression profiles in Nanostring, with the samples ranked by their probability of belonging to the HR group. Of notice, the genes that are up-regulated in the HR class are stronger and more robust across the two sets, which is in concordance with the results we saw on the three-way CCC classifier. As a term of comparison, we show the corresponding heatmap for the 44 samples profiled on Affymetrix in Figure III.7.

In addition to the classification based on the Affymetrix model we also used leave-one-out cross-validation (LOOCV) within the Nanostring datasets. The models in the LOOCV are trained on 43 samples as opposed to the 141 samples that we used to train the Affymetrix model, and as for the three-class problem, the predictive accuracy was reduced to 84.1% and 79.6%, with an AUC of 0.893 for the frozen and 0.828 for the FFPE samples (

Table III.9). The heatmaps and ROC curves are shown in Figure III.8 and Figure III.9.

Table 4.2 – HR prediction results across all datasets.

All predictions in this table are derived by building an elastic net model on the discovery set I, which we then used to predict the class labels across all sets after using gene specific normalization to reduce the batch effect. (ACC: accuracy, SENS: sensitivity, SPEC: specificity, PPV: positive predictive value, NPV: negative predictive value, FDR: false discovery rate, AUC: area under the receiver operating characteristic (ROC) curve)

	Discovery II	Lenz CHOP	Lenz R-CHOP	Lenz Lohr	Validation Affymetrix	Validation Nanostring frozen	Validation Nanostring FFPE
Technology	Affymetrix					Nanostring	
Samples	72	181	233	57	44	44	44
ACC	91.78	90.61	92.70	94.74	90.91	90.91	84.09
SENS	95.45	79.17	94.12	89.47	81.25	87.5	68.75
SPEC	90.20	98.17	92.12	97.37	96.43	92.86	92.86
PPV	80.77	96.61	83.12	94.44	92.86	87.5	84.62
NPV	97.87	87.70	97.44	94.87	90	92.86	83.87
FDR	19.23	3.34	16.88	5.56	7.14	12.5	15.39

AUC	0.987	0.984	0.978	0.991	0.967	0.94	0.877

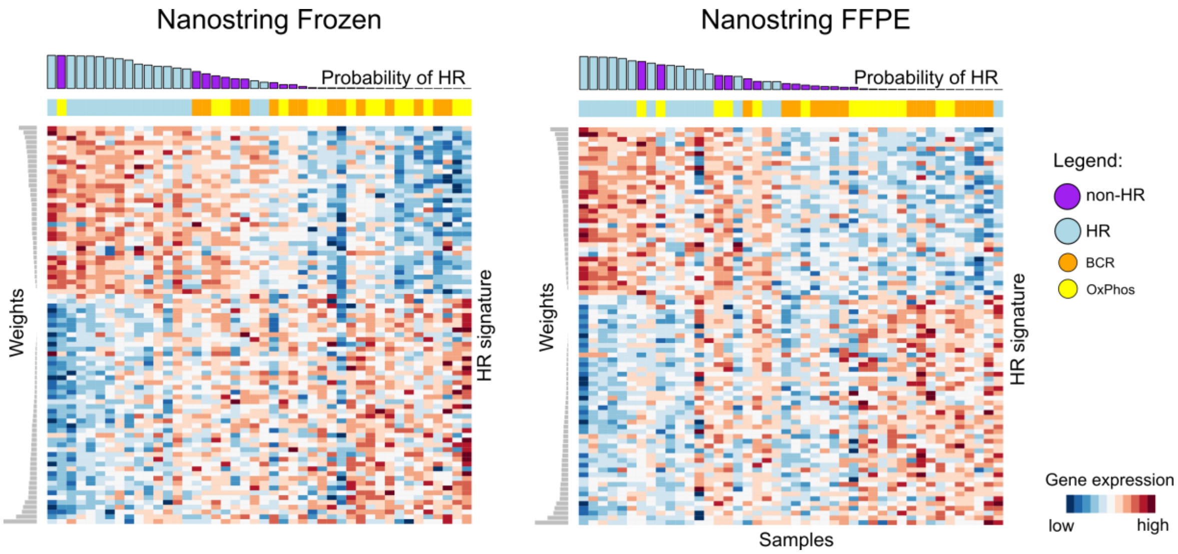


Figure 4.6: Heatmaps for Nanostring frozen and FFPE datasets.

On the left we show a heatmap of the Nanostring fresh frozen samples, which are ordered by the HR class probabilities resulting from an elastic net model trained on the discovery set I. The top barplot shows the probabilities of the classifier, the color-bar below shows the CCC subgroups, which are our gold-standard. Each row corresponds to a gene in the HR signature and the barplot on the left indicates the coefficient weights of the elastic net model. On the right side we show the same heatmap for the FFPE Nanostring set.

4.3.11 Learning Curves for host response classifier

We built learning curves again, this time relating classification AUC to compound's sample size using the same scheme as for the CCC classifier, with the one difference that we use the more robust AUCs instead of the accuracies. The estimated

AUCs and their corresponding number of samples for both frozen and FFPE sets are shown in Figure III.10 together with linear regression lines fitted on the [sample size; AUC] pairs.

4.4 Discussion

The CCC and HR biomarker genes were selected based on a purely data-driven approach. This resulted in the inclusion of several well-known genes in their respective subtypes. For the HR subtype, genes that reduce the activity of the specific immune system based on T-cell were included. Namely PD-L1 that was found to be relevant in a host of recent studies, e.g. (Green et al. 2010; Zitvogel & Kroemer 2012; Herbst et al. 2014) and CTLA4, which restrains the adaptive immune response of T cells towards tumor associated antigens 27. For the BCR subtype the signature most notably includes SYK, a well-known gene that helps to promote survival in hematopoetic malignancies (Friedberg et al. 2010), while the OxPhos portion includes several genes related to oxidative phosphorylation and the electron transport chain such as SUCLG1, NDUFAB1, ATP6V1D, MRPS16.

From the learning curves in Figure 4.4 and Figure III.10 we can conclude that 44 samples are not sufficient to achieve maximum prediction performance. All three curves show an upward trend with no ‘plateauing’, suggesting that an increased sample size would indeed lead to increased classification accuracy. This also explains the considerable difference in accuracy between the predictions based on the models trained on the 141-sample discovery set and those based on LOOCV within the validation set, where each model is trained only on 43 samples.

Interestingly, Figure 2 shows that the host response portion of the CCC signature appears to be more robust to the translation from Affytmetrix to Nanostring, which is even more pronounced when going across tissue preservation technologies. The overall CCC accuracy within the FFPE set drops to 59.1%, while the prediction of the HR samples has a sensitivity of 68.8% and a specificity of 89.3%.

Based on these findings, we showed that a simple host response biomarker was very robust across both platforms and preservation technologies; we showed that predictions based on a model trained on Affymetrix and tested on Nanostring, as well as across preservation methods (training on frozen, testing on fixed samples), still resulted in an AUC of 0.877. Furthermore, with an adequate sample size, we expect that a model trained on FFPE samples profiled on Nanostring would increase prediction performance even further.

4.5 Conclusions

We were successfully able to build parsimonious biomarkers for both CCC and HR. However, this study also showed the difficulty of validating molecular subtypes purely based on their gene expression profiles. A fact that is often glossed over is that absolute gene expression value for a sample is heavily dependent on many external factors, such as amount of RNA, preservation technology and assaying technology. This means that a classifier that is trained on a particular dataset is also dependent on all the external factors that are used to create that dataset and as a consequence, predictions based on gene expression profiles of the same biological samples using different technologies can lead to different results. We showed an example of this when switching from fresh frozen tissue on Nanostring to FFPE, which works for the HR classifier, but not for the more challenging three-class CCC classifier. When external gold standards are available they can be used to show inconsistencies, however in our case this was not as straightforward. Careful study design (i.e. the same samples profiled using both preservation technologies as well as Affymetrix and Nanostring) was necessary to translate our findings from fresh frozen samples in Affymetrix to frozen and FFPE in Nanostring.

Going forward, we plan to use patient derived xenograft (PDX) models – primary tumors grafted onto immunocompromised mouse models – to functionally validate our *in silico* predictions. To this end, we will profile the PDX samples on the Nanostring platform and treat them with the selective inhibitors specific to each class (e.g., with SYK inhibitors to test BCR class membership) (Chen et al. 2013; Chen, et al. 2008). This approach will allow for an independent functional validation of our predictions.

5 IDENTIFYING TUMORS DEPENDENT ON OXIDATIVE PHOSPHORYLATION ACROSS DIFFERENT CANCER TYPES

I co-developed ASSIGN, a method used in this chapter:

Shen Y, Rahman M, Piccolo SR, Gusenleitner D, EI-Chaar NN, Cheng L, Monti S, Bild AH and Johnson WE, 2015. ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. *Bioinformatics (Oxford, England)*, 31(11), pp.1745–53.

5.1 **Introduction**

In the presence of oxygen, non-proliferating tissues rely primarily on oxidative phosphorylation (OxPhos) to produce ATP. In this process pyruvate, which is metabolized from glucose during glycolysis, is completely oxidized into carbon di-oxide. In anaerobic condition, i.e. in the absence of oxygen, oxidative phosphorylation is not possible and lactate is generated instead after glycolysis. Fast proliferating cells, on the other hand meet most of their energy demands primarily through glycolysis even in the presence of sufficient oxygen and most glucose is converted into lactate. This aerobic glycolysis is well established and termed the Warburg effect (Warburg 1956). In comparison to oxidative phosphorylation the ATP output of glycolysis is minimal, however, recent studies suggest that the excess lactate can be used as chemical ‘‘building blocks’’ required for the anabolic processes that must occur prior to cell division (Vander Heiden et al. 2009). Cancers cells are typically glycolytic since one of the hallmarks of cancer is an increased proliferation, however, they usually retain a degree of mitochondrial respiration and derive a significant fraction of their ATP from oxidative phosphorylation (Ward & Thompson 2012).

Besides efficient ATP production, mitochondrial oxidative phosphorylation is also a major cellular source of reactive oxygen species (ROS) and enhanced and unbalanced metabolic activity can lead to an increase in ROS (Hanahan & Weinberg 2011). This increase can be both a gift and a curse for a cancer cell. It can confer advantages in proliferation (Gatenby & Gillies 2004; Locasale & Cantley 2011) and can also promote genomic instability through oxidative damage (Weinberg & Chandel 2009), but high levels of ROS are toxic to cells, which can lead to damage and eventual cell death (Trachootham et al. 2009; Diehn et al. 2009). Thus cancer cells are usually adaptive in their response to oxidative stress.

As described in Chapter 4, we identified three molecular subtypes in diffuse large B-Cell lymphoma, one of which heavily relies on oxidative phosphorylation (OxPhos) and co-regulated mechanisms that protect tumors from ROS (Monti et al. 2005). A follow-up study (Caro et al. 2012) expanded on these findings and showed OxPhos-dependent DLBLC cell-lines are targetable by PPARG inhibitors, suggesting a therapeutic relevance. Furthermore, DLBCL was not the only cancer type where OxPhos dependency has been shown. Vazquez et al. 2013 demonstrated that high expression levels of PGC1a, a key regulator of mitochondrial respiration, metabolically define melanomas with high levels of ROS detoxification capacities. This subtype of high PGC1a expressing tumors shows a higher survival rate under increased levels of oxidative stress. Given that this dependence on metabolic pathways to maintain survival rates was found in at least two tissue types we hypothesize that this might be a broader mechanism of cancer biology that might be relevant also in other tissue types.

Here we show our attempt to test this hypothesis by investigating four additional tissue types. An overview of the analyses done in this chapter is provided in Figure 5.1.

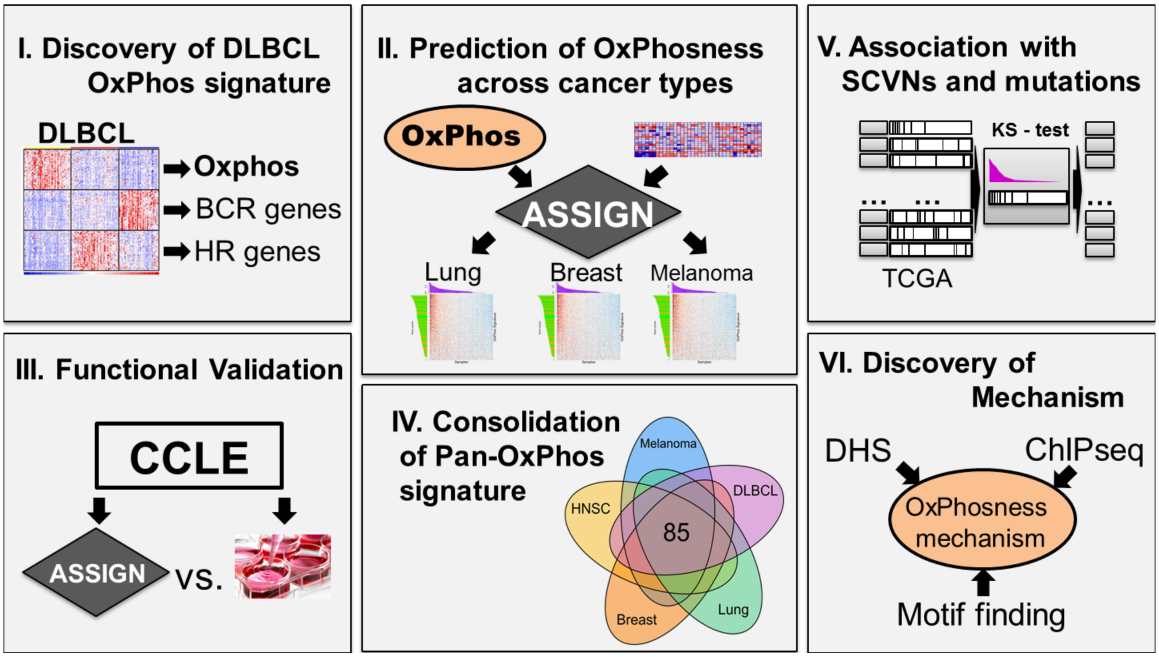

I. Discovery of DLBCL OxPhos signature
DLBCL
Oxphos
BCR genes
HR genes
II. Prediction of OxPhosness across cancer types
OxPhos
ASSIGN
Lung
Breast
Melanoma
V. Association with SCVNs and mutations
KS - test
TCGA
III. Functional Validation
CCLE
ASSIGN
vs.
IV. Consolidation of Pan-OxPhos signature
Melanoma
DLBCL
HNSC
85
Breast
Lung
VI. Discovery of Mechanism
DHS
ChIPseq
OxPhosness mechanism
Motif finding

Figure 5.1: Analysis workflow

I. Derivation of a lymphoma centered OxPhos signature based on the DLBCL dataset that was used to derive the CCC subtypes. II. Use of ASSIGN to look for differences in OxPhos activity in a variety of different tumor types. III. Functional validation on cell-lines, assessing OxPhos activity and OxPhos dependency. IV. Determination of the overlap among tissue specific OxPhos signatures to derive a PanOxPhos signature. V. Search for associations between OxPhos activity and somatic mutations or copy number alterations. VI. Investigation of the potential mechanism of actions driving the differences in OxPhos activity.

5.2 Materials

For this study we used only publicly available datasets. As shown in Table 5.1, we focused on gene expression profiles (GEP) derived both from fresh frozen primary tissue and cancer cell-lines. Four of the eight primary tumor datasets are part of The Cancer Genome Atlas (TCGA): BRCA (TCGA 2012b), LUAD (Collisson et al. 2014), LUSC (TCGA 2012a) and HNSC (Lawrence et al. 2015). For these four sets we downloaded the FPKM normalized level 3 RNASeq data from https://tcga-data.nci.nih.gov/tcga/. We also acquired the matching somatic copy-number alteration (SCNA) data, which were called using GISTIC 2.0 (Mermel et al. 2011) and mutation (MUT) data which were called using MutSigCV (Lawrence et al. 2013). Both MUT and SCNA data were downloaded from the Firehose pipeline (Marx 2013) at the Broad Institute (http://gdac.broadinstitute.org/).

In addition to the TCGA sets we downloaded the Curtis breast cancer dataset (Curtis et al. 2012), another FPKM normalized RNASeq dataset, and three Affymetrix datasets: Lymphoma 2003 (Monti et al. 2005), which was profiled on both Affymetrix U133A and B, Lymphoma 2010 (Monti et al. 2012), profiled on Affymetrix U133Plus2.0 and the Melanoma Riker set (Riker et al. 2008), profiled on Affymetrix U133A.

Furthermore, we used the breast, lung and melanoma cell-lines from the cancer cell-line encyclopedia (CCLE) (Barretina et al. 2012) assayed on Affymetrix U133Plus2.0, and an in-house 22-sample DLBCL dataset (Polo et al. 2007), which was also profiled on the Affymetrix U133A/B pair.

All Affymetrix dataset were normalized using frozen Robust Multiarray Analysis (fRMA) (McCall et al. 2010). Probes' annotation by Ensembl gene identifiers was based on custom Brainarray CDFs version 18 (Dai et al. 2005).

Table 5.1: Cancer datasets

On the top we list all primary tumor datasets we used for the study, on the bottom all cancer cell-line sets. We show the sample numbers for gene expression profiles and indicate the availability (X) of somatic copy number alteration data (SCNA) and mutation data (MUT).

Fresh frozen	Samples	SCNA	MUT	Platform
TCGA Breast	977	X	X	Illumina RNASeq
TCGA Head and neck	303	X	X	Illumina RNASeq
TCGA Lung adenocarcinoma	455	X	X	Illumina RNASeq
TCGA Lung squamous	408	X	X	Illumina RNASeq
Melanoma Riker	86			Affymetrix U133A
Lymphoma 2003	176			Affymetrix U133A/B
Lymphoma 2010	116			Affymetrix U133Plus2.0
Breast Curtis	1981	X		Illumina RNASeq
Cell-lines				
CCLE breast	58			Affymetrix U133Plus2.0
CCLE lung	179			Affymetrix U133Plus2.0
CCLE melanoma	61			Affymetrix U133Plus2.0
DFCI - DLBCL	22			Affymetrix U133A/B

5.3 **METHODS**

5.3.1 Stratification using ASSIGN

We applied *Context-specific Genomic Profiling of Multiple Heterogeneous Biological Pathways* – ASSIGN (Shen et al. 2015) to estimate relative OxPhos activity within all datasets. ASSIGN utilizes a flexible Bayesian factor analysis approach similar to clustering, which adapts to differences in backgrounds among samples while scoring the OxPhos activity in each sample.

ASSIGN assigns weights to each gene indicating their importance for the phenotype stratification. This allowed us to rank not only the samples by their respective signature activity, but also the importance of the involved genes. Unlike standard supervised classification or regression methods, ASSIGN can be applied to a dataset without prior knowledge of the phenotype, i.e. it discovers the dimension of the largest variance within a given gene signature space and ranks samples accordingly, while determining gene importance. This has the significant advantage that a gene signature that was derived with a certain background, e.g. tissue type or assaying technology, can be adapted to other backgrounds, overcoming tissue specific effects and differences in assaying technology. However, the underlying assumption of the method is that the main source of variation in the subspace defined by the studied signature is associated to the phenotype of interest. If this assumption is violated (e.g., if the main source of variation in the signature-projected dataset were to be a batch effect), than the algorithm will likely fail. A related potential downside to ASSIGN is that it will always force a separation of samples, regardless of whether or not this stratification is biologically meaningful.

5.3.2 Establishing significance for ASSIGN

Figure 5.2 shows an example of two signatures in the TCGA LUAD dataset. On the left we show the set in the OxPhos signature space (Table IV.2), described in the next section. On the right we show the same dataset in BCR signature space (Table IV.1), as described in Chapter 4. For both signatures we used only genes that are up-regulated in the respective subtype. The two example signatures were chosen, because BCR signaling is not expected to show a significant activity in breast cancer, while differences in OxPhos activity are.

Both heatmaps show a stratification within the datasets, indicating some samples with high signature activity scores (purple bar plots on top) and some with no activity. However, while the set in OxPhos space shows a very clear up-regulation in the OxPhos samples, the same is not true when looking at the BCR signature space. Furthermore, unlike the left heatmap, the right heatmap contains not only up-regulated genes (green bar plots on the left), but also a significant portion of down-regulated genes (in grey). While it is possible for a signature to have both up and down-regulated genes, the BCR signature contains only genes that are up-regulated in the DLBCL BCR subtype.

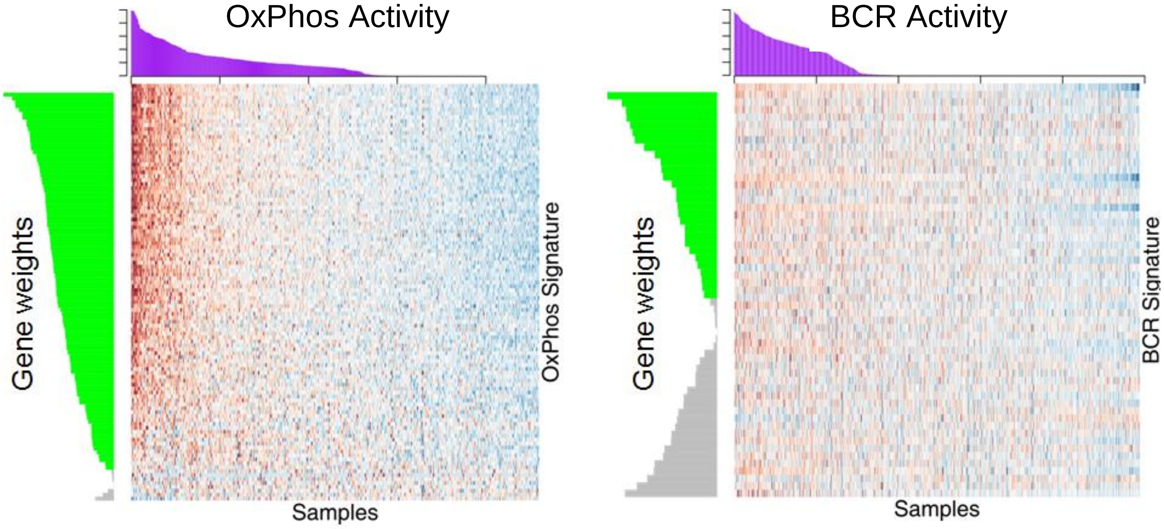


Figure 5.2 Comparison of OxPhos and BCR signatures in breast cancer

Both heatmaps show the gene expression of the TCGA LUAD samples; in OxPhos signature space on the left and BCR signature space on the right side. Up-regulated genes are in red, down-regulated in blue. The samples are ordered according to activity scores (in purple on top) derived from ASSIGN, the genes are ordered by gene weights derived from the ASSIGN models. The gene weights are colored by significance; in green we show significant genes with positive weights.

We tested the difference between actual stratifications and random noise more systematically, by comparing the distribution of the gene weights resulting from 1000 ASSIGN reruns based on random signatures with the absolute scaled gene weights of the OxPhos and BCR signatures. The left panel of Figure 5.3 shows that the gene weights from the OxPhos signature can be easily distinguished from the random weights, whereas the BCR gene signature weights on the right are indistinguishable from the random gene signature weights. For active signatures or pathways many genes contribute to the stratification, as indicated by a large number of genes with a higher weights, while the weights of randomly picked genes follow an exponential distribution, where the stratification is driven by very few genes while most others have a minor contribution. The weights of active pathways do not follow an exponential distribution, which can be exploited to test for an active pathway. Hence, for the rest of this study we used a Kolmogorov-Smirnov (KS) test to determine whether the absolute scaled gene weights of a signature follow an exponential distribution, in order to get a significance of stratification.

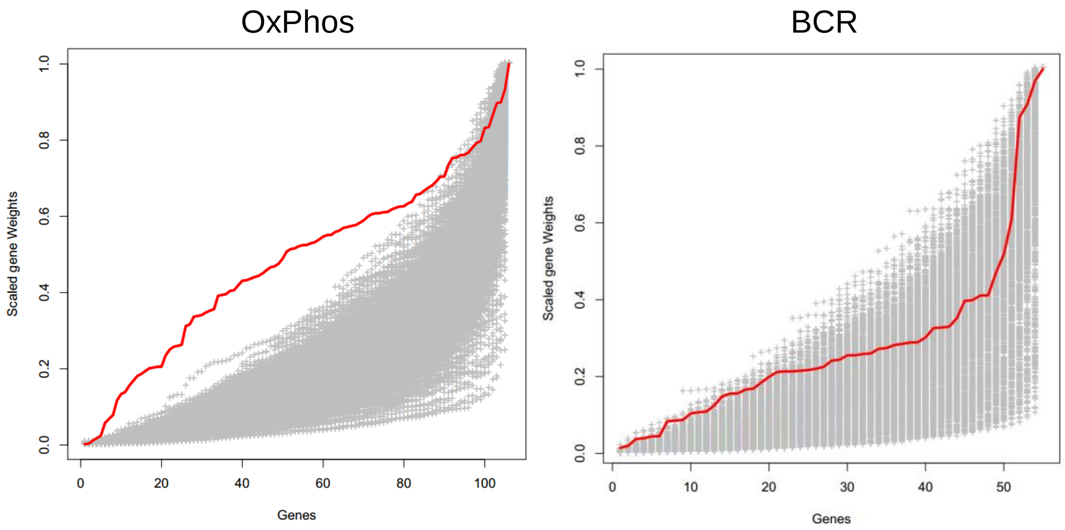


Figure 5.3: Gene weights of the actual gene signature vs. random sets

In these plots we show the absolute scaled (to 1.0) gene weights as derived by ASSIGN on TCGA LUAD versus gene signatures ordered by their weights. In grey we show the weights for 1000 randomly chosen gene sets, whereas in red we show the weights for the DLBCL OxPhos gene signature on the left and the DLBCL BCR signature on the right.

5.3.3 Testing for associations with mutations and SCNAs

With the Oxphos activity established in our datasets, we were interested in potential drivers for the OxPhos phenotype. Somatic alterations are among the potential candidates, so we looked for associations with mutation data from MutSigCV (Lawrence et al. 2013) and somatic copy number alteration (SCNA) from GISTIC 2.0 (Mermel et al. 2011). To that end, we assembled mutational and SCNA profiles across all TCGA datasets, only including calls with a false discovery rate <0.25, and we used a two sample Kolmogorov-Smirnov test to test for association between signature activity score and somatic alterations. Figure 5.4 shows the workflow of this analysis.

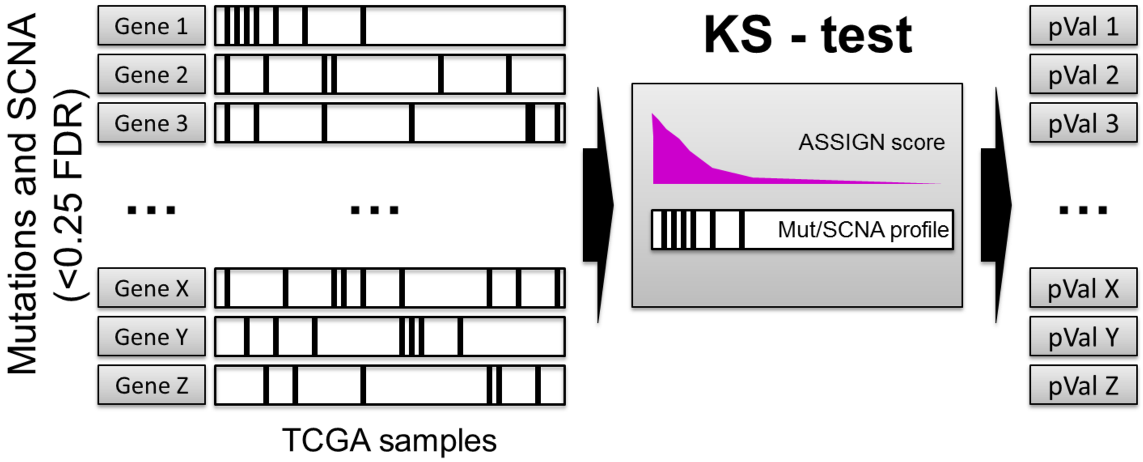


Figure 5.4: Workflow of the association between OxPhos activity and mutations

First we derive only significantly recurrent (FDR<0.25) mutations or SCNA for each sample in a given dataset. We then assemble a mutational/SCNA profile for each gene in a given dataset and use a Kolmogorov-Smirnov test to look for associations with gene signature activity, which allows us to derive a p-value for each gene.

5.4 **Results**

5.4.1 Deriving an OxPhos signature from DLBCL

In Chapter 4, I described the development of a parsimonious classifier for the CCC (Monti et al. 2005) classes in DLBCL. As part of this parsimonious classifier we derived streamlined signatures for each of the three subtypes. Table IV.2 includes the 108 genes of the CCC-OxPhos signature, which we used as a starting point for the rest of this chapter.

***5.4.2 Assessing OxPhos activity using ASSIGN* across tissue types**

Next we set to systematically determine whether we could find the same subtype in more than these two tissue types using ASSIGN (Shen et al. 2015). Figure 5.5 shows a comparison between the two methods on the Lymphoma 2010 gene expression dataset. On the left we show a simple hierarchical clustering in OxPhos signature space. It is evident, that the samples with an up-regulation, correspond to the OxPhos CCC type, however, due to the nature of hierarchical clustering there is not ranking in terms of “oxphos-ness” as well as gene importance. The panel on the right shows the same data, ordered by the ASSIGN output and gives a much clearer indication on how active OxPhos is in each sample and as a bonus it gives weight to the importance of each gene.

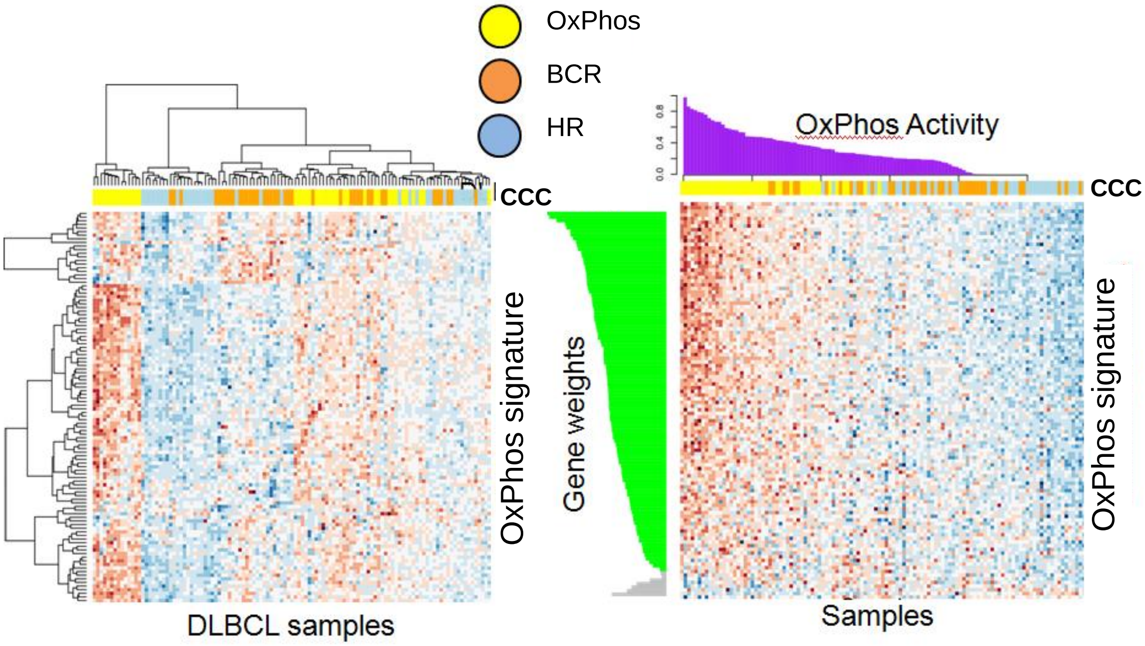


Figure 5.5: Heatmaps of the lymphoma 2003 dataset in OxPhos gene signature space.

Both heatmaps show the 116 samples from the Lymphoma 2010 dataset in the 108 OxPhos gene signature space. Red indicates up-regulated, blue indicates down-regulated genes. The color bars on top show the comprehensive consensus clustering (CCC) classification for each sample. On the left we show the samples clustered by hierarchical clustering. On the right the samples are ordered by OxPhos activity and genes are ordered by gene weights derived from an ASSIGN model.

Next, we used ASSIGN on each of our cancer dataset and calculated p-values for each set testing whether the observed stratification is significantly different from randomly drawn gene set (as described in the methods section). The results are reported in Table 5.2. All fresh frozen datasets showed significant results, the only not significant results was observed in a DLBCL cell-line dataset, which we suspect might be due to the small sample size.

Table 5.2: Significance of OxPhos activity across different cancer datasets

Here we show the p-values indicating the significance of sample stratification (see methods section) within OxPhos signature space in comparison to random signatures across all primary cancer and cancer cell-line dataset.

Dataset	p-value
DLBCL 2003	9.77E-09
DLBCL 2010	6.36E-08
TCGA Lung adenocarcinoma	3.28E-06
CCLE lung	5.53E-06
TCGA Lung squamous carcinoma	8.84E-06
TCGA Head and neck squamous carcinoma	1.30E-05
Breast Curtis	0.000111
CCLE breast	0.00013
CCLE melanoma	0.00088
Melanoma Riker	0.000881
TCGA Breast cancer	0.00162
DFCI -DLBCL	0.15789

5.4.3 Generalizing the OxPhos signature

One of the major advantages of ASSIGN is that it adapts the weights of a gene signature to a context, i.e. when we use our signature derived from DLBCL in another tissue type it will down-weight genes that are specific to lymphoma. In a similar manner, ASSIGN can also be used to add genes that might be relevant to our signature, but were not included in the context the signature was originally derived from. For that we assembled a generalized oxphos gene set from 6 gene sets of MSigDB 3.0 (Liberzon et al. 2011): oxidative phosphorylation, respiratory electron transport, TCA cycle, mitochondria pathway, electron transport chain and oxidoreductase activity acting on

NADH or NADPH. The union of these 6 sets results in a total of 232 genes. There is an overlap with the 108 gene DLBCL OxPhos signature, leading to a total of 287 gene superset. ASSIGN allows for the specification of prior probabilities of genes being significant through the parameter θ_1. By default this parameter is set to 0.9 for genes in a signature, which we still used for the 108 genes that were included in the DLBCL OxPhos signature. For the other 179 genes in the generalized signature we used a prior probability of 0.1, giving them the chance to be included in the resulting list of genes, but making sure that the stratification is still driven by the original DLBCL OxPhos gene signature. We ran this modified version of our ASSIGN analysis on every dataset; graphical representations of most of the primary tumor sets can be found in Figure 5.6 (TCGA LUAD) and Figure IV.1 (TCGA BRCA, LUSC, HNSC and Melanoma Riker).

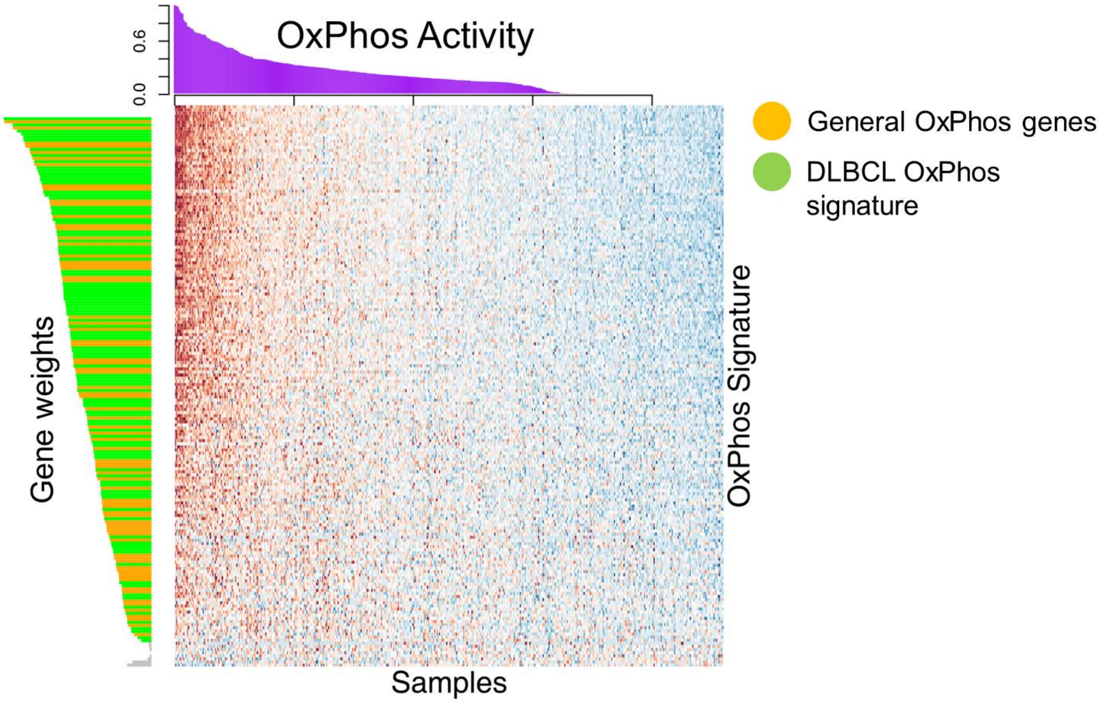


Figure 5.6: Heatmap of generalized OxPhos signature in breast cancer

Here we show the 455 samples in the TCGA LUAD dataset. Red indicates up-regulated genes, whereas blue indicates down-regulated ones. The samples are ordered by OxPhos activity scores derived from an ASSIGN model, which is also shown on top in purple. On the left we show the gene weights from the ASSIGN model, in green the genes from the original DLBCL OxPhos signature, in orange the genes that were added from the generalized OxPhos gene signature and in grey the genes that have either a negative weight or are insignificant.

5.4.4 DLBCL OxPhos signature vs. PGC1α subgroup in melanoma

Our first test of the hypothesis that OxPhos dependency may be relevant in tumor types other than DLBCL was performed in melanoma, based on the results of a recent study that characterized a melanoma subtype exhibiting a strong up-regulation of PPARGC1 (PGC1α) gene expression and dependency on oxygen supply for survival (Vazquez et al. 2013). In particular, the study showed that hypoxia conditions or drugs inhibiting oxidative phosphorylation such as PPARγ inhibitors selectively killed PGC1α high cell-lines.

We started by downloading and processing the 86 primary melanoma samples used by the authors and carried out hierarchical clustering within the space of our 108 DLBCL-OxPhos signature space (see Figure 5.7 on the left side). This heatmap shows 3 distinct clusters, which we color coded: red for the samples that seem up-regulated, light blue for the down-regulated ones and grey for the intermediate cluster. On the right side of Figure 5.7, we then show boxplots for PGC1α expression in each of these three clusters. A t-test comparing the difference between OxPhos high (red) and low (light blue) group resulted in a significant p-value of 0.003711.

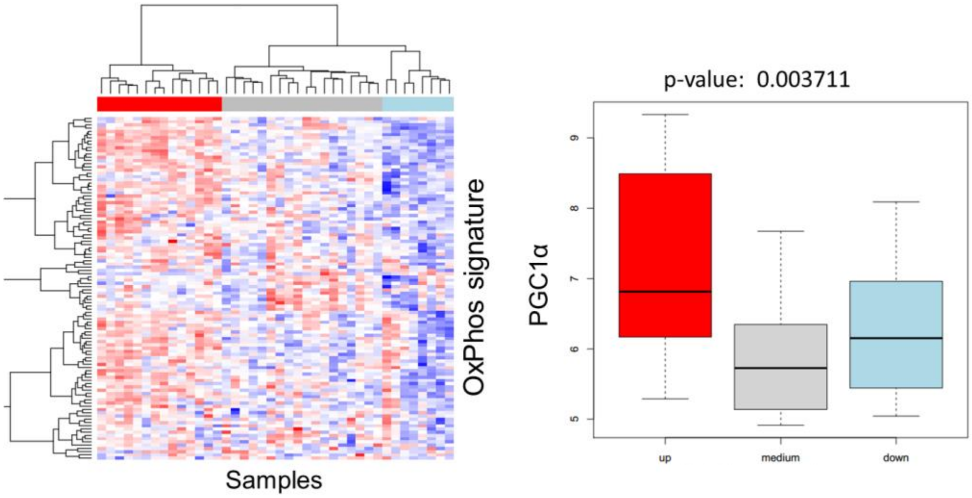


Figure 5.7: Differences in melanoma samples in OxPhos signature space

On the left we show a heatmap of the 86 melanoma Riker samples. Each row corresponds to one of the 108 genes in the OxPhos signature, whereas each column represents one sample. Red indicates up-regulation, blue indicates down-regulation. We used hierarchical clustering to cluster similar samples and colored the three main clusters in the color bar on top. On the right side we show the same three groups of samples, only looking at the expression of PGC1α.

We also preformed gene set enrichment analysis (GSEA) (Subramanian et al. 2005) using the canonical pathways of collection 2 (C2) of the molecular signature database (MSigDB v.3.0) (Liberzon et al. 2011) between the samples in the red cluster and the light blue cluster. Out of the 1256 gene sets there were no significantly down-regulated gene sets. 199 gene sets were significantly (FDR <25%) up-regulated and 40 gene sets had a nominal p-value <1%. See Table IV.3 for the top 25 gene sets; these include oxidative phosphorylation, TCA cycle and respiratory electron transport and glyoxylate and dicarboxylate metabolism, which are all related to our OxPhos signature. This confirmed that the differences in the clusters were not only driven by a potential batch effect, but by differences in metabolism.

5.4.5 Validation in Cell-lines

For the first round of validation we used the cell-line annotations for melanoma cell-lines from Vazquez et al. 2013. In this publication the authors distinguish between a PCG1α positive (OxPhos) and PCG1α negative (non-OxPhos) phenotype, which they functionally validated by measuring the differences in glucose, lactate and ATP levels. All cell-lines tested in the publication are also in the cancer cell-line encyclopedia (CCLE) (Barretina et al. 2012), for which gene expression data are publicly available. We used ASSIGN to predict the OxPhos activity in each cell-line and then compared the predictions with the functional validation (see Figure 5.8). There was only one cell-line (A-375) where the prediction is not concordant with the functional validation.

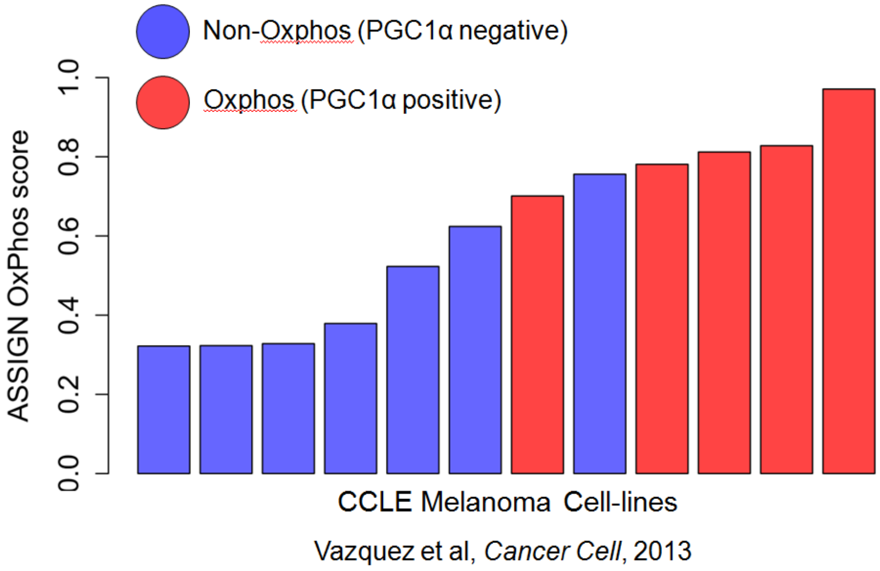


Figure 5.8: OxPhos predictions in melanoma cell lines

Here we show the OxPhos activity predictions of the melanoma cell-lines that were functionally validated in Vasquez et al 2013. The PGC1α positive cell-lines, which are functionally shown to be OxPhos dependent, are shown in red, the OxPhos independent cell-lines in blue. The cell-lines are ordered by our OxPhos activity predictions.

Based on these first successful results, we next turned to the functional characterization of lung and breast cancer cell lines. To this end, we selected the cell lines corresponding to the 3 top OxPhos predictions and the 3 top non-OxPhos predictions from each of the CCLE lung and breast datasets, and performed functional validation on these cell-lines through our collaborators in the Danial laboratory. Three metabolic assays were performed on the selected cell lines, with the inclusion of DLBCL cell lines as additional control in every experiment. Based on the result of previous studies with DLBCL cell lines (Caro et al. 2012), OxPhos dependency can be functionally defined by:

1) sensitivity to hypoxia; 2) higher mitochondrial oxygen consumption rates (OCR) in the presence of palmitate as a sole substrate; and 3) lower glucose derived lactate.

Survival in Hypoxia conditions (Figure 5.9 and Figure IV.2): Hypoxia is a way to trigger anaerobic glycolysis due to oxygen deprivation. Cells with non-OxPhos phenotypes are predicted to survive hypoxic conditions as in Warburg-type cancers. OxPhos-dependent cells are expected to selectively die under the same conditions (as seen with DLBCL cell lines). However, as shown in the Figures, tests performed in both tissue types (lung and breast) did not show any significant difference in the sensitivity of the predicted OxPhos and non-OxPhos cell lines to hypoxia.

OCR assays (Figure 5.10): Due to the morphological and growth rate differences between all cell lines, each experiment was optimized per cell line and normalized for protein content. There is no marked difference in OCR when lung or breast cancer cells were provided with the fatty acid Pamitate vs. no substrate. For breast cancer cells, the predicted non-OxPhos subgroup did not respire on Palmitate and was not suitable for OCR experiments. As a comparison, OCR traces for Ly4 and U2932 DLBCL cell lines are shown on the right.

Lactate assays (Figure 5.11): In this assay, the non-OxPhos phenotype is predicted to have higher glucose-derived lactate that OxPhos cells. However, there are no significant differences between the cells predicted to be OxPhos vs. non-OxPhos in either lung or breast cancers.

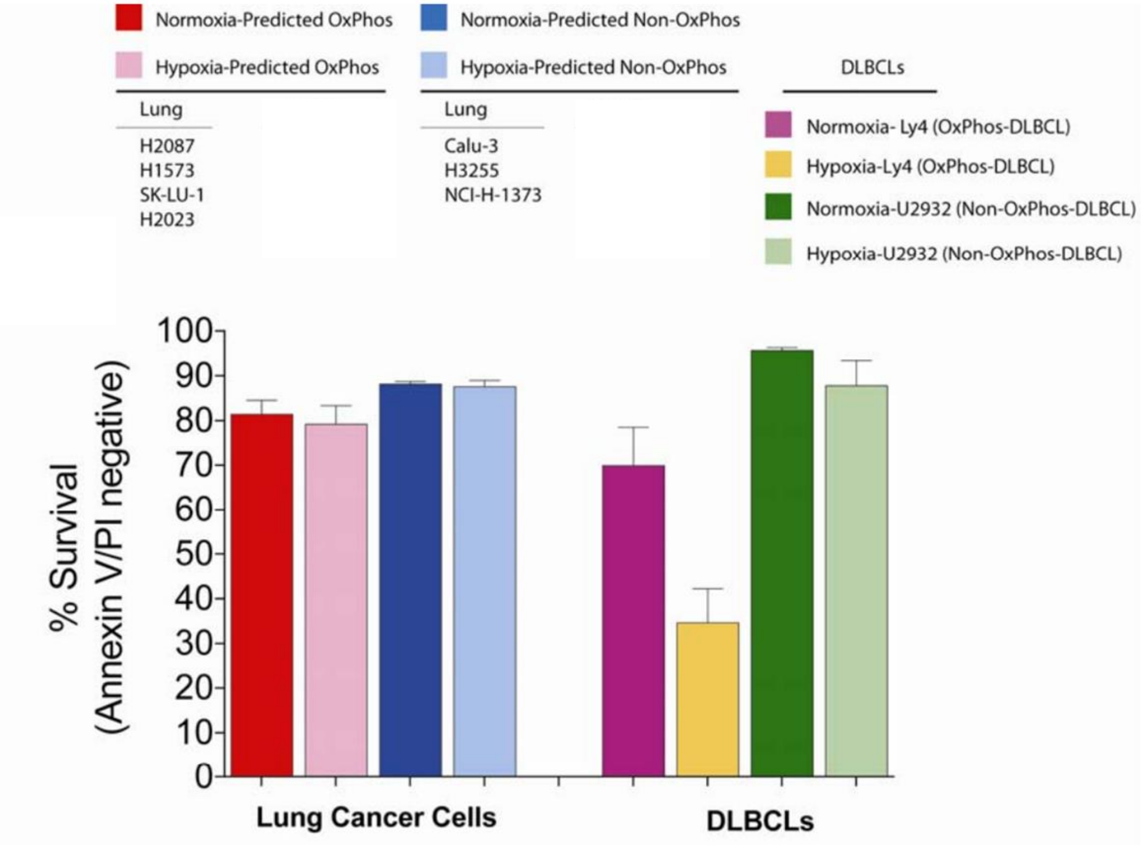


Figure 5.9: Survival of lung cancer cell-lines after 72h in 1% hypoxia and normoxia conditions

(Courtesy of Nika Danial) In this barplot we show the survival rate of lung and DLBCL cell-lines in dependence on oxygen availability. On the right we show the gold standard DLBCL cell-lines. Ly4 a OxPhos dependent cell-line shows a strong decrease in survival rate when comparing normal oxygen levels (purple) and hypoxia conditions (yellow), while U2932, which is not OxPhos dependent does not show a similar drop in survival rates (dark vs. light green) On the left we show the comparison between normoxia and hypoxia conditions for 7 lung cancer cell-lines that were predicted as OxPhos dependent (red) and cell-lines that were predicted as nonOxPhos (blue).

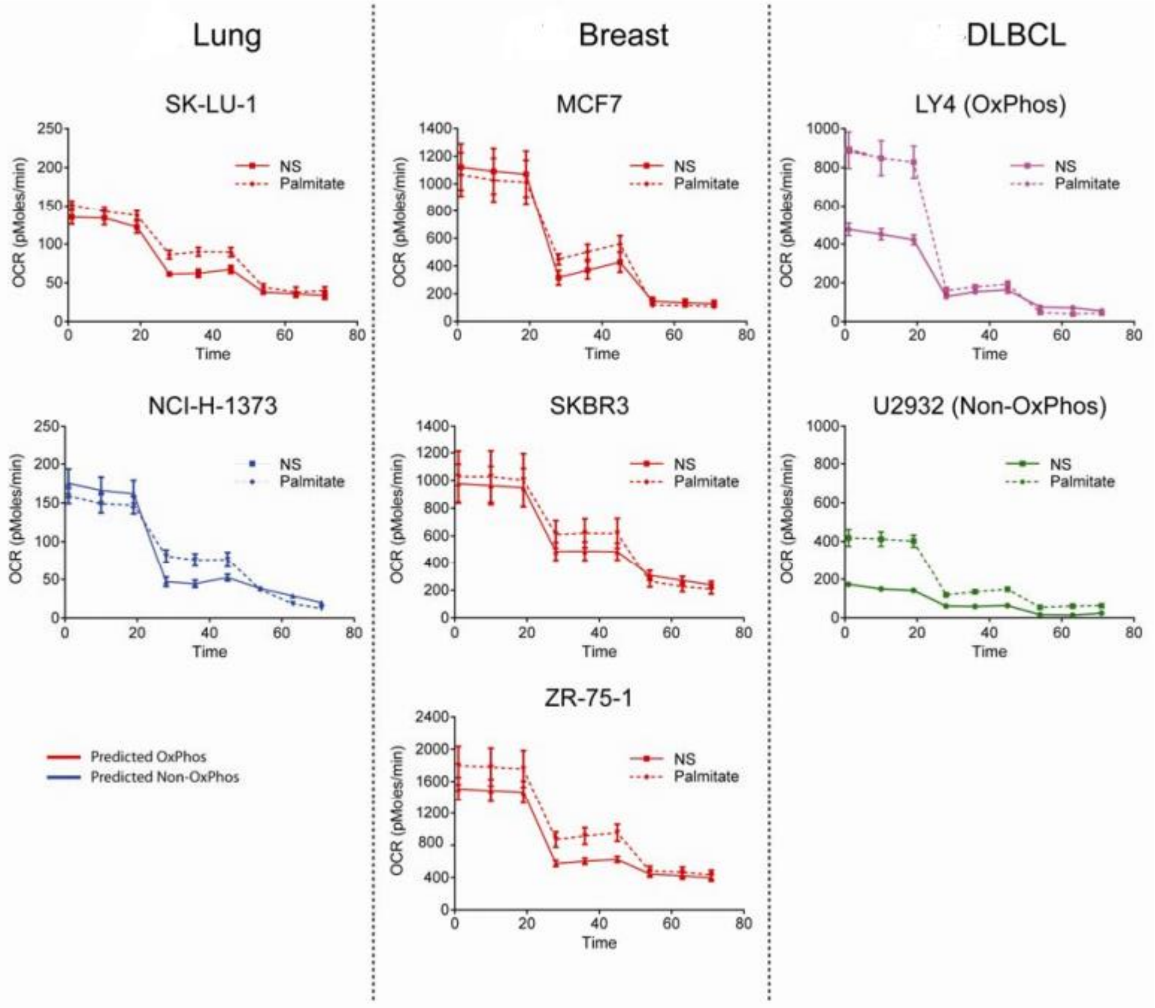


Figure 5.10: Representative oxygen consumption rate (OCR) assay in cell-lines

(Courtesy of Nika Danial) OCR traces in the presence (dotted line) or absence (solid line) of Palmiate for lung, breast and DLBCL cell-lines. Red and blue traces the predicted OxPhos and nonOxPhos, respectively. For each OCR trace the first dip ~25 min indicates addition of the ATP synthase inhibitor Oligomycin and the second dip ~55 min indicates addition of rotenone and antimycine.

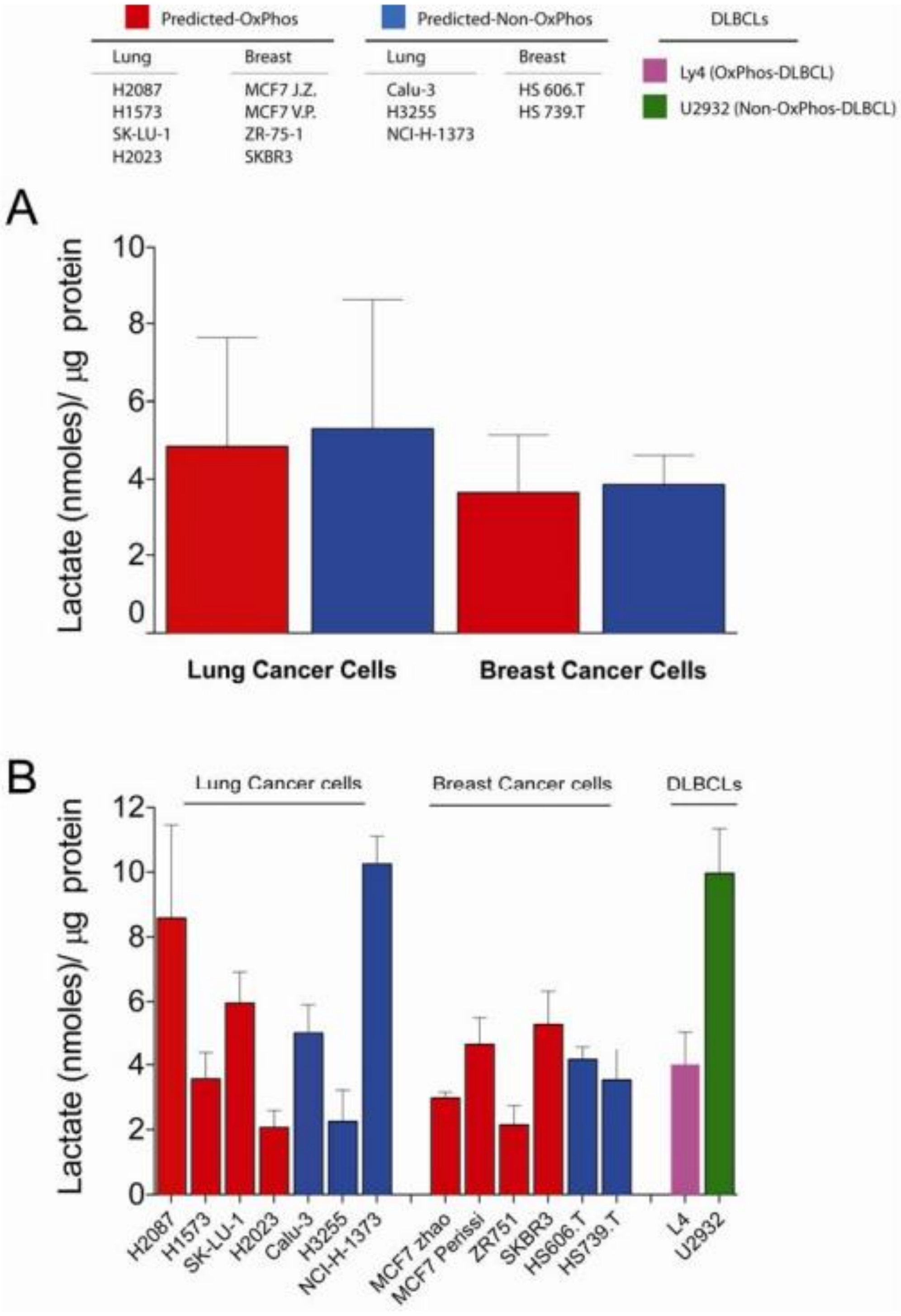


Figure 5.11: Glucose-derived lactate in the indicated lung and breast cancer cell lines.

(Courtesy of Nika Danial) (A) Cumulative data per predicted subtypes for each cancer is shown. (B) Lactate production per individual cell lines used to derive 3A. Ly4 and U2932 DLCBL cell lines were used as OxPhos and non-OxPhos controls, respectively.

5.4.6 Deriving a Pan-OxPhos signature

Adapting the generalized version of the DLBCL OxPhos signature to different tissue types gave us the opportunity to determine how many of the genes were concordantly up-regulated in a variety of different cancers and to also identify tissue specific genes. Figure 5.12 shows a Venn-Diagram of all TCGA datasets as well as the lymphoma and melanoma sets. There is only one gene difference between the TCGA lung adenocarcinomas and lung squamous carcinomas, so we intersected these two sets as representative for lung cancer. There is a large agreement among all sets and an overlap of 85 genes, which we define as our tissue-independent Pan-OxPhos signature. The genes in this signature are shown in Table IV.4 and were used for most of the analyses in the following sections. Interestingly, the second biggest overlap of 23 genes is between all four TCGA sets, excluding the Affymetrix datasets. We suspect that this is due to differences in detection levels, since RNASeq as a technology can detect genes at a lower expression level than one-color microarrays.

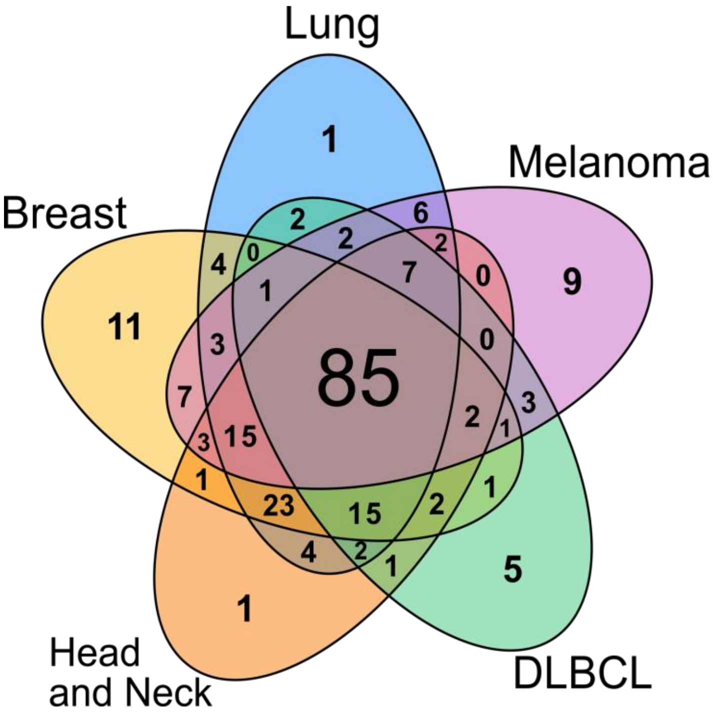


Figure 5.12: Overlap of tissue specific OxPhos signatures

In this Venn-diagram we show intersects of all tissue specific OxPhos gene sets as derived from ASSIGN. Lung represents both TCGA LUSC (lung squamous carcinoma) and LUAD (lung adenocarcinoma), since these two differed only by one gene. Lung, breast and head and neck represent RNASeq datasets, whereas DLBCL and melanoma are Affymetrix datasets.

5.4.7 Association with mutations and somatic copy number variations

As described in the methods section, we preprocessed mutation (MutSigCV) (Lawrence et al. 2013) and somatic copy number variation (GISTIC 2.0) (Mermel et al. 2011) data for all TCGA sample and tested for association between somatic alterations and the OxPhos phenotype by a Kolmogorov-Smirnov (KS) test. After correcting for multiple testing by the false discovery rate method, we found only one significant (FDR<0.25) mutation in breast cancer in the TBL1XR1 gene. We show the OxPhos activity and the corresponding TBL1XR1 mutations for the TCGA BRCA samples in Figure 5.13. There are only 10 mutations of that gene in that dataset and they seem to be enriched in the non-OxPhos side.

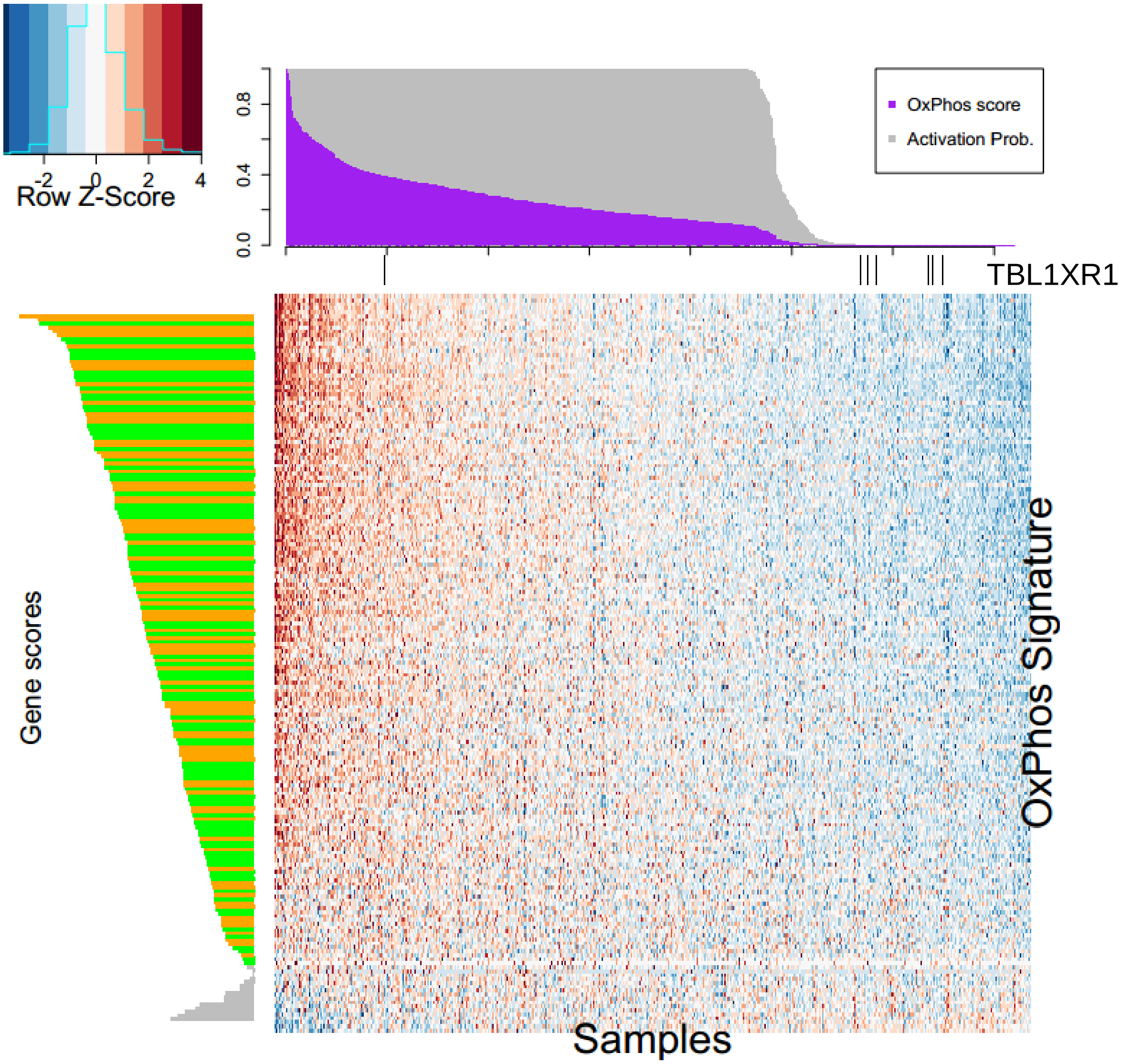


Figure 5.13: Association between OxPhos activity and TBL1XR1 mutation in breast cancer

In this heatmap, we show the gene expression of the TCGA BRCA dataset, in OxPhos gene signature space. Red values show up-regulated gene expression, blue down-regulated. The samples are ordered according to predicted OxPhos activity (purple) by ASSIGN. The genes are ordered by ASSIGN gene weight (original DLBCL OxPhos gene signature members in green, expanded set in orange). The black and white bar on top of the heatmap shows the presence (black) of TBL1XR1 mutations, which are enriched in the OxPhos active samples.

5.4.8 Finding upstream regulators using Ingenuity

Next, we performed Ingenuity Pathway Analysis (Krämer et al. 2014) to investigate possible mechanisms of action contributing to the OxPhos phenotype. Ingenuity offers a suite of tools that allows for the annotation of gene signatures, but also includes more sophisticated regulatory network based approaches such as the Upstream Regulator Analysis (URA). URA determines likely upstream regulators and is not limited to transcription factors, but includes any gene or small molecule that has been observed experimentally to affect gene expression in some direct or indirect way (Krämer et al. 2014). Table 2.1 shows the results of an URA on our 85 gene Pan-OxPhos signature. Of note is that the gene RICTOR, which is the binding partner of MTOR in the MTORC2 complex, has a very significant p-value of 5.06E-60.

Table 5.3: Ingenuity upstream regulators

Predicted up-stream regulators of the PanOxPhos signature as derived by Ingenuity.

Upstream Regulator	Molecule Type	p-value
RICTOR	Other	5.06E-60
guanidinopropionic acid	chemical - endogenous	3.35E-15
IGF1R	transmembrane receptor	6.90E-14
5-fluorouracil	chemical drug	1.69E-13
sirolimus	chemical drug	3.15E-13
Esrra	transcription regulator	5.35E-13
MAPT	Other	1.37E-11
CD 437	chemical drug	2.32E-11
INSR	kinase	7.06E-10
mono-(2-ethylhexyl)phthalate	chemical toxicant	8.34E-10
MYCN	transcription regulator	2.56E-09
FOXO1	transcription regulator	3.39E-09
VEGFA	growth factor	1.30E-08
interferon beta-1a	biologic drug	2.52E-08
PSEN1	peptidase	2.32E-07
HTT	transcription regulator	4.70E-07
NRF1	transcription regulator	1.18E-06
APP	Other	4.51E-06
MYC	transcription regulator	2.01E-05

5.4.9 Identification of related compounds using connectivity map 2.0

An additional approach of annotating a list of genes is through the Connectivity Map or cMap (Lamb et al. 2006; Lamb 2007). The cMap 2.0 (http://www.lincscloud.org/) contains more than a million gene expression signatures corresponding to hundreds of thousands of perturbations in multiple cell lines. Perturbations include treatments with a large panel of drugs, as well as knock-down and over-expression experiments of single genes. These perturbation signatures can be queried with a custom list of genes, such as our PanOxPhos signature, yielding a list of signatures that are most correlated or anti-correlated. We show the outcome of this cMap 2.0 analysis for chemical compounds in Table 5.4. Of note is that the top anti-correlated signature belongs to the compound wortmannin, which is known to inhibit PI3K related enzymes such as mTor (Feldman & Shokat 2010) and torin-2 is a potent and selective mTOR inhibitor (Liu et al. 2013).

Table 5.4: Compounds signatures (anti-)correlated to the PanOxPhos signature

For this analysis we submitted our 85 gene PanOxPhos signature to the connectivity map 2.0 and looked for compounds that show correlated gene expression signatures. On the left we show the compounds that are anti-correlated and on the right we show compounds that are correlated.

pert_iname	mean rank
wortmannin	-98.3578
torin-2	-98.2349
withaferin-a	-97.9292
Oxetane	-97.5076
ST-4029573	-97.3765
Salermide	-96.7146
BRD-K58214070	-96.5766
SCH-79797	-96.5565
Calyculin	-96.4921
EMF-sumo1-11	-96.1698

pert_iname	mean rank
RHO-kinase-inhibitor-III[rockout]	98.1762
SJ-172550	97.5339
Triptolide	96.5122
Linezolid	96.03
AS-703026	96.0037
Dexamethasone	95.8667
Orteronel	95.8092
Isoxicam	95.644
BRD-K11671649	95.5445
Nifedipine	95.4223

5.4.10 Looking at RICTOR and MTOR expression

Since both the ingenuity pathway analysis and the connectivity map analysis showed a potential link between our 85 gene PanOxPhos signature and the mTOR (mammalian target of rapamycin) and its binding partner RICTOR (rapamycininsensitive companion of mTOR), we next investigated the correlation between the gene expression of these two genes and the OxPhos activity in all of our datasets. The results for RICTOR, shown in Figure IV.3, indicate a consistent weak negative correlation between -0.18 and -0.682 in all of the TCGA datasets. The results for mTOR, shown Figure 5.14, indicate a strong negative correlation between -0.511 and -0.654, which is consistent with the results in the previous sections. However, when we performed the same analysis in cell-lines we saw a positive correlation between mTOR and OxPhos activity between 0.138 and 0.478. This indicates a systematic difference of the metabolic characteristics between cancer cell-lines and primary tumors, possibly suggesting in-vitro growth conditions that do not adequately mimic their in-vivo counterpart..

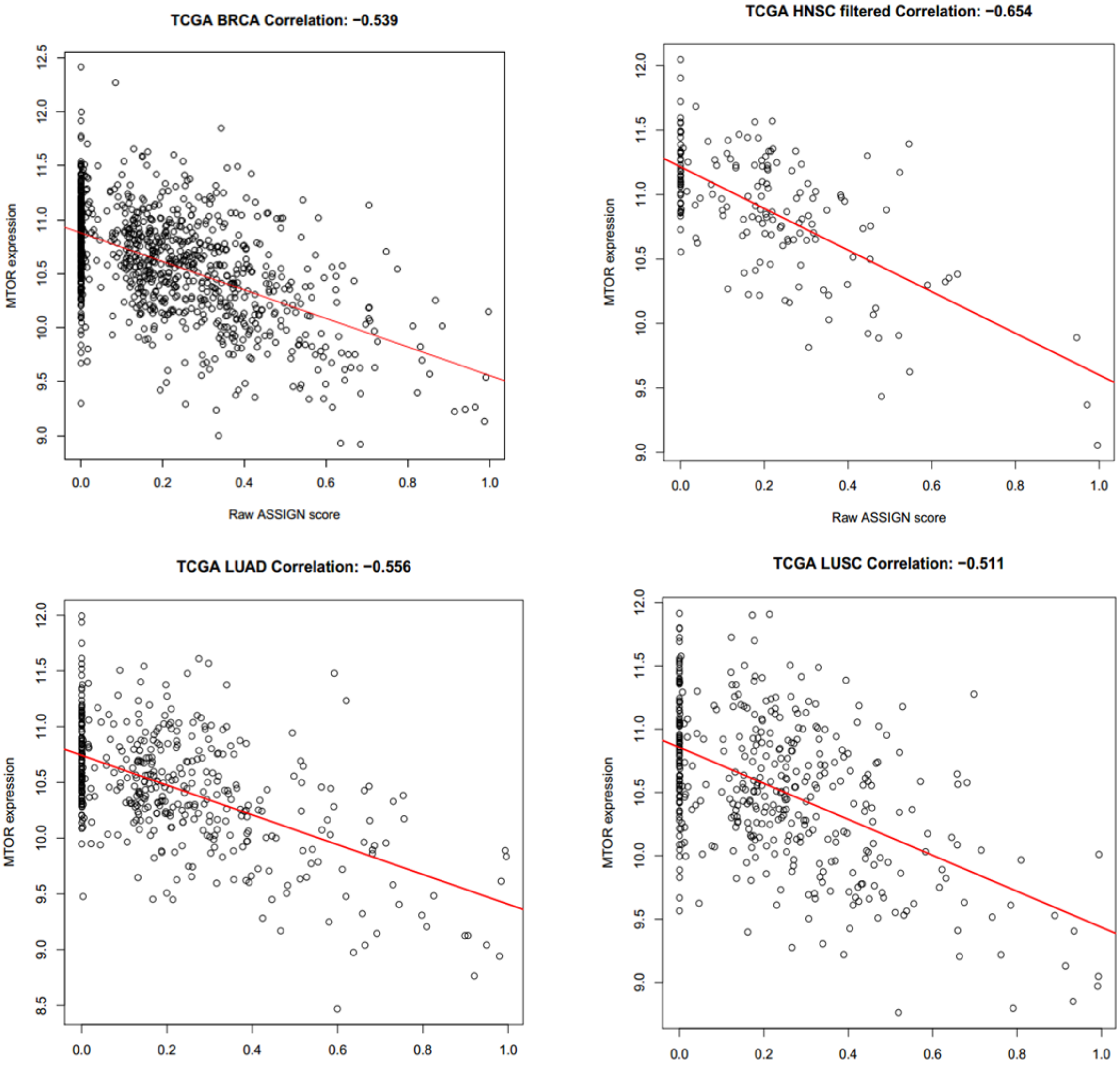


Figure 5.14: Expression of MTOR vs. OxPhos activity across primary tumor datasets

We show the gene expression of MTOR in comparison to OxPhos activity (Raw ASSIGN score) in TCGA BRCA, the oral cavity samples in TCGA HNSC, TGCA LUAD and TCGA LUSC. Red shows the trend-line based on a linear model.

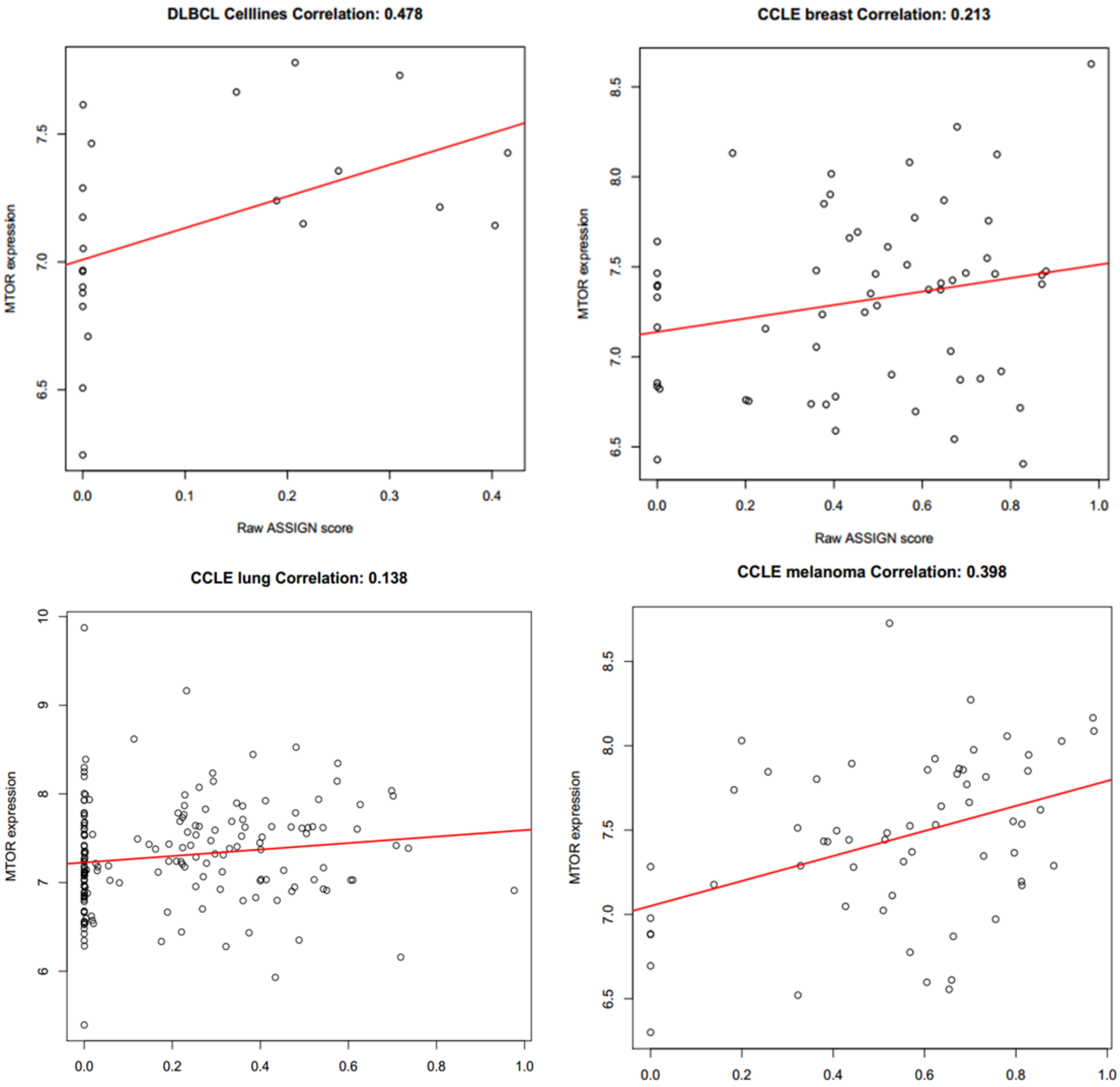


Figure 5.15: Expression of MTOR vs. OxPhos activity across cell-line datasets

We show the gene expression of MTOR in comparison to OxPhos activity (Raw ASSIGN score) in TCGA BRCA, the oral cavity samples in our in-house DLBCL cell-lines, CCLE breast, lung and melanoma lines.

5.5 Discussion

Based on our analysis we show that the therapeutically relevant OxPhos subtype described in (Monti et al. 2005; Polo et al. 2007) is associated with the PGC1α-high subtype described in (Vazquez et al. 2013). Both the heatmap in Figure 5.7 and the gene set enrichment analysis in Table IV.3 support this conclusion. With the link between those two subtypes established, we expanded our analysis to additional cancer types, which would suggest a broad relevance in cancer biology.

For this purpose we applied ASSIGN, a pathway-activity scoring method that is capable of adapting its predictions to different molecular contexts, as well as of assigning importance weights to a signature genes. ASSIGN has several advantages over simple clustering, but most importantly it provides a ranking of both the analyzed samples and genes with respect to the phenotype of interest. The generalized use of ASSIGN we adopted, where we add potentially interesting genes with a low prior probability of significance, makes it capable of fully adapting gene signatures not only across different tissue types, but also across quantification assays. We show an instance of these capabilities in the Venn diagram of Figure 5.12, where the RNAseq-based TCGA datasets include 23 genes that are either not present or very lowly expressed in the Affymetrix profiled datasets.

Using ASSIGN we were able to quantify OxPhos activity and to compare the predicted activation levels with functional validations in 12 melanoma cell-lines (Vazquez et al. 2013), which showed a strong concordance with our predictions.

However, when extending our validation to 6 breast cancer cell-lines and 7 lung cancer cell-lines, we did not find a similar concordance. Neither differences in OxPhos dependency, as measured by growth rates in hypoxia conditions, nor in OxPhos activity, as measured by oxygen consumption rate (OCR) assays, were detected.

We investigated potential links between the OxPhos activity and genomic alterations such as mutations and somatic copy number variations, using a KS-test as described in the methods section, but could not find any strong associations. The only significant result with an FDR of less than 25% was the TBL1XR1 gene in breast cancer. However, literature research did not yield any evidence that supports this link.

Finally, we used the intersection of all tissue specific OxPhos signatures to establish a tissue independent 85 gene PanOxPhos signature. We used Ingenuity Pathway Analysis and the Connectivity Map to annotate this list of genes. Both analyses highlighted a potential link between OxPhos activity and the mTor2 complex, which involves the proteins RICTOR and mTOR and has been shown to modulate resistance against oxidative stress in cancer cells (Lu et al. 2015; Cai & Andres 2014). Closer inspection of the correlation between OxPhos activity and mTor/RICTOR (Figure 5.14 and Figure IV.3) confirmed the association with mTorC2 in primary tumors of different tissue types. However, when looking at the same correlations in cell-lines (Figure 5.15), we detected a positive yet not significant correlation, instead of the strong negative correlation observed in primary tissues, indicating that the cell lines do not adequately capture in vivo conditions. We suspect this might be due to differences in available nutrients; cell-lines are usually grown with an abundance of nutrients, however, short of reprofiling the gene expression of cell-lines grown in conditions that resemble primary tumors more closely, this remains speculation.

APPENDIX I

Genomic models of short-term exposure accurately predict long-term chemical carcinogenicity and identify putative mechanisms of action

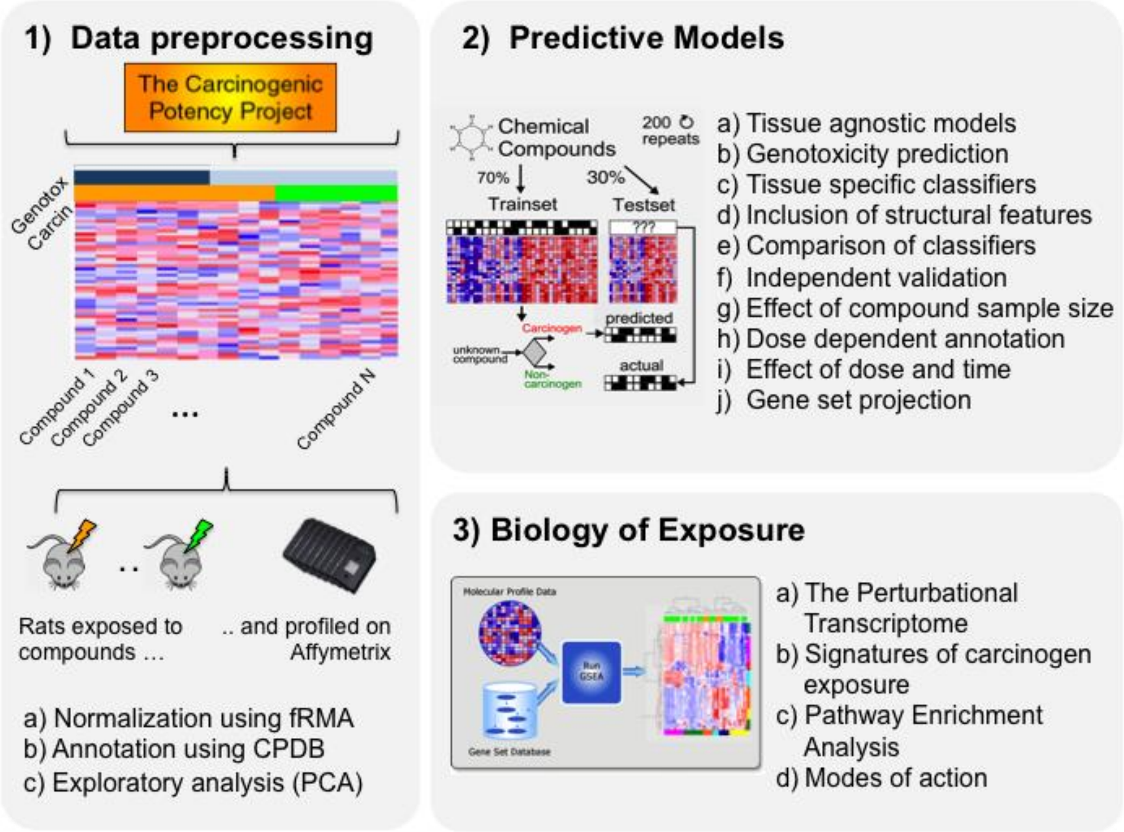

1) Data preprocessing
The Carcinogenic Potency Project
Genotox
Carcin
Compound 1
Compound 2
Compound 3
...
Compound N
Rats exposed to compounds …
.. and profiled on Affymetrix
a) Normalization using fRMA
b) Annotation using CPDB
c) Exploratory analysis (PCA)
2) Predictive Models
Chemical Compounds
200 repeats
70%
30%
Trainset
Testset
???
predicted
actual
Carcinogen
unknown compound
Non-carcinogen
a) Tissue agnostic models
b) Genotoxicity prediction
c) Tissue specific classifiers
d) Inclusion of structural features
e) Comparison of classifiers
f) Independent validation
g) Effect of compound sample size
h) Dose dependent annotation
i) Effect of dose and time
j) Gene set projection
3) Biology of Exposure
Molecular Profile Data
Run GSEA
Gene Set Database
a) The Perturbational Transcriptome
b) Signatures of carcinogen exposure
c) Pathway Enrichment Analysis
d) Modes of action

Figure I.1: Overview of the analysis.

The study presented here consists of three parts: 1) Preprocessing, annotation and exploration of the data. 2) Building classification models to predict carcinogenicity in rats, which includes the investigation of the effects of dose-, time-, and tissue-specificity, effects of sample size, and others. 3) Biology of exposure, where we defined carcinogenicity signatures, investigated enriched pathways and derived putative modes of action.

Discussion

***Cost-benefit analysis*.**

If we assume that about 10% of the 84,000 chemicals currently in commercial use are carcinogens (Waters et al. 2010), classification of the complete set based on our classifier optimized on a 1:1 FP/FN cost function would yield approximately 4400 predicted carcinogens – of which 1285 would be expected false positive (based on the sensitivity/specificity as assessed by training on DM and testing on TGG, see Figure 2.3) – and about 5200 carcinogens would be missed (FN). If we wished to reduce the number of FPs to 500, corresponding to a specificity of ~99.3%, this would translate into a sensitivity of ~20.9%, and lead to the detection of 1756 out of the expected 8400 true carcinogens. Conversely, adopting a 1:2 FP/FN cost function would lead to an increased sensitivity of 88.4% and a drop in specificity to 36.3%. These scenarios are presented to show the considerable flexibility afforded by the classifier, and to emphasize that the appropriate specificity/sensitivity trade-off will be determined by the main purpose for which the classifier is used. If its primary purpose is to prioritize compounds for further screening, a high sensitivity (few FNs) would be preferable, even at the cost of a lower specificity (more FPs). On the other hand, if its purpose is to prove conclusively that a compound is carcinogenic (e.g., for regulatory purposes), then increasing the specificity even at the cost of a lower sensitivity might be preferable.

Structural features as predictors.

Evaluation of the relative predictive power of gene expression and chemicals' structural features conclusively shows the higher information content of the former over the latter, but also shows that augmenting the prediction models with such structural information marginally improves classification, in particular genotoxicity. The top structural features as ranked by the Random Forest variable importance include chloride.p.alkyl, halde..p.alkyl, nitrosamine, nitrose and benzene.1.alkyl.4.carbonyl, among others, which enable compound-DNA interaction and consequently are predictive of genotoxicity. Since the 3D structural features are easily accessible for most compounds, it seems sensible to incorporate these in any future classifier.

Material

For the Gene set enrichment analysis as well as the projection into pathway space we used the gene sets of the canonical pathways in the second compendium of the molecular signature database (MSigDB) (Subramanian et al. 2005) version 3.0, which includes 880 gene sets. All gene sets were mapped from human gene symbols to rat Ensembl gene identifiers using the R/Bioconductor package `BiomaRt`.

For the DrugMatrix, each compound is annotated with 1,902 dichotomous chemical structure descriptors extracted from the Leadscope Enterprise 3.0 software package (Columbus, Ohio). All samples were profiled on the Affymetrix Rat 230.2 microarray.

Methods

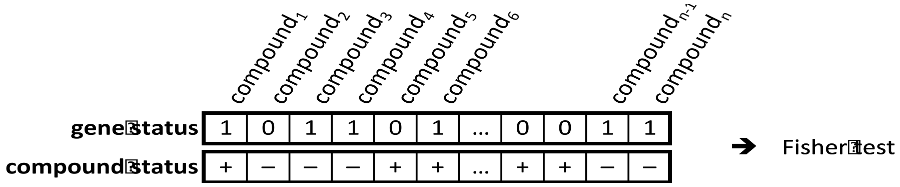
Exploratory analysis

In order to reduce the dimension of the dataset and have a 2 or 3-dimensional representation of the dataset we used Principal Component Analysis (PCA) using the R package *prcomp* and Multidimensional Scaling (MDS) using the R package *ggplot2*.

Defining the Perturbational Transcriptome

The list of genes that significantly respond to chemical perturbation was identified by carrying out a two-group moderated t-test between the control samples and the corresponding treatment samples *for each* compound (at a given dose) separately, while correcting for the confounding effect of time. Only the genes with FDR-corrected q-value≤0.01 and fold-change≥1.5 (in either direction) in at least five compounds were included. A gene-by-compound matrix was then constructed, with each column representing the vector of "control *vs*. treatment" t-scores for the corresponding compound. A total of 191 compound-dose instances, corresponding to 138 distinct compounds for which either carcinogenicity or genotoxicity information was available, were included in this analysis. Hierarchical clustering of both the compounds and the genes based on the t-scores' matrix was performed, and the results visualized in a heatmap with the color-coding based on the t-test's q-values and the direction of the up-regulation (Figure 2.2). The procedure yielded a clear two-cluster stratification, with one of the clusters highly enriched for carcinogenic compounds. Association between cluster membership and carcinogenicity (genotoxicity) status of the compounds was assessed by Fisher test.

Each gene was tested for its association with carcinogenicity, by performing a Fisher test between the gene status (0: not differentially expressed; 1: differentially expressed) and the compound status (+: carcinogenic; –: non-carcinogenic) across compounds, and the nominal p-values were corrected for multiple hypothesis testing by the FDR procedure (Figure 2.2b, columns grouped under ‘Enrichment’).



	compound$_1$	compound$_2$	compound$_3$	compound$_4$	compound$_5$	compound$_6$				compound$_{n-1}$	compound$_n$
gene status	1	0	1	1	0	1	...	0	0	1	1
compound status	+	–	–	–	+	+	...	+	+	–	–

To test whether the number of genes up-/down-regulated by each compound was significantly higher in carcinogens than in non-carginogens, a Kolmogorov-Smirnoff test was performed as shown in Figure I.2. The test evaluates whether the distribution of carcinogenic compounds is significantly skewed toward either ends of the list of compounds sorted according to the number of genes they up-/down-regulate. The results show a significant over-representation of carcinogenic compounds toward the high-end of the sorted list.

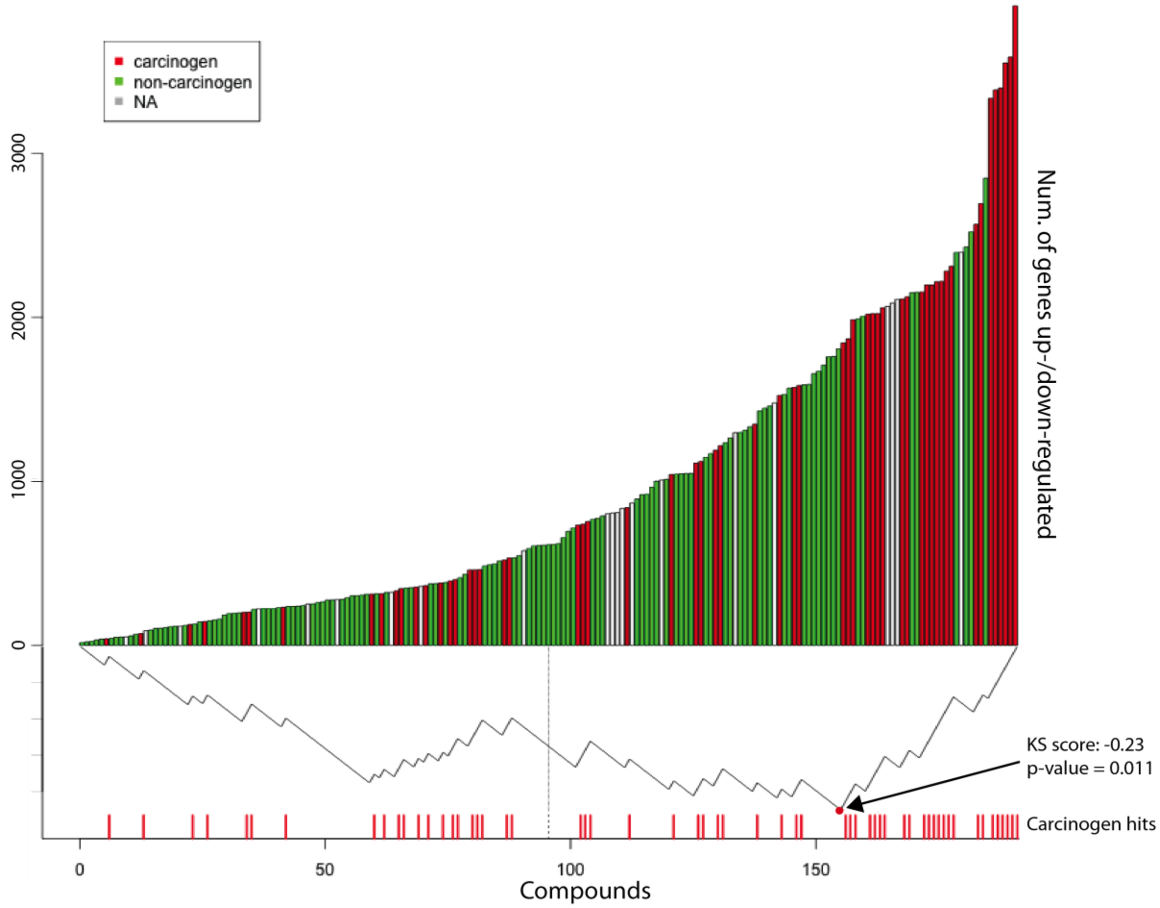


Figure I.2 - Distribution of differentially regulated genes

Number of up-/down-regulated genes across compounds. The carcinogenic compounds (red ticks) are significantly skewed toward the right-end of the distribution, as measured by a KS test (bottom).

Tissue-agnostic carcinogenicity classifiers

We first assessed whether it is possible to predict the carcinogenicity of a compound independent of the tumor site. To this end, Random Forest classifiers were built from the DrugMatrix liver samples using tissue agnostic carcinogenicity labels, whereby a compound is labeled as carcinogenic if it is found to induce cancer in any tissue type at any dose. The random resampling-based estimation of classification performance yielded an AUC of 64.8% when predicting carcinogenicity in this fashion (Table I.1 and corresponding ROC curves in Figure I.3).

Mode of Action Figure

For Figure 2.6b we used the top 50 pathways as ranked their variable importance for classifying the carcinogenic potential of a chemical compound. The pathways as well as the chemical compound were grouped using hierarchical clustering. In order to acquire the driving genes for each cluster or mode of action we clustered the chemical compounds only in the space of the pathways of a given mode of action. We then split these hierarchical clusters in two groups at the top node of the dendrogram and went back to the actual gene expression data for these two groups, where we performed differential gene expression analysis (limma) between those groups in order to get a gene ranking. We then reduced the list of genes to those that are present in any of the pathways that defined a given mode of action and reported the top ranking genes (Figure 2.6c – right column).

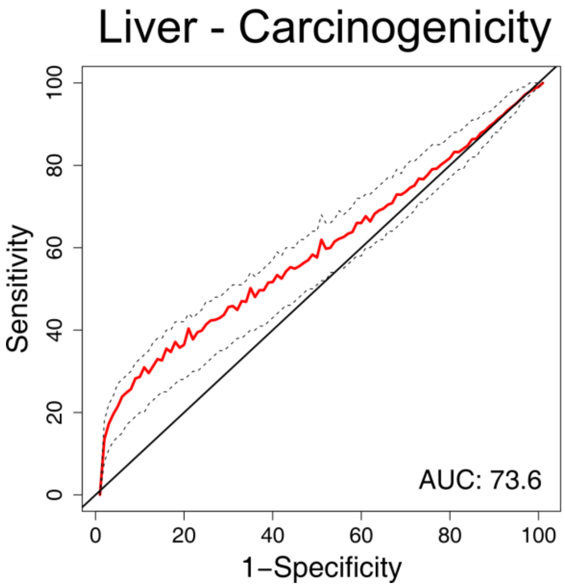


Figure I.3 - Tissue-agnostic carcinogenicity prediction

ROC curves corresponding to random forest classifiers trained on liver samples but using tissue-agnostic carcinogenicity labels. The red curves show the means over 200 iterations of a 70%/30% train/test dataset split, whereas the dashed curves indicate the first and third quartiles respectively.

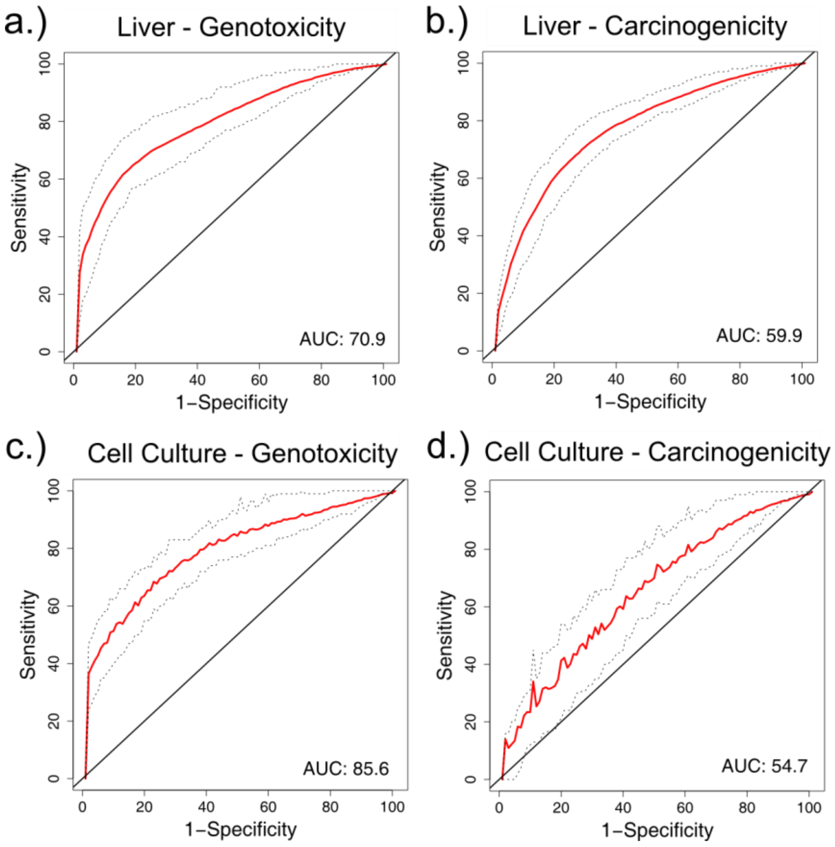


Figure I.4 – Prediction based on chemicals' structural features

ROC curves corresponding to random forest classifiers using chemicals' structural features as predictors. The red curves show the means over 200 iterations of a 70%/30% train/test dataset split, whereas the dashed curves indicate the first and third quartiles respectively.

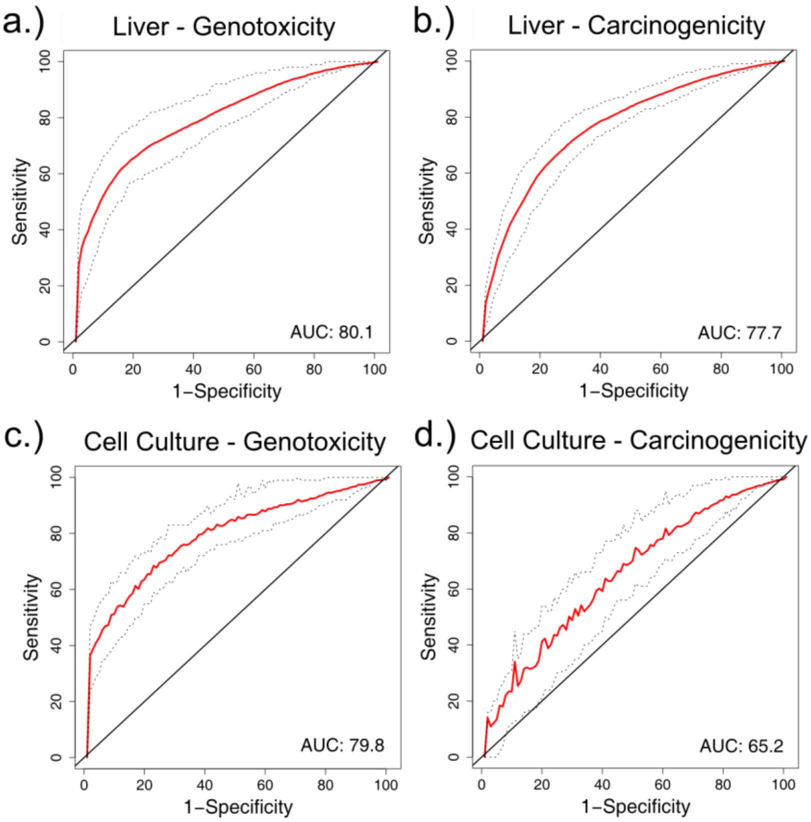


Figure I.5 – Prediction based on gene expression and chemicals' structural features

ROC curves corresponding to random forest classifiers using the expression of the 500 genes with highest variance *and* chemicals' structural features as predictors. The red curves show the means over 200 iterations of a 70%/30% train/test dataset split, whereas the dashed curves indicate the first and third quartiles respectively.

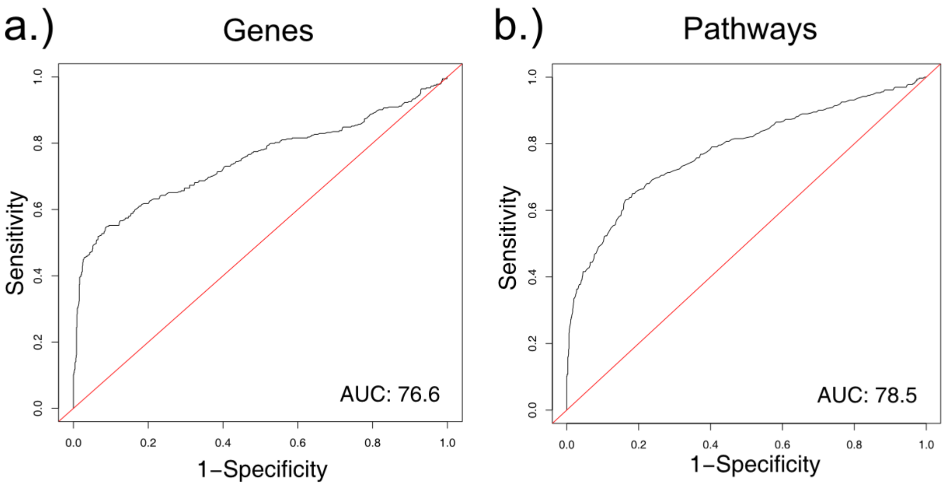


Figure I.6 – ROC of models trained on the DrugMatrix and tested on TG-GATEs

We trained a prediction model on all liver samples in the DrugMatrix and predicted the class labels of samples in the TG-GATEs treated with chemicals not included in the DrugMatrix. a) ROC curve for the gene-based predictions and b) ROC curve for the pathway-based predictions (see Methods).

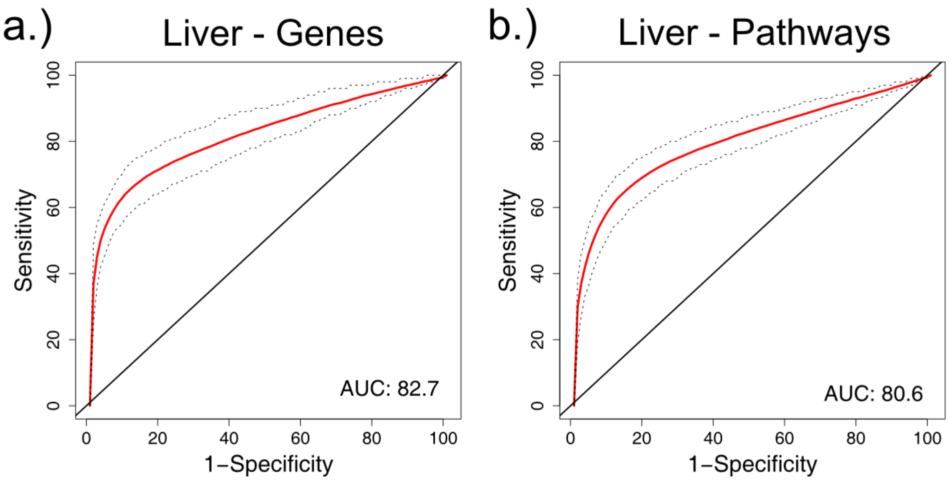


Figure I.7 – ROC of TG-GATEs cross-validation tests

ROC curves corresponding to random forest classifiers trained and tested on TG-GATEs. The train/test split was repeated 200 times to get estimates on the 95% confidence interval. a) results of the gene-based predictions and b) results of the pathway-based predictions (see Methods). The red curves show the means over 200

iterations of a 70%/30% train/test dataset split, whereas the dashed curves indicate the first and third quartiles respectively.

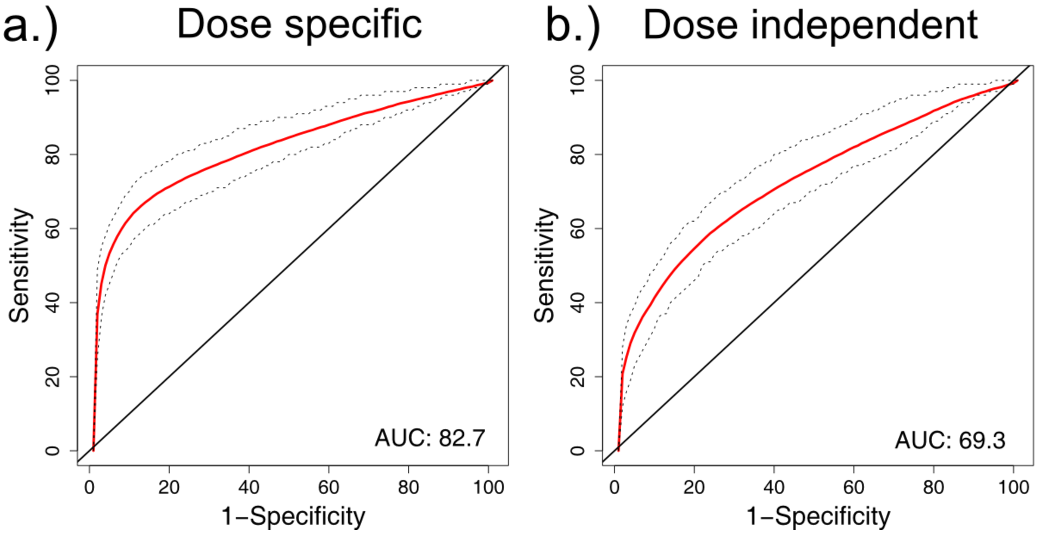


Figure I.8 – Effect of dose dependence on prediction

ROC curves corresponding to random forest classifiers trained on a) dose-specific carcinogenicity labels; and b) dose-independent carcinogenicity labels. For the dose-independent labels we used the annotation at the maximum dose and used it for all other doses. The red curves show the means over 200 iterations of a 70%/30% train/test dataset split, whereas the dashed curves indicate the first and third quartiles respectively.

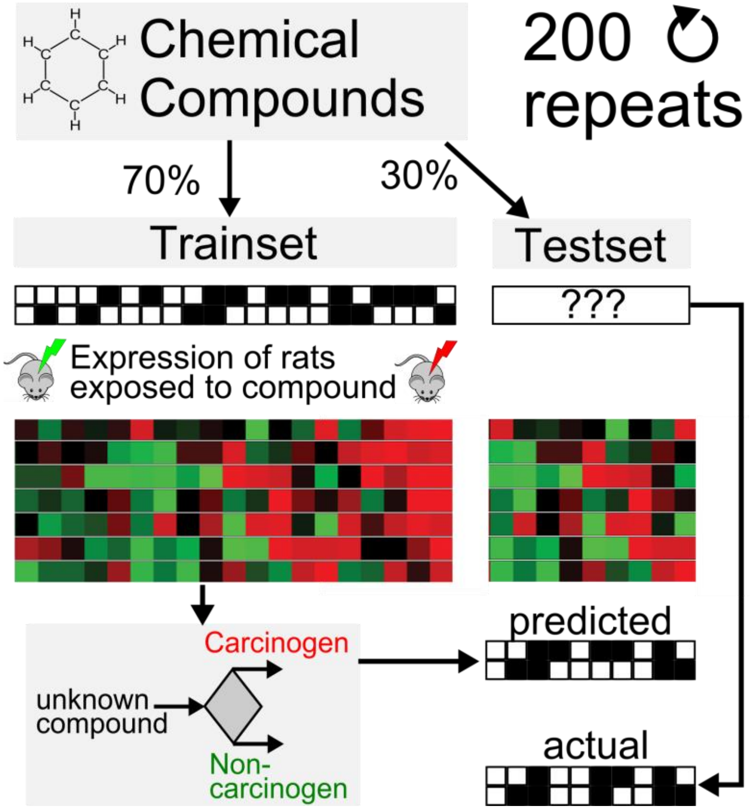


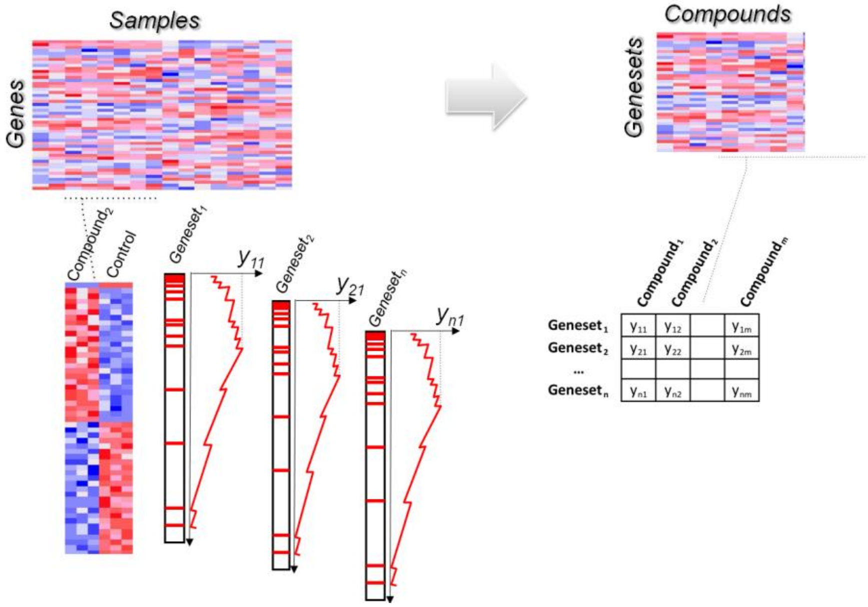
Figure I.9 – Random resampling scheme

Chemical compounds are split into a 70% training set and a 30% test set (stratified with respect to the phenotype to be predicted). The gene expression profiles associated with the training set are then used to train a classification model, which is used to predict the class labels of the test set. The predicted class labels are then compared with the actual labels and the prediction performance (AUC) can be evaluated. To achieve a robust evaluation and get an estimate of the standard error the random 70%/30% split is repeated 200 times.

.

Geneset Projection

increasing interpretability and robustness



	Compound$_1$	Compound$_2$		Compound$_m$
Geneset$_1$	y_{11}	y_{12}		y_{1m}
Geneset$_2$	y_{21}	y_{22}		y_{2m}
...				
Geneset$_n$	y_{n1}	y_{n2}		y_{nm}

Figure I.10 - Overview gene set projection

For each compound, a vector of *n* gene set enrichment scores were computed based on the “Compound vs. control” phenotype, where *n* is the number of gene sets. The original matrix of gene-by-compound is thus transformed into a gene set-by-compound matrix.

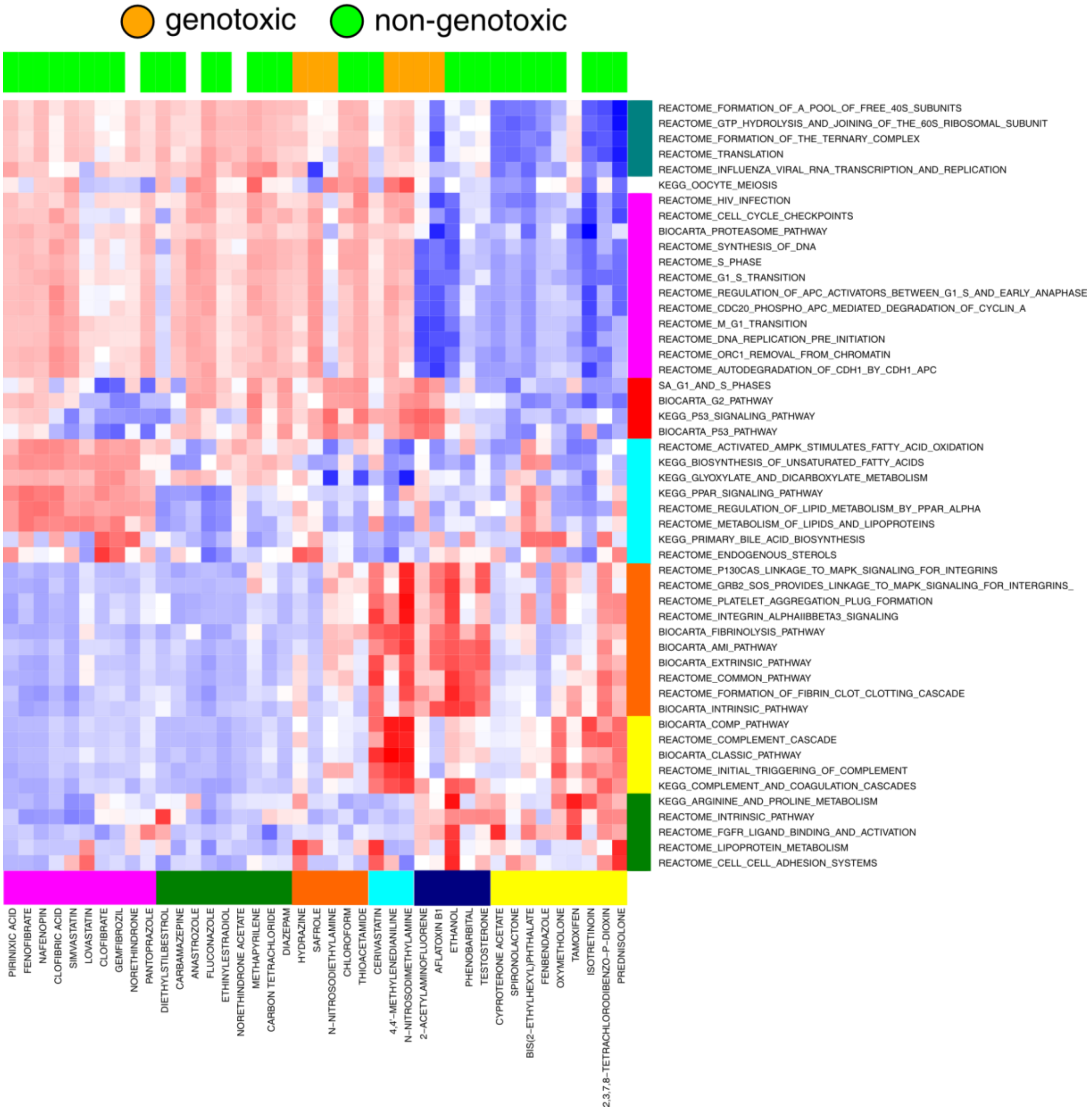


Figure I.11- Detailed Modes of Action of carcinogenic chemical compounds

Heatmaps of the top 50 pathways as ranked by their variable importance derived from a random forest classifier of hepato-carcinogenicity. Rows correspond to pathways, clustered into biological processes; columns correspond to chemical compounds. The heatmap shows all carcinogenic compounds in the DrugMatrix,

respectively. Only profiles corresponding to maximum duration and dose treatments, with replicates averaged, are displayed.

Table I.1 - Differential expression of carcinogens vs. non-carcinogens

Comparison of gene expression between rats exposed to carcinogens and non-carcinogens in the Drug Matrix. Multiple replicates were averaged while controlling for the exposure time.

Class	FC	adj.P.Val	Name	Description
CARC	1.69	3.91E-21	DACT2	dapper, antagonist of beta-catenin, homolog 2 (Xenopus laevis)
CARC	1.72	1.95E-20	ZDHHC2	zinc finger, DHHC-type containing 2
CARC	1.42	7.37E-17	PTER	phosphotriesterase related
CARC	1.83	1.76E-16	CIDEA	cell death-inducing DFFA-like effector a
CARC	1.38	5.36E-16	ANXA7	annexin A7
CARC	1.58	7.44E-16	HSDL2	hydroxysteroid dehydrogenase like 2
CARC	1.78	8.08E-16	ACOT1	acyl-CoA thioesterase 1
CARC	1.47	2.64E-15	HEBP2	heme binding protein 2
CARC	1.52	5.80E-15	MYO5B	myosin VB
CARC	1.41	2.03E-14	PQLC3	PQ loop repeat containing 3
CARC	1.79	2.55E-14	ACOT1	acyl-CoA thioesterase 1
CARC	1.39	1.21E-13	NUDT7	nudix (nucleoside diphosphate linked moiety X)-type motif 7
CARC	1.89	3.20E-13	CPT1B	carnitine palmitoyltransferase 1B (muscle)
CARC	3.99	6.13E-13	ACOT1	acyl-CoA thioesterase 1
CARC	1.65	1.87E-12	AQP7	aquaporin 7
CARC	1.6	2.49E-12	ECI1	enoyl-CoA delta isomerase 1
CARC	1.54	3.05E-12	ME1	malic enzyme 1, NADP(+)-dependent, cytosolic
CARC	1.45	5.16E-12	SNX10	sorting nexin 10
CARC	1.42	1.12E-11	POLR3G	polymerase (RNA) III (DNA directed) polypeptide G (32kD)
CARC	1.7	2.01E-11	PEX11A	peroxisomal biogenesis factor 11 alpha
CARC	1.75	3.03E-11	AIG1	androgen-induced 1
CARC	1.35	3.63E-11	CYP2J2	cytochrome P450, family 2, subfamily J, polypeptide 2
CARC	1.38	1.22E-10	GNAI1	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 1
CARC	1.65	1.30E-10	PDK4	pyruvate dehydrogenase kinase, isozyme 4
CARC	1.47	7.67E-10	CCND1	cyclin D1
CARC	1.61	1.08E-09	VNN1	vanin 1
CARC	1.42	1.15E-09	SLC22A5	solute carrier family 22 (organic cation/carnitine transporter), member 5
CARC	1.37	1.22E-09	TMBIM1	transmembrane BAX inhibitor motif containing 1
CARC	1.42	2.34E-09	ECH1	enoyl CoA hydratase 1, peroxisomal
CARC	1.51	3.12E-09	HSPB1	heat shock 27kDa protein 1
CARC	1.56	3.60E-09	RAB30	RAB30, member RAS oncogene family
CARC	1.42	5.72E-09	CRAT	carnitine O-acetyltransferase
CARC	1.66	8.63E-09	HDC	histidine decarboxylase
CARC	1.37	2.21E-08	SPC24	SPC24, NDC80 kinetochore complex component, homolog (S. cerevisiae)
CARC	1.36	3.55E-08	SLC25A30	solute carrier family 25, member 30
CARC	1.36	4.66E-08	ACSL3	acyl-CoA synthetase long-chain family member 3
CARC	1.41	5.06E-08	MCM6	minichromosome maintenance complex component 6
NONCARC	0.48	5.07E-08	STAC3	SH3 and cysteine rich domain 3
NONCARC	0.73	3.11E-08	IL1R1	interleukin 1 receptor, type I
NONCARC	0.64	1.60E-08	NOX4	NADPH oxidase 4
NONCARC	0.7	1.30E-08	FMO1	flavin containing monooxygenase 1
NONCARC	0.73	8.87E-09	IL33	interleukin 33
NONCARC	0.69	8.36E-09	XPNPEP2	X-prolyl aminopeptidase (aminopeptidase P) 2, membrane-bound
NONCARC	0.71	3.83E-09	INHBC	inhibin, beta C
NONCARC	0.52	3.51E-09	CXCL1	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating

				activity, alpha)
NONCARC	0.73	1.71E-12	FAM46C	family with sequence similarity 46, member C
NONCARC	0.74	7.41E-13	HSD3B2	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2
NONCARC	0.73	1.46E-13	ARMC9	armadillo repeat containing 9
NONCARC	0.73	3.62E-14	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2
NONCARC	0.64	3.30E-14	CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2
NONCARC	0.68	2.55E-14	LIN7A	lin-7 homolog A (C. elegans)
NONCARC	0.68	2.26E-14	SLC16A10	solute carrier family 16, member 10 (aromatic amino acid transporter)
NONCARC	0.71	3.90E-16	NTF3	neurotrophin 3
NONCARC	0.52	3.77E-16	SEZ6	seizure related 6 homolog (mouse)
NONCARC	0.39	3.54E-18	A2M	alpha-2-macroglobulin

Table I.2 – Differential analysis of genotoxic carcinogens vs. non-genotoxic carcinogens

Comparison of gene expression between rats exposed to genotoxic carcinogens and non-genotoxic carcinogens in the DrugMatrix. Multiple replicates were averaged while controlling for the exposure time.

Class	FC	adj.P.Val	Name	Description
CARC_GT	1.37	9.02E-07	FAM49A	family with sequence similarity 49, member A
CARC_GT	1.69	9.02E-07	JAM3	junctional adhesion molecule 3
CARC_GT	1.76	6.06E-06	C8orf46	chromosome 8 open reading frame 46
CARC_GT	1.47	0.000148	PLN	phospholamban
CARC_GT	1.37	0.000188	SDC4	syndecan 4
CARC_GT	1.5	0.000203	CAV2	caveolin 2
CARC_GT	1.73	0.000402	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
CARC_GT	1.52	0.000502	MDM2	Mdm2, p53 E3 ubiquitin protein ligase homolog (mouse)
CARC_GT	1.39	0.000906	NFKBIZ	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, zeta
CARC_GT	1.42	0.000906	EDNRB	endothelin receptor type B
CARC_GT	1.37	0.000927	SULF2	sulfatase 2
CARC_GT	1.6	0.001673	CTGF	connective tissue growth factor
CARC_GT	1.35	0.001983	ZFP36	zinc finger protein 36, C3H type, homolog (mouse)
CARC_GT	1.45	0.002101	DUSP6	dual specificity phosphatase 6
CARC_GT	1.4	0.002585	HYAL3	hyaluronoglucosaminidase 3
CARC_GT	1.37	0.002585	NHEJ1	nonhomologous end-joining factor 1
CARC_GT	1.39	0.003549	AHR	aryl hydrocarbon receptor
CARC_GT	1.63	0.00428	CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2
CARC_GT	1.37	0.005501	PHLDA3	pleckstrin homology-like domain, family A, member 3
CARC_GT	1.39	0.00833	CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5
CARC_GT	1.44	0.00833	SLC25A25	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 25
CARC_GT	2.26	0.008475	CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1
CARC_GT	1.42	0.008501	RGS2	regulator of G-protein signaling 2, 24kDa
CARC_GT	1.43	0.008933	TP53INP1	tumor protein p53 inducible nuclear protein 1
CARC_GT	1.36	0.009854	CCNG1	cyclin G1
CARC_GT	1.74	0.010164	BCL6	B-cell CLL/lymphoma 6
CARC_GT	1.75	0.010827	CYP2C18	cytochrome P450, family 2, subfamily C, polypeptide 18
CARC_GT	1.57	0.012712	BTG2	BTG family, member 2
CARC_GT	1.37	0.012782	HLA-DRA	major histocompatibility complex, class II, DR alpha
CARC_GT	1.53	0.013334	DUSP1	dual specificity phosphatase 1
CARC_GT	1.64	0.01764	EGR1	early growth response 1
CARC_GT	1.63	0.02163	TSKU	tsukushi small leucine rich proteoglycan homolog (Xenopus laevis)
CARC_GT	1.35	0.022342	CCND1	cyclin D1
CARC_GT	1.78	0.022968	CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5
CARC_GT	1.38	0.024595	PPP1R3C	protein phosphatase 1, regulatory subunit 3C
CARC_GT	1.58	0.03165	SLC6A6	solute carrier family 6 (neurotransmitter transporter, taurine), member 6
CARC_GT	1.51	0.033326	CDH17	cadherin 17, LI cadherin (liver-intestine)

CARC_GT	1.46	0.041069	ZNF354A	zinc finger protein 354A
CARC_GT	1.46	0.041205	KLF6	Kruppel-like factor 6
CARC_GT	1.43	0.043475	USP2	ubiquitin specific peptidase 2
CARC_NGT	0.56	0.049227	AQP3	aquaporin 3 (Gill blood group)
CARC_NGT	0.55	0.044921	HDC	histidine decarboxylase
CARC_NGT	0.66	0.039928	EPHX2	epoxide hydrolase 2, cytoplasmic
CARC_NGT	0.67	0.038885	PRLR	prolactin receptor
CARC_NGT	0.67	0.032504	ABHD1	abhydrolase domain containing 1
CARC_NGT	0.57	0.03165	CYP8B1	cytochrome P450, family 8, subfamily B, polypeptide 1
CARC_NGT	0.52	0.025887	QPCT	glutaminyl-peptide cyclotransferase
CARC_NGT	0.65	0.023538	CRAT	carnitine O-acetyltransferase
CARC_NGT	0.7	0.022888	DACT2	dapper, antagonist of beta-catenin, homolog 2 (Xenopus laevis)
CARC_NGT	0.59	0.020686	PDK4	pyruvate dehydrogenase kinase, isozyme 4
CARC_NGT	0.56	0.017979	ACOT1	acyl-CoA thioesterase 1
CARC_NGT	0.63	0.017556	PNPLA3	patatin-like phospholipase domain containing 3
CARC_NGT	0.55	0.016964	EHHADH	enoyl-CoA, hydratase/3-hydroxyacyl CoA dehydrogenase
CARC_NGT	0.64	0.01516	CIDEA	cell death-inducing DFFA-like effector a
CARC_NGT	0.63	0.015132	ECI1	enoyl-CoA delta isomerase 1
CARC_NGT	0.57	0.013334	AQP7	aquaporin 7
CARC_NGT	0.62	0.013334	HSD3B2	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2
CARC_NGT	0.57	0.013013	VNN1	vanin 1
CARC_NGT	0.64	0.012782	MYO5B	myosin VB
CARC_NGT	0.73	0.012188	DDHD1	DDHD domain containing 1
CARC_NGT	0.46	0.011624	CPT1B	carnitine palmitoyltransferase 1B (muscle)
CARC_NGT	0.63	0.011563	FADS2	fatty acid desaturase 2
CARC_NGT	0.68	0.011184	GALE	UDP-galactose-4-epimerase
CARC_NGT	0.69	0.010164	NUDT7	nudix (nucleoside diphosphate linked moiety X)-type motif 7
CARC_NGT	0.68	0.009674	ABHD3	abhydrolase domain containing 3
CARC_NGT	0.62	0.009674	ANGPTL4	angiopoietin-like 4
CARC_NGT	0.72	0.008501	TOR3A	torsin family 3, member A
CARC_NGT	0.71	0.00833	CYP2J2	cytochrome P450, family 2, subfamily J, polypeptide 2
CARC_NGT	0.72	0.008325	MIOX	myo-inositol oxygenase
CARC_NGT	0.67	0.008276	ACSM2A	acyl-CoA synthetase medium-chain family member 2A
CARC_NGT	0.66	0.008157	SLC25A30	solute carrier family 25, member 30
CARC_NGT	0.73	0.007972	ACOX1	acyl-CoA oxidase 1, palmitoyl
CARC_NGT	0.66	0.007972	G6PD	glucose-6-phosphate dehydrogenase
CARC_NGT	0.52	0.007972	PEX11A	peroxisomal biogenesis factor 11 alpha
CARC_NGT	0.62	0.006138	ECH1	enoyl CoA hydratase 1, peroxisomal
CARC_NGT	0.55	0.005281	CYP4A11	cytochrome P450, family 4, subfamily A, polypeptide 11
CARC_NGT	0.59	0.005185	ACSM5	acyl-CoA synthetase medium-chain family member 5
CARC_NGT	0.65	0.004088	C2orf88	chromosome 2 open reading frame 88
CARC_NGT	0.54	0.003376	ACOT1	acyl-CoA thioesterase 1
CARC_NGT	0.47	0.003363	AIG1	androgen-induced 1
CARC_NGT	0.16	0.001673	ACOT1	acyl-CoA thioesterase 1
CARC_NGT	0.72	0.001063	DECR1	2,4-dienoyl CoA reductase 1, mitochondrial
CARC_NGT	0.67	0.000773	IMPA2	inositol(myo)-1(or 4)-monophosphatase 2
CARC_NGT	0.73	0.00077	CLYBL	citrate lyase beta like
CARC_NGT	0.74	0.000511	SLC22A25	solute carrier family 22, member 25
CARC_NGT	0.55	0.000148	ME1	malic enzyme 1, NADP(+)-dependent, cytosolic

Table I.3 - Random forest with tissue agnostic labels

Random forest cross-validation results using tissue agnostic class labels for genotoxicity and carcinogenicity in liver and cell culture in the DrugMatrix. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	LIVER	CELL CULTURE	LIVER	CELL CULTURE
	GenTox	GenTox	Carcinogen	Carcinogen

AUC	**73.64 ± 1.0**	**79.56 ± 1.0**	**64.76 ± 1.0**	**63.35 ± 1.2**
ACC	75.3 ± 0.8	76.56 ± 0.8	61.27 ± 0.8	61.93 ± 1.0
SENS	41.76 ± 1.8	56.77 ± 2	72.43 ± 1.4	71.91 ± 1.6
SPEC	87.14 ± 0.8	86.72 ± 1.0	43.15 ± 2.0	45.24 ± 2.5
PPV	52.65 ± 2.2	67.39 ± 2.2	70.31 ± 1.2	71.11 ± 1.6
NPV	81.45 ± 1.0	80.62 ± 1.2	45.6 ± 1.6	46.2 ± 1.8
FDR	47.35 ± 2.2	32.61 ± 2.2	29.69 ± 1.2	28.89 ± 1.6

Table I.4 - Prediction using tissue-specific labels

Random forest cross-validation results for genotoxicity and carcinogenicity in liver and cell culture in the DrugMatrix. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	LIVER GenoToxicity	LIVER Carcinogenicity
#Samples	1260	1221
#Chemicals	130	127
AUC	**75.08 ± 1.2**	**76.73 ± 1.0**
ACC	75.62 ± 0.8	72.95 ± 0.8
SENS	42.82 ± 2.2	56.78 ± 1.8
SPEC	87.25 ± 0.8	82.91 ± 1.0
PPV	52.79 ± 2.4	66.61 ± 1.8
NPV	81.88 ± 1.0	76.37 ± 1.2
FDR	47.21 ± 2.4	33.39 ± 1.8

Table I.5 – Prediction with tissue specific labels using SVM

Support Vector Machine (SVM) cross-validation results for genotoxicity and carcinogenicity in liver and cell culture in the DrugMatrix. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	LIVER GenTox	CELL CULTURE GenTox	LIVER Carcinogen_liv	CELL CULTURE Carcinogen_liv
AUC	**65.63 ± 4.3**	**75.15 ± 5.5**	**61.31 ± 4.1**	**56.4 ± 7.3**
ACC	73.05 ± 3.3	78.83 ± 4.7	63.94 ± 3.7	65.99 ± 5.5
SENS	49.42 ± 8.8	63.24 ± 11.6	50.16 ± 8.4	35.14 ± 14.9
SPEC	81.83 ± 4.1	87.06 ± 6.3	72.46 ± 5.9	77.65 ± 6.5
PPV	48.3 ± 10.6	70.34 ± 12.5	50.6 ± 9.6	35.4 ± 13.1
NPV	82.15 ± 5.5	83.07 ± 6.1	71.97 ± 6.5	76.97 ± 7.1

FDR	51.7 ± 10.8	29.66 ± 12.5	49.4 ± 9.6	64.6 ± 13.1

Table I.6 - Prediction with tissue specific labels using PAMR

Shrunken centroid (PAMR) cross-validation results for genotoxicity and carcinogenicity in liver and cell culture in the DrugMatrix. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	LIVER GenTox	CELL CULTURE GenTox	LIVER Carcinogen_liv	CELL CULTURE Carcinogen_liv
AUC	**70.36 ± 1.0**	**76.73 ± 1.2**	**77.3 ± 0.8**	**58.79 ± 1.8**
ACC	73.22 ± 0.8	75.69 ± 1.0	72.66 ± 0.8	66.59 ± 1.2
SENS	16.11 ± 1.4	47.36 ± 2.2	53.29 ± 1.6	21.53 ± 2.2
SPEC	93.87 ± 0.8	90.64 ± 1.4	84.4 ± 0.8	86.76 ± 1.4
PPV	53.02 ± 3.3	74.97 ± 2.7	66.45 ± 1.8	43.97 ± 3.7
NPV	76.03 ± 1.0	77.67 ± 1.2	75.31 ± 1.2	71.93 ± 1.4
FDR	46.98 ± 3.3	25.03 ± 2.7	33.55 ± 1.8	56.03 ± 3.7

Table I.7 - Prediction with tissue specific labels using structural features alone

Random forest cross-validation results for genotoxicity and carcinogenicity in liver and cell culture in the DrugMatrix based on structural features. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	LIVER GenTox	CELL CULTURE GenTox	LIVER Carcinogen_liv	CELL CULTURE Carcinogen_liv
AUC	**70.94 ± 4.1**	**85.59 ± 2.2**	**59.89 ± 8.8**	**54.68 ± 9.2**
ACC	82.33 ± 2.2	73.09 ± 14.1	56.72 ± 3.1	58.65 ± 5.1
SENS	44.91 ± 12.7	93.75 ± 12.2	30.58 ± 9.4	25 ± 29.4
SPEC	96.4 ± 2.9	68.72 ± 16.1	73.63 ± 16.5	76.84 ± 32.9
PPV	83.01 ± 9.2	46.1 ± 34.3	42.7 ± 16.5	35 ± 29.4
NPV	82.42 ± 4.1	96.51 ± 6.9	63.28 ± 5.9	71.12 ± 17.4
FDR	16.99 ± 9.2	53.9 ± 34.3	57.3 ± 16.5	65 ± 29.4

Table I.8 - Prediction with tissue specific labels using gene expression and structural features

Random forest cross-validation results for genotoxicity and carcinogenicity in liver and cell culture in the DrugMatrix based on structural features and gene expression profiles. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	LIVER GenTox	CELL CULTURE GenTox	LIVER Carcinogen_liv	CELL CULTURE Carcinogen_liv
AUC	**80.11 ± 1.8**	**79.76 ± 1.8**	**77.74 ± 1.4**	**65.22 ± 2.2**

ACC	81.39 ± 1.2	75.7 ± 1.4	72.61 ± 1.0	68.08 ± 1.4
SENS	53.33 ± 3.5	59.78 ± 3.1	59.63 ± 2.7	29.62 ± 3.3
SPEC	91.05 ± 1.2	84.13 ± 1.8	81.4 ± 1.6	84.87 ± 2.0
PPV	67.37 ± 3.3	64.12 ± 3.5	66.12 ± 2.5	45.93 ± 4.5
NPV	85.16 ± 1.4	81.75 ± 1.8	76.84 ± 1.8	74.18 ± 1.8
FDR	32.63 ± 3.3	35.88 ± 3.5	33.88 ± 2.5	54.07 ± 4.5

Table I.9: Prediction results on TG-GATEs of a model trained on the DrugMatrix

Random forest classification results including the 95% confidence interval for carcinogenicity in liver, based on genes and pathways. The model was trained on the DrugMatrix and tested on TG-GATEs.

	Genes	Pathways
#Samples	2064	2064
#Chemicals	47	47
AUC	**76.64 ± 1.8**	**78.50 ± 1.8**
ACC	81.62 ± 1.8	80.56 ± 1.8
SENS	37.36 ± 2.2	48.48 ± 2.2
SPEC	98.25 ± 0.6	92.57 ± 1.2
PPV	88.89 ± 1.4	70.97 ± 2.0
NPV	80.68 ± 1.8	82.75 ± 1.6
FDR	11.11 ± 1.4	29.03 ± 2.0

Table I.10 - Cross-validation results in the TG-GATEs dataset

Random forest cross-validation results for carcinogenicity in liver, based on genes and pathways in the TG-GATEs dataset. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	Genes	Pathways
AUC	**82.67 ± 1.0**	**80.6 ± 0.8**
ACC	80.07 ± 0.8	78.99 ± 0.6
SENS	63.35 ± 1.8	56.72 ± 1.6
SPEC	90.22 ± 0.8	91.75 ± 0.6
PPV	78.88 ± 1.6	78.93 ± 1.4
NPV	80.99 ± 1.0	79 ± 1.0
FDR	21.12 ± 1.6	21.07 ± 1.4

Table I.11 – Classification performance with and without dose specific annotation

Random forest cross-validation results for carcinogenicity in liver, based on genes and pathways in the TG-GATEs dataset. Classification results of both dose-specific and -unspecific carcinogenicity labels are included. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	Dose dependent		Dose Independent
AUC	82.67 ± 1.0		69.26 +/- 0.9
ACC	80.07 ± 0.8		80.9 +/- 0.4
SENS	63.35 ± 1.8		31.97 +/- 1.3
SPEC	90.22 ± 0.8		93.06 +/- 0.6
PPV	78.88 ± 1.6		52.26 +/- 1.5
NPV	80.99 ± 1.0		84.99 +/- 0.5
FDR	21.12 ± 1.6		47.74 +/- 1.5

Table I.12 - Gene set projection of the DrugMatrix samples

Random forest cross-validation results for tissue genotoxicity and carcinogenicity in liver and cell culture based on pathway projected profiles in the DrugMatrix. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split.

	LIVER	CELL CULTURE	LIVER	CELL CULTURE
	GenTox	GenTox	Carcinogen_liv	Carcinogen_liv
AUC	68.32 ± 1.0	79.86 ± 1.2	73.27 ± 0.8	64.87 ± 1.4
ACC	72.62 ± 0.8	78.36 ± 1.0	71.52 ± 0.7	66.19 ± 1.0
SENS	27.54 ± 1.7	59.2 ± 2.0	51.96 ± 1.6	38.11 ± 2.5
SPEC	88.9 ± 0.8	88.53 ± 1.2	83.91 ± 0.9	78.82 ± 1.3
PPV	46.76 ± 2.1	72.22 ± 2.3	66.33 ± 1.8	43.06 ± 2.4
NPV	77.95 ± 1.1	81.51 ± 1.2	74.08 ± 1.1	75.23 ± 1.3
FDR	53.24 ± 2.1	27.78 ± 2.3	33.67 ± 1.8	56.94 ± 2.4

Table I.13 – Comparison with published signatures

Comparison of classification results for tissue carcinogenicity in liver. The random forest model is compared to two published signatures that were tested with a support vector machine. The first three columns show models trained on the DrugMatrix and tested on TG-GATEs, while the fourth columns shows the mean over 200 iterations of a 70%/30% train/test dataset split of the non-genotoxic compounds in the DrugMatrix.

	Random Forest	**Ellinger-Ziegelbauer 2008**	**Fielden 2011**	**Fielden 2011 (Non-GT)**
AUC	76.64	61.75	69.56	62.59 ± 0.6
ACC	81.62	71.57	83.05	66.16 ± 1.0
SENS	37.36	40.11	39.84	37.76 ± 2.0
SPEC	98.25	83.38	99.28	87.42 ± 1.1
PPV	88.89	47.56	95.39	67.49 ± 2.3
NPV	80.68	78.75	81.46	67.99 ± 1.5
FDR	11.11	52.44	4.61	32.51 ± 2.2

Table I.14 - Testing different numbers of features using a variance filter

Random forest cross-validation results for carcinogenicity in liver using different numbers of features, based on variance ranking. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split in the DrugMatrix.

	200 Genes	500 Genes	1000 Genes	2000 Genes
AUC	**76 ± 0.8**	**76.1 ± 0.8**	**75.50 ± 0.8**	**75.8 ± 1.0**
ACC	72 ± 0.8	72.8 ± 0.8	72.50 ± 0.8	72.5 ± 0.8
SENS	52.2 ± 1.8	52.1 ± 1.8	51.30 ± 1.6	54.1 ± 1.8
SPEC	83.8 ± 1.2	85.00 ± 1.0	84.90 ± 1.0	83.4 ± 1.2
PPV	64.1 ± 2.0	66.00 ± 2.0	66.40 ± 1.8	64.3 ± 2.0
NPV	76.2 ± 1.2	75.60 ± 1.2	75.40 ± 1.2	76.8 ± 1.2
FDR	35.9 ± 2.0	34.00 ± 2.0	33.60 ± 1.8	35.6 ± 2.0

Table I.15– Testing different numbers of features using differential expression

Random forest cross-validation results for carcinogenicity in liver using different numbers of features, based on differential expression ranking. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split in the DrugMatrix.

	200 Genes	500 Genes	1000 Genes	2000 Genes
AUC	**75.2 ± 0.8**	**75.4 ± 0.8**	**74.8 ± 0.8**	**74.3 ± 0.8**
ACC	73 ± 0.8	73.1 ± 0.8	72.3 ± 0.8	72 ± 0.8
SENS	51.3 ± 1.8	53.6 ± 1.6	50.4 ± 1.8	48.5 ± 1.8
SPEC	85.3 ± 0.8	84.2 ± 1.0	85.1 ± 1.0	86 ± 1.0
PPV	65.7 ± 1.8	64.8 ± 1.8	65.6 ± 1.8	66.3 ± 1.8
NPV	76 ± 1.2	77 ± 1.0	75.3 ± 1.2	74.5 ± 1.2
FDR	34.3 ± 1.8	35.2 ± 1.8	34.4 ± 1.8	33.7 ± 1.8

Table I.16 – Prediction results with lower variance features

Random forest cross-validation results for carcinogenicity in liver using 500 features with decreasing variance. Each value represents the mean and 95% confidence interval over 200 iterations of a 70%/30% train/test dataset split in the DrugMatrix.

	Features 1-500	Features 501-1000	Features 1001-1500
AUC	**77.74 ± 1.4**	**75.16 ± 0.8**	**74.58 ± 0.8**
ACC	72.61 ± 1.0	72.09 ± 0.8	71.95 ± 0.8
SENS	59.63 ± 2.7	53.16 ± 1.8	53.04 ± 1.8
SPEC	81.4 ± 1.6	83.63 ± 1.0	83.7 ± 1.0
PPV	66.12 ± 2.5	65.29 ± 2.0	66.17 ± 1.8
NPV	76.84 ± 1.8	75.55 ± 1.2	75.09 ± 1.2
FDR	33.88 ± 2.5	34.71 ± 2.0	33.83 ± 1.8

Table I.17 – Samples in Drugmatrix with carcinogenicity annotation

Overview of samples in the DrugMatrix with either carcinogenicity or genotoxicity annotation, according to tissue type.

	LIVER	CELL CULTURE	KIDNEY	HEART	THIGH MUSCLE	All
All samples	2195	813	1410	862	158	5438
Untreated	279	113	335	231	36	994
Treated	1916	700	1075	631	122	4444
Non-Genotoxic	942	362	463	339	77	2183
Genotoxic	318	171	245	125	77	936
Non-Carcinogen	765	341	51	/	/	1157
Carcinogen	456	141	51	/	/	648
Compounds	199	104	139	88	21	551

Table I.18 – Samples in TG-GATEs

Overview of samples in the TG-GATEs with either carcinogenicity or genotoxicity annotation, according to tissue type.

	Liver			Kidney	
	single	repeat	*in-vitro*	single	Repeat
All samples	6264	6249	3140	1872	1856
Untreated	1572	1572	768	468	468
# Compounds	131	131	131	39	39

Table I.19 – Overlapping compounds between TG-GATEs and DrugMatrix

Overview of 25 compounds that were both tested in the TG-GATEs and DrugMatrix, showing the differences in treatment doses.

	TG-GATEs doses (mg/kg)				DrugMatrix doses (mg/kg)			
acetaminophen	300	600	1000		100	-	-	-
allyl alcohol	3	10	30		16	25	32	-
aspirin	45	150	450		35	167	375	-
carbamazepine	30	100	300		490	-	-	-
carbon tetrachloride	30	100	300		400	1175	-	-
clofibrate	30	100	300		130	500	-	-
clomipramine	10	30	100		115	-	-	-
diazepam	25	75	250		710	-	-	-

diclofenac	1	3	10		10	-	-	-
ethanol	400	1200	4000		6000	-	-	-
fenofibrate	10	100	1000		43	100	215	430
gemfibrozil	30	100	300		100	700	-	-
indomethacin	0.5	1.6	5		12	-	-	-
ketoconazole	10	30	100		114	227	-	-
meloxicam	3	10	30		33	-	-	-
methapyrilene	10	30	100		100	-	-	-
methimazole	10	30	100		100	-	-	-
naproxen	6	20	60		10	-	-	-
phenobarbital	10	30	100		25	54	-	-
promethazine	20	60	200		2.3	113	-	-
propylthiouracil	10	30	100		625		-	-
simvastatin	40	120	400		15	1200	-	-
tamoxifen	6	20	60		2.5	64	-	-
thioacetamide	4.5	15	45		200		-	-
valproic acid	45	150	450		1340	1500	-	-

Table I.20 – Performance measurements

Equations to calculate the performance measurements. True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)

Accuracy	(TP+TN)/(TP+TN+FP+FN)
Sensitivity	TP / (TP+FN)
Specificity	TN / (TN + FP)
Positive Predictive Value (PPV)	TP / (TP+FP)
Negative Predictive Value (NPV)	TN / (TN + FN)
False Discovery Rate (FDR)	FP/ (TP+FP)

APPENDIX II

Figure II.1 - Schematic depicting target gene selection.

(i) TCF3 genes were derived from Schmitz et al1 and then validated in silico by differential analysis against GEPs of BL and DLBCL from 2 prior publications2,3. Transcripts with "false discovery rates" (FDR) <0.25 were ranked by signal to noise ratio and the top 37 within the union of the two analyses were selected for the initial profiling panel. Analysis of expression data from the training cohort and the construction of molecular classifiers resulted in the inclusion of 7 gene targets in the final probe set that was tested in the test cohort. (ii) MYC biological activity signature was derived from the differential analysis of 457 published MYC targets4-10 against the global GEP of frozen DLBCLs with corresponding MYC IHC class (MYC IHC high versus MYC IHC low) from the training cohort. DA analysis of the entire 18,400 transcripts within the GEP of the frozen tissue with respect to the corresponding MYC IHC class was also performed.

Supplementary Figure 1

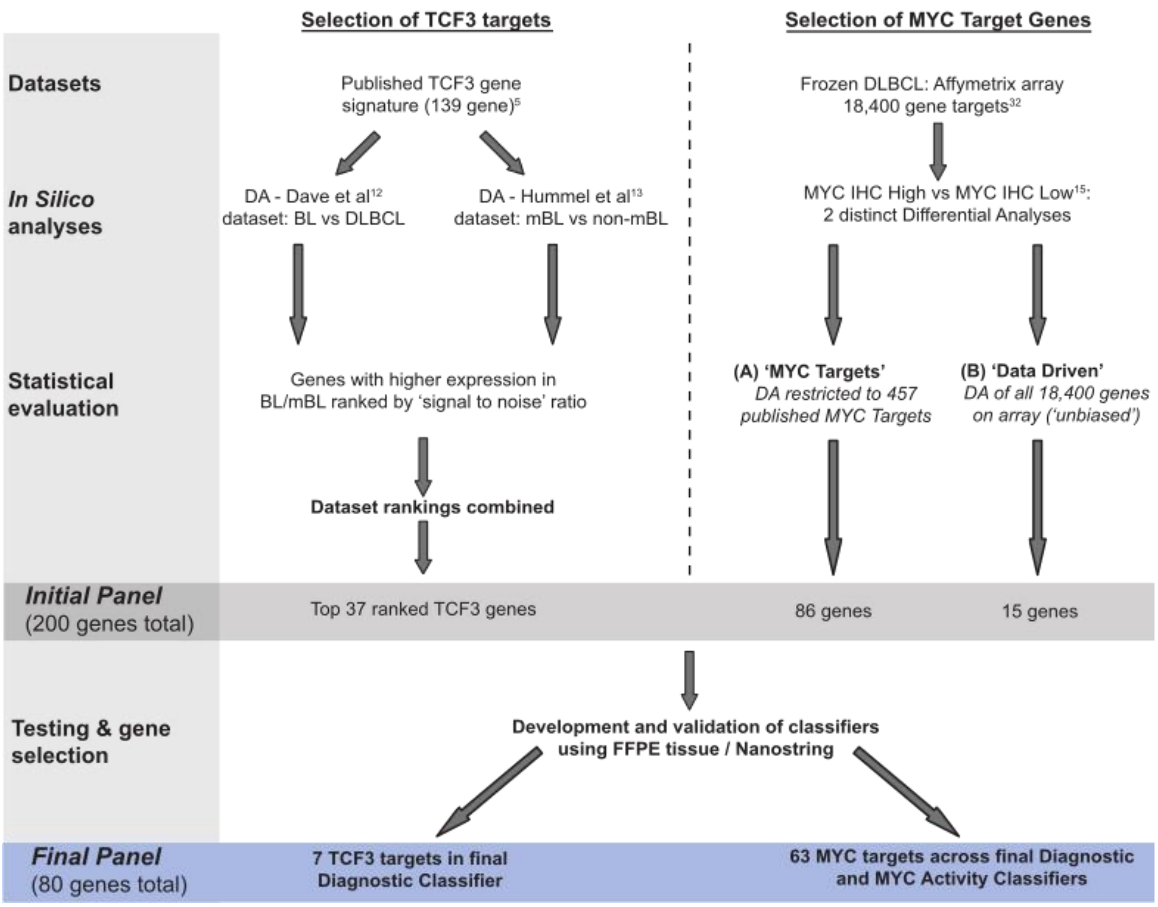

Selection of TCF3 targets
Selection of MYC Target Genes
Datasets
Published TCF3 gene signature (139 gene)[5]
Frozen DLBCL: Affymetrix array 18,400 gene targets[32]
In Silico analyses
DA - Dave et al[12] dataset: BL vs DLBCL
DA - Hummel et al[13] dataset: mBL vs non-mBL
MYC IHC High vs MYC IHC Low[15]: 2 distinct Differential Analyses
Statistical evaluation
Genes with higher expression in BL/mBL ranked by 'signal to noise' ratio
(A) 'MYC Targets' DA restricted to 457 published MYC Targets
(B) 'Data Driven' DA of all 18,400 genes on array ('unbiased')
Dataset rankings combined
Initial Panel (200 genes total)
Top 37 ranked TCF3 genes
86 genes
15 genes
Testing & gene selection
Development and validation of classifiers using FFPE tissue / Nanostring
Final Panel (80 genes total)
7 TCF3 targets in final Diagnostic Classifier
63 MYC targets across final Diagnostic and MYC Activity Classifiers

Figure II.2 - Unsupervised hierarchical clustering of data

for 3 tumors tested on more than one occasion during the test study. Heatmap data are normalized to the 6 housekeeping genes but are not normalized between 'profiling panel builds'. The 'profiling panel build' and experiment number (first line), the relative expression of the transcripts used in the final profiling panel are shown (heatmap, housekeeping gene data not shown). *The mean MYC activity score (third line) and bar charts of respective MYC activity scores by build and experiment number (fourth line) use the final classifier output, following normalization of data between profiling panel builds.

Supplementary Figure 2

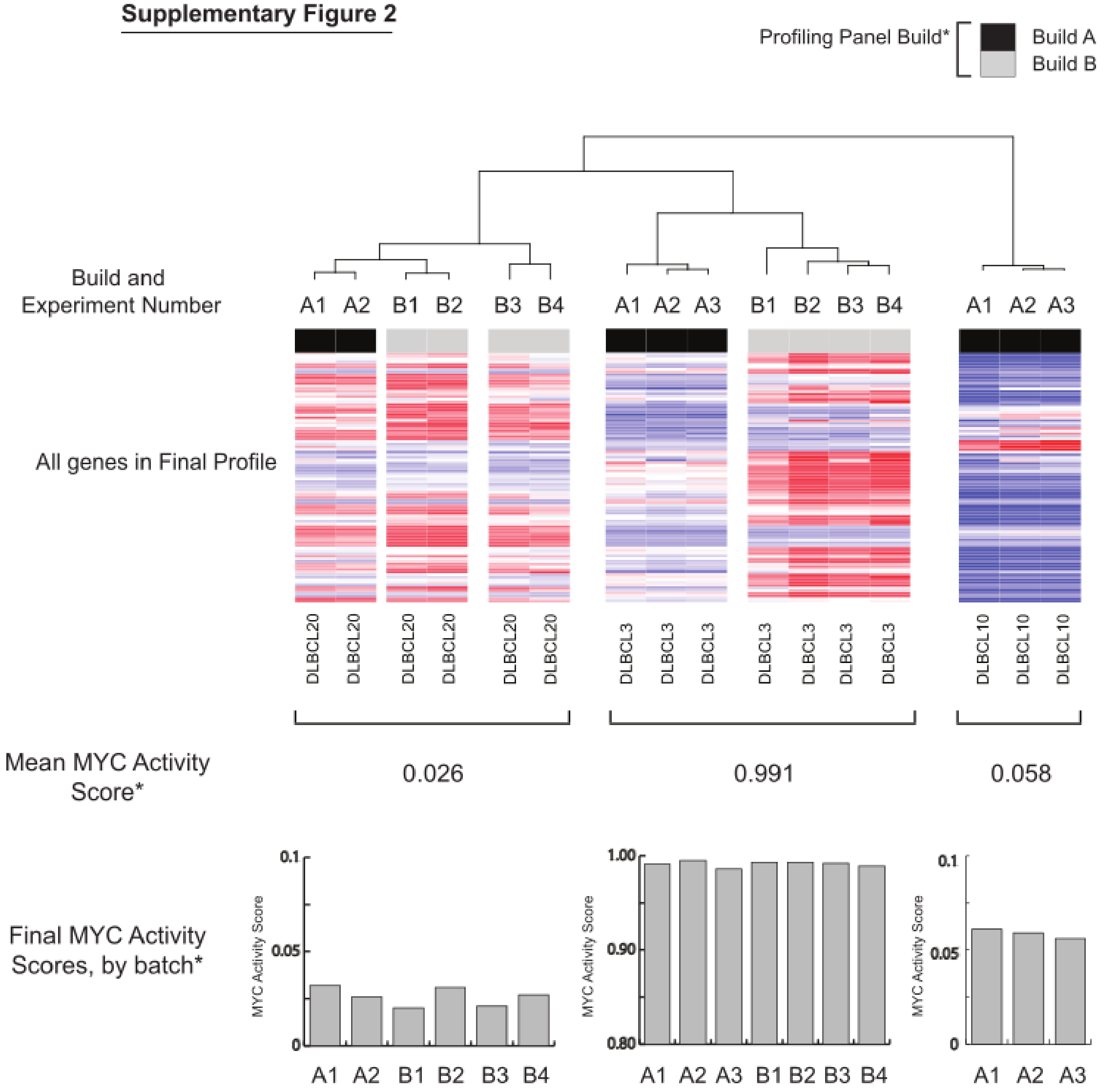
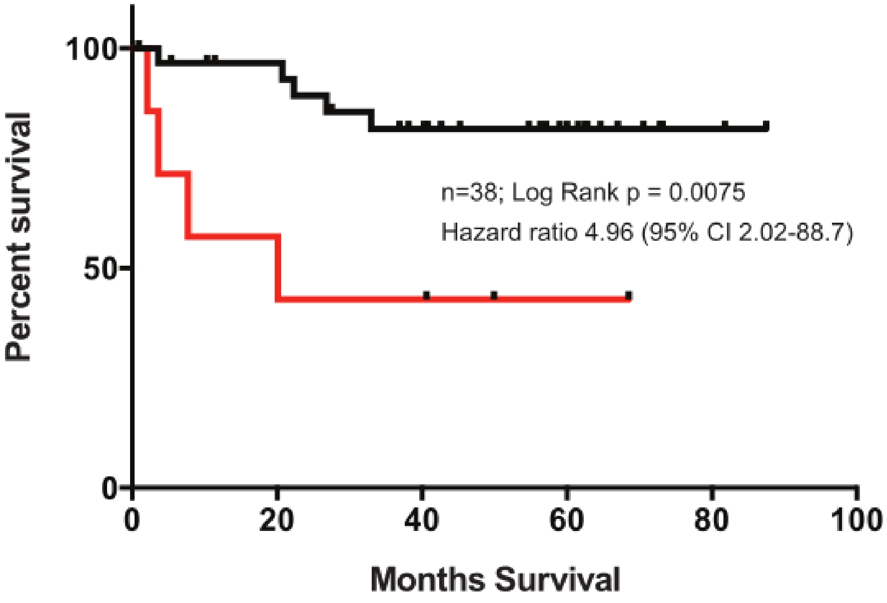


Figure II.3 - Kaplan-Meier (KM) curves showing Overall Survival (OS) for the outcome series where matched MYC IHC and MYC Activity score data are available. Two cases lacked MYC IHC score therefore n=38 rather than n=40 in Figure 6. (A) Segregated by MYC IHC: MYC IHC-High >50% (red line) and MYC IHC-Low ≤50% (black line). (B) Segregated by MYC Activity Score: MYC Activity High (>0.5, red line) and MYC Activity Low (<0.5, black line).

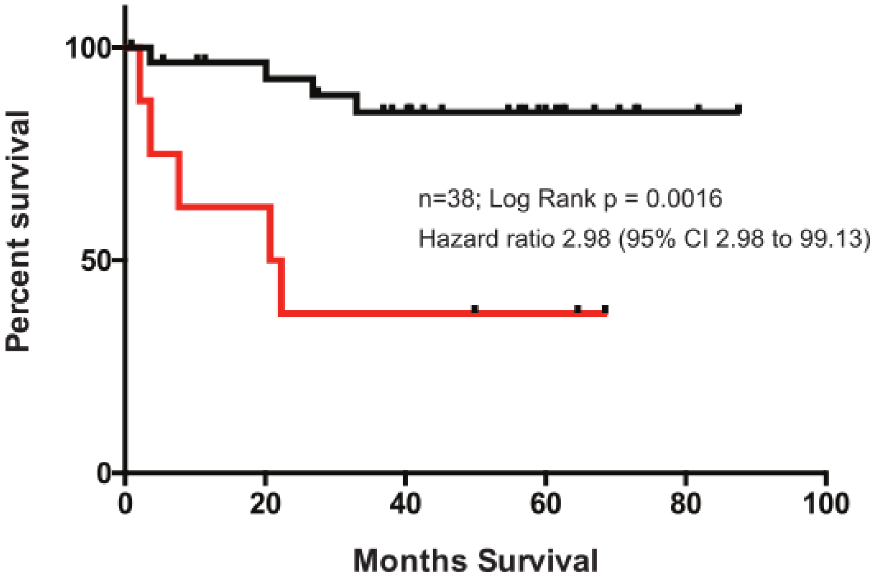
Supplementary Figure 3

MYC IHC



n=38; Log Rank p = 0.0075
Hazard ratio 4.96 (95% CI 2.02-88.7)

MYC Activity Score



n=38; Log Rank p = 0.0016
Hazard ratio 2.98 (95% CI 2.98 to 99.13)

Table II.1 Clinical details of the training and test cohorts are shown.

Tumors with no recorded values for the diagnostic and MYC activity classifiers failed analytical quality control.

Training Cohort: Burkitt Lymphoma

Case	Figure Code	Diagnosis	Age of Biopsy (years)	Diagnostic Score	MYC Activity Score	MYC IHC (%)	MYC Rearrangement*
1	BL1	Burkitt Lymphoma	6	0.999	0.945	80	NA
2	BL2	Burkitt Lymphoma	12	0.996	1	60	1
3	BL3	Burkitt Lymphoma	6	0.996	0.998	NA	1
4	BL4	Burkitt Lymphoma	6	0.996	0.994	60	NA
5	BL5	Burkitt Lymphoma	9	0.893	0.982	70	1
6	BL6	Burkitt Lymphoma	9	0.85	0.996	80	1
7	BL7	Burkitt Lymphoma	5	0.845	0.997	100	1
8	BL8	Burkitt Lymphoma	4	0.841	0.994	100	1
9	BL9	Burkitt Lymphoma	9	0.829	1	100	1
10	BL10	Burkitt Lymphoma	9	0.632	0.629	80	1
11	BL11	Burkitt Lymphoma	6	0.551	0.999	100	1
12	BL12	Burkitt Lymphoma	6	0.306	0.952	70	1

Training Cohort: DLBCL

Case	Figure Code	Diagnosis	Age of Patient (years)	Diagnostic Score	MYC Activity Score	MYC IHC (%)	BCL2 IHC (%)	MYC Rearrangement*
1	DLBCL1	DLBCL-NOS	64	0.527	0.91	80	10	1
2	DLBCL2	DLBCL-NOS	57	0.348	1	90	100	1
3	DLBCL3	DLBCL-NOS	49	0.305	1	90	60	1
4	DLBCL4	DLBCL-NOS	83	0.207	0.58	70	0	1
5	DLBCL5	DLBCL-NOS	49	0.164	0.84	70	0	0
6	DLBCL6	DLBCL-NOS	70	0.122	0.76	60	100	0
7	DLBCL7	DLBCL-NOS	66	0.113	0.49	70	0	0
8	DLBCL8	DLBCL-NOS	65	0.107	0.84	70	10	1
9	DLBCL9	DLBCL-NOS	-	0.08	0.22	30	0	0
10	DLBCL10	DLBCL-NOS	63	0.08	0.11	30	0	0
11	DLBCL11	DLBCL-NOS	75	0.065	0.98	70	100	0
12	DLBCL12	DLBCL-NOS	68	0.05	0.02	10	100	0
13	DLBCL13	DLBCL-NOS	46	0.046	0.42	20	10	0
14	DLBCL14	DLBCL-NOS	72	0.027	0.46	70	100	0

15	DLBCL15	DLBCL-NOS	40	0.022	0.46	40	100	0
16	DLBCL16	DLBCL-NOS	81	0.021	0.88	60	70	1
17	DLBCL17	DLBCL-NOS	-	0.019	0.12	30	40	0
18	DLBCL18	DLBCL-NOS	69	0.018	0.19	10	90	0
19	DLBCL19	DLBCL-NOS	60	0.016	0.01	20	70	0
20	DLBCL20	DLBCL-NOS	76	0.014	0.02	20	100	0
21	DLBCL21	DLBCL-NOS	75	0.013	0.98	30	100	0
22	DLBCL22	DLBCL-NOS	62	0.013	0.95	80	90	1
23	DLBCL23	DLBCL-NOS	45	0.011	0.13	30	0	0
24	DLBCL24	DLBCL-NOS	55	0.011	0.63	90		0
25	DLBCL25	DLBCL-NOS	38	0.008	0.1	20	70	0
26	DLBCL26	DLBCL-NOS	30	0.006	0.01	20	0	0
27	DLBCL27	DLBCL-NOS	78	0.005	0.01	20	100	0
28	DLBCL28	DLBCL-NOS	67	0.003	0.02	30	10	0
29	DLBCL29	DLBCL-NOS	57	0.001	0	30	0	0
30	DLBCL30	DLBCL-NOS	68	-	-	80		0

Test Cohort: Burkitt Lymphoma

Case	Figure Code	Diagnosis	Age	Diagnostic Score	MYC Activity Score	MYC IHC (%)	MYC Rearrangement*
1	tBL1	Burkitt Lymphoma	3	0.946	0.996	80	1
2	tBL2	Burkitt Lymphoma (Atypical)	10	0.883	0.996	0	1
3	tBL3	Burkitt Lymphoma	0.5	0.877	0.955	0	1
4	tBL4	Burkitt Lymphoma	11	0.845	0.862	90	1
5	tBL5	Burkitt Lymphoma	7	0.811	0.919	60	1
6	tBL6	Burkitt Lymphoma	10	0.802	0.963	90	1
7	tBL7	Burkitt Lymphoma	10	0.766	0.969	10	1
8	tBL8	Burkitt Lymphoma	7	0.682	0.991	80	NA
9	tBL9	Burkitt Lymphoma	9	0.641	0.975	90	1
10	tBL10	Burkitt Lymphoma	1	-	-	100	1
11	tBL11	Burkitt Lymphoma	1	-	-	100	1
12	tBL12	Burkitt Lymphoma	2	-	-	90	1

Test Cohort: Genetic Double Hit Lymphoma

Case	Figure Code	Diagnosis	Age	Diagnostic Score	MYC Activity Score	MYC IHC (%)	BCL2 IHC (%)	MYC-R*	BCL2-R*	BCL6-R*
1	tDHL1	BCL-U	55	0.901	0.981	100	0	1	1	0
2	tDHL2	BCL-U	86	0.849	1	90	0	1	0	1
3	tDHL3	BCL-U	64	0.773	0.986	100	100	1	1	0
4	tDHL4	BCL-U	40	0.305	0.983	80	50	1	1	0
5	tDHL5	BCL-U	68	0.119	0.598	60	100	1	1	0
6	tDHL6	DLBCL	46	0.046	0.997	80	0	1	1	0
7	tDHL7	DLBCL	41	0.02	0.259	70	100	1	1	0
8	tDHL8	DLBCL	62	0.015	0.177	70	100	1	1	0

Test Cohort: DLBCL

Case	Figure Code	Diagnosis	Age	Diagnostic Score	MYC Activity Score	MYC IHC (%)	BCL2 IHC (%)	MYC-R*	BCL2-R*	BCL6-R*
1	tDLBCL1	DLBCL, NOS	43	0.501	0.629	70	0	1	0	0
2	tDLBCL2	DLBCL, NOS	76	0.332	0.122	90	80	0	0	NA
3	tDLBCL3	DLBCL, NOS	58	0.167	0.638	90	100	0	NA	NA
4	tDLBCL4	DLBCL-NOS	72	0.15	0.51	40	0	0	0	0
5	tDLBCL5	DLBCL-NOS	65	0.09	0.85	50	0	0	0	0
6	tDLBCL6	DLBCL, Immunoblastic	91	0.086	0.661	80	100	0	0	0
7	tDLBCL7	DLBCL-NOS	80	0.06	0.38	40	10	0	0	0
8	tDLBCL8	DLBCL, NOS	71	0.062	0.35	60	100	0	0	0
9	tDLBCL9	DLBCL, NOS	44	0.055	0.816	90	100	0	0	NA
10	tDLBCL10	DLBCL, HIV+	41	0.052	0.076	10	0	0	0	0
11	tDLBCL11	DLBCL, NOS	78	0.049	0.06	20	60	0	0	0
12	tDLBCL12	DLBCL-NOS	53	0.05	0.18	30	20	0	NA	NA
13	tDLBCL13	DLBCL, NOS	54	0.046	0.024	60	10	0	0	0
14	tDLBCL14	DLBCL-NOS	80	0.04	0.27	50	0	0	0	0
15	tDLBCL15	DLBCL, NOS	46	0.04	0.517	40	10	0	0	0

Case	Figure Code	Diagnosis	Age	Diagnostic Score	MYC Activity Score	MYC IHC (%)	BCL2 IHC (%)	MYC-R*	BCL2-R*	BCL6-R*
16	tDLBCL16	DLBCL, NOS	68	0.035	0.594	30	90	0	NA	NA
17	tDLBCL17	DLBCL, NOS	73	0.034	0.047	40	80	0	1	0
18	tDLBCL18	DLBCL, NOS	44	0.031	0.004	10	0	0	0	0
19	tDLBCL19	DLBCL-NOS	77	0.03	0.09	20	100	0	NA	NA
20	tDLBCL20	DLBCL, NOS	62	0.026	0.032	20	100	0	0	0
21	tDLBCL21	DLBCL, NOS	76	0.024	0.121	20	0	0	0	1
22	tDLBCL22	DLBCL-NOS	33	0.02	0.3	20	0	0	0	1
23	tDLBCL23	DLBCL-NOS	78	0.02	0.46	40	30	0	0	0
24	tDLBCL24	DLBCL, NOS	54	0.016	0.983	70	100	0	0	
25	tDLBCL25	DLBCL-NOS	72	0.02	0.02	40	90	0	NA	NA
26	tDLBCL26	DLBCL-NOS	27	0.02	0.04	40	80	0	0	0
27	tDLBCL27	DLBCL, NOS	58	0.015	0.378	30	90	0	0	1
28	tDLBCL28	DLBCL-NOS	69	0.01	0.02	50	0	0	NA	NA
29	tDLBCL29	DLBCL-NOS	46	0.01	0.04	30	50	0	NA	NA
30	tDLBCL30	DLBCL, NOS	49	0.012	0.038	10	90	0	0	0
31	tDLBCL31	DLBCL-NOS	32	0.01	0	10	100	0	0	0
32	tDLBCL32	DLBCL, NOS	60	0.008	0.009	30	30	0	0	1
33	tDLBCL33	DLBCL-NOS	68	0.01	0.03	40	30	0	NA	NA
34	tDLBCL34	DLBCL-NOS	69	0.01	0	-	90	0	0	0
35	tDLBCL35	DLBCL-NOS	40	0	0.003	40	0	0	0	0
36	tDLBCL36	DLBCL-NOS	59	0	0.27	50	90	0	0	0
37	tDLBCL37	DLBCL-NOS	62	0	0.1	40	0	0	0	0
38	tDLBCL38	DLBCL-NOS	79	0	0	-	0	0	0	0
39	tDLBCL39	DLBCL, NOS	72	-	-	-	80	0	0	0

Table II.2 - 200-gene Initial Profiling panel.

The 200 genes targets used in the initial profiling panel are listed and are organized into groups as derived (see Supplementary Fig.1 and Supplementary methods). 'Data driven' targets are genes that were not previously published as MYC targets but that were differentially expressed in the training set in either MYC IHC-High or MYC IHC-Low DLBCL.

Published MYC Targets

	Gene Symbol	Accession Number	Target Region (base pairs)
1	ACHY	NM_000687.2	1805-1905
2	AKAP1	NM_139275.1	2725-2825
3	AMD1	NM_001634.4	810-910
4	APEX1	NM_001641.2	727-827
5	APITD1	NM_199294.2	950-1050
6	AURKA	NM_003600.2	405-505
7	BUB1B	NM_001211.4	835-935
8	FAM216A / C12ORF24	NM_013300.2	722-822
9	CCNB1	NM_031966.2	715-815
10	CDC25A	NM_001789.2	690-790
11	CDK4	NM_000075.2	1055-1155
12	CHN1	NM_001025201.2	1965-2065
13	CIRH1A	NM_032830.2	84-184
14	CTPS	NM_001905.2	2570-2670
15	CYCS	NM_018947.4	1735-1835
16	DDX21	NM_004728.2	685-785
17	DDX47	NM_016355.3	1180-1280
18	DHX33	NM_001199699.1	2873-2973
19	DKC1	NM_001363.3	2255-2355
20	DLEU1	NR_002605.1	173-273
21	EIF1AX	NM_001412.3	3818-3918
22	ETFA	NM_001127716.1	630-730
23	EXOSC8	NM_181503.1	655-755
24	FBL	NM_001436.3	883-983
25	FKBP4	NM_002014.3	2755-2855
26	FXN	NM_001161706.1	515-615
27	GEMIN4	NM_015721.2	1925-2025
28	GEMIN5	NM_015465.3	4760-4860
29	GINS2	NM_016095.2	990-1090
30	GOT2	NM_002080.2	2145-2245
31	GPD1L	NM_015141.2	2565-2665
32	HSPE1	NM_002157.2	65-165
33	IDH3A	NM_005530.2	1521-1621
34	IMPA2	NM_014214.1	545-645
35	KIAA0101	NM_014736.4	65-165
36	LDHB	NM_001174097.1	1190-1290
37	LMNB2	NM_032737.2	3630-3730
38	LRPPRC	NM_133259.3	6220-6320
39	LSM7	NM_016199.2	150-250
40	LYAR	NM_001145725.1	230-330
41	MCC	NM_001085377.1	5578-5678
42	MDM1	NM_017440.2	1360-1460
43	MGST1	NM_145764.1	330-430

	Gene Symbol	Accession Number	Target Region (base pairs)
44	MKI67IP	NM_032390.4	215-315
45	MRPS2	NR_051968.1	1512-1612
46	MRPS34	NM_023936.1	719-819
47	MYB	NM_005375.2	3145-3245
48	MYC	NM_002467.3	1610-1710
49	NAP1L1	NM_004537.4	543-643
50	NME1	NM_000269.2	500-600
51	NOLC1	NM_004741.3	3405-3505
52	NOP2	NM_001033714.1	1800-1900
53	NPM1	NM_002520.5	10-110
54	NUDCD2	NM_145266.4	368-468
55	PA2G4	NM_006191.2	2475-2575
56	PAICS	NM_001079524.1	2604-2704
57	PDHA1	NM_000284.3	1080-1180
58	PDLIM3	NM_014476.4	897-997
59	PHB	NM_002634.2	1270-1370
60	PHB2	NM_007273.3	1210-1310
61	POLR3K	NM_016310.2	395-495
62	PPAT	NM_002703.3	1210-1310
63	PPRC1	NM_015062.3	4640-4740
64	PRMT1	NM_001536.4	444-544
65	PSMG1	NM_203433.1	655-755
66	RAB8B	NM_016530.2	4157-4257
67	RANBP1	NM_002882.2	380-480
68	RFC3	NM_002915.3	740-840
69	RIN2	NM_018993.2	690-790
70	RPIA	NM_144563.2	1588-1688
71	RPL22	NM_000983.3	1270-1370
72	RPL23	NM_000978.3	71-171
73	RRS1	NM_015169.3	1247-1347
74	SFRS7	NM_001031684.2	532-632
75	SRM	NM_003132.2	254-354
76	SSBP1	NM_003143.1	235-335
77	STRAP	NM_007178.3	1535-1635
78	STRBP	NM_001171137.1	1150-1250
79	TFDP1	NM_007111.4	1826-1926
80	TIPIN	NM_017858.2	230-330
81	TMEM97	NM_014573.2	2055-2155
82	TRAP1	NM_016292.2	1293-1393
83	TYMS	NM_001071.1	555-655
84	UBE2CBP (UBE3D)	NM_198920.1	834-934
85	UCHL3	NM_006002.3	375-475
86	WDR3	NM_006784.2	90-190

Additional Published

	Gene Symbol	Accession Number	Target Region (base pairs)
1	ABCE1	NM_001040876.1	635-735
2	AIMP2	NM_006303.3	507-607
3	BRD2	NM_005104.2	1890-1990
4	BRD3	NM_007371.3	2645-2745
5	BRD4	NM_014299.2	745-845
6	CAD	NM_004341.3	2380-2480
7	CD44	NM_001001392.1	429-529
8	CDK4	NM_000075.2	1055-1155
9	EBNA1BP2	NM_006824.2	318-418
10	EEF1A2	NM_001958.2	1045-1145
11	EXOSC8	NM_181503.1	655-755
12	FASN	NM_004104.4	5387-5487
13	HNRNPA2B1	NM_002137.3	435-535
14	IARS	NM_002161.3	3952-4052
15	LDHA	NM_001165414.1	1690-1790
16	LRP8	NM_033300.2	1590-1690
17	MAT2A	NM_005911.4	805-905
18	MAX	NM_002382.3	240-340
19	MITF	NM_000248.3	3240-3340
20	MRPL3	NM_007208.2	350-450
21	MYCL1	NM_001033081.2	568-668
22	MYCN	NM_005378.4	1545-1645
23	NCL	NM_005381.2	1492-1592
24	p50 (NFKB1)	NM_003998.2	1675-1775
25	p65 (GORASP1)	NM_031899.2	2755-2855
26	PEBP1	NM_002567.2	1335-1435
27	POLD2	NM_006230.1	505-605
28	POLR2H	NM_006232.2	317-417
29	PRDX4	NM_006406.1	540-640
30	PYCR1	NM_006907.2	513-613
31	RPL23	NM_000978.3	71-171
32	RRP1B	NM_015056.2	1070-1170
33	SLC16A1	NM_003051.3	635-735
34	SLC39A14	NM_001128431.2	1245-1345
35	SLC39A6	NM_012319.2	1580-1680
36	TBL3	NM_006453.2	1070-1170
37	UCK2	NM_012474.3	730-830

Data-driven MYC High

	Gene Symbol	Accession Number	Target Region (base pairs)
1	KIAA1737	NM_033426.2	3868-3968
2	FAM211A-AS1 / C17orf76-AS1	NR_027164.1	214-314
3	PCDH9	NM_020403.3	3580-3680
4	SAMD13	NM_001010971.2	672-772
5	SERHL2	NM_014509.3	637-737
6	TCL1A	NM_001098725.1	867-967
7	TMEM100	NM_018286.2	655-755

Data-driven MYC Low

	Gene Symbol	Accession Number	Target Region (base pairs)
1	SHISA8	NM_001207020.1	1111-1211
2	EGFL6	NM_015507.2	1495-1595
3	IGFBP2	NM_000597.2	675-775
4	P2RY12	NM_022788.3	230-330
5	SLAMF1	NM_003037.2	580-680
6	SLC12A8	NM_024628.5	770-870
7	TDO2	NM_005651.1	0-100
8	TMEM119	NM_181724.2	1490-1590

TCF3

	Gene Symbol	Accession Number	Target Region (base pairs)
1	ALDH5A1	NM_001080.3	455-555
2	ATF4	NM_001675.2	1151-1251
3	BMP7	NM_001719.1	525-625
4	KIAA0226L / C13orf18	NM_025113.2	1071-1171
5	CBFA2T3	NM_005187.5	3195-3295
6	CCRL1	NM_178445.1	2200-2300
7	CD38	NM_001775.2	1035-1135
8	CXCR4	NM_003467.2	1335-1435
9	NSG1 / D4S234E	NM_014392.3	1860-1960
10	DNMT3B	NM_175850.1	1950-2050
11	DVL2	NM_004422.2	1025-1125
12	DYRK3	NM_003582.2	1310-1410
13	E2F2	NM_004091.2	3605-3705
14	FUT1	NM_000148.3	3660-3760
15	GPLD1	NM_001503.2	465-565
16	GRAP	NM_006613.3	1918-2018
17	ICOSLG	NM_015259.4	1190-1290
18	ID3	NM_002167.3	195-295
19	IGLL1	NM_020070.2	188-288
20	LHFP	NM_005780.2	460-560
21	LZTS1	NM_021020.2	3970-4070
22	MME	NM_000902.2	5059-5159
23	MRPS35	NM_021821.2	250-350
24	N4BP3	NM_015111.1	5435-5535
25	NEIL1	NM_024608.2	1675-1775
26	NEIL3	NM_018248.2	842-942
27	PPM1A	NM_021003.4	550-650
28	PPP2R5C	NM_002719.3	1240-1340
29	PRKAR2B	NM_002736.2	1350-1450
30	RAPGEF5	NM_012294.3	3420-3520
31	RIMS3	NM_014747.2	3580-3680
32	SLC1A4	NM_003038.4	3030-3130
33	SYNE2	NM_182914.2	20435-20535
34	TBC1D1	NM_001253915.1	1926-2026
35	TCF3	NM_003200.2	4325-4425
36	TNFSF8	NM_001244.3	518-618
37	YPEL1	NM_013313.3	2270-2370

BCL2

	Gene Symbol	Accession Number	Target Region (base pairs)
1	BCL-W/BCL2L2	NM_004050.2	2300-2400
2	BCL2	NM_000657.2	5-105
3	BCL2A1	NM_004049.2	80-180
4	BCL2L1	NM_138578.1	1560-1660
5	MCL1	NM_021960.3	1260-1360

Lineage-specific (‘Constitutional’)

	Gene Symbol	Accession Number	Target Region (base pairs)
1	CD3E	NM_000733.2	75-175
2	CD19	NM_001770.4	1770-1870
3	CD20/MS4A1	NM_152866.2	620-720
4	CD68	NM_001251.2	1140-1240
5	CD79A	NM_001783.3	695-795

Housekeeping Genes

	Gene Symbol	Accession Number	Target Region (base pairs)
1	AAMP	NM_001087.3	1646-1746
2	ACTB	NM_001101.2	1010-1110
3	FTL	NM_000146.3	85-185
4	GAPDH	NM_002046.3	972-1072
5	GNB2L1	NM_006098.4	375-475
6	H3F3A	NM_002107.3	190-290
7	HMBS	NM_000190.3	315-415
8	KARS	NM_005548.2	1885-1985
9	PPIA (Cyclophyllin A)	NM_021130.2	925-1025
10	PSMB3	NM_002795.2	340-440
11	PSMD2	NM_002808.3	771-871
12	PTDSS1	NM_014754.1	2375-2475
13	TBP	NM_001172085.1	587-687
14	TCFL1 (VPS72)	NM_005997.1	1112-1212
15	TUBB	NM_178014.2	320-420

Table II.3 - 80-gene Final Profiling Panel

The 80 gene targets used in the final profiling panel are listed and are organized into groups based on biological pathways.

	Gene Symbol	Accession Number	Target Region (base pairs)
1	AHCY	NM_000687.2	1805-1905
2	AKAP1	NM_139275.1	2725-2825
3	APEX1	NM_001641.2	727-827
4	APITD1	NM_199294.2	950-1050
5	BUB1B	NM_001211.4	835-935

6	FAM216A / C12ORF24	NM_013300.2	722-822
7	CDC25A	NM_001789.2	690-790
8	CDK4	NM_000075.2	1055-1155
9	CIRH1A	NM_032830.2	84-184
10	CTPS	NM_001905.2	2570-2670
11	DDX21	NM_004728.2	685-785
12	DHX33	NM_001199699.1	2873-2973
13	DLEU1	NR_002605.1	173-273
14	ETFA	NM_001127716.1	630-730
15	FBL	NM_001436.3	883-983
16	GEMIN4	NM_015721.2	1925-2025
17	GOT2	NM_002080.2	2145-2245
18	KIAA0101	NM_014736.4	65-165
19	LDHB	NM_001174097.1	1190-1290
20	LYAR	NM_001145725.1	230-330
21	MRPS2	NR_051968.1	1512-1612
22	MRPS34	NM_023936.1	719-819
23	MYC	NM_002467.3	1610-1710
24	NME1	NM_000269.2	500-600
25	NOLC1	NM_004741.3	3405-3505
26	NOP2	NM_001033714.1	1800-1900
27	PAICS	NM_001079524.1	2604-2704
28	PHB	NM_002634.2	1270-1370
29	PHB2	NM_007273.3	1210-1310
30	PPAT	NM_002703.3	1210-1310
31	PPRC1	NM_015062.3	4640-4740
32	PRMT1	NM_001536.4	444-544
33	RANBP1	NM_002882.2	380-480
34	RFC3	NM_002915.3	740-840
35	RRS1	NM_015169.3	1247-1347
36	SRM	NM_003132.2	254-354
37	SSBP1	NM_003143.1	235-335
38	STRBP	NM_001171137.1	1150-1250
39	TMEM97	NM_014573.2	2055-2155
40	TRAP1	NM_016292.2	1293-1393
41	UCHL3	NM_006002.3	375-475
42	WDR3	NM_006784.2	90-190

Selected as MYC Targets

	Gene Symbol	Accession Number	Target Region (base pairs)
1	CAD	NM_004341.3	2380-2480
2	EBNA1BP2	NM_006824.2	318-418
3	FASN	NM_004104.4	5387-5487
4	LRP8	NM_033300.2	1590-1690
5	NCL	NM_005381.2	1492-1592
6	POLD2	NM_006230.1	505-605
7	PYCR1	NM_006907.2	513-613
8	SLC16A1	NM_003051.3	635-735
9	UCK2	NM_012474.3	730-830

Data Driven MYC High

	Gene Symbol	Accession Number	Target Region (base pairs)
1	FAM211A-AS1 / C17orf76-AS1	NR_027164.1	214-314
2	KIAA0226L	NM_025113.2	1071-1171
3	PCDH9	NM_020403.3	3580-3680
4	SAMD13	NM_001010971.2	672-772
5	TCL1A	NM_001098725.1	867-967

Data Driven MYC Low

	Gene Symbol	Accession Number	Target Region (base pairs)
1	SHISA8	NM_001207020.1	1111-1211
2	IGFBP2	NM_000597.2	675-775
3	P2RY12	NM_022788.3	230-330
4	SLAMF1	NM_003037.2	580-680
5	SLC12A8	NM_024628.5	770-870
6	TDO2	NM_005651.1	0-100
7	TMEM119	NM_181724.2	1490-1590

Housekeeping Genes

	Gene Symbol	Accession Number	Target Region (base pairs)
1	AAMP	NM_001087.3	1646-1746
2	H3F3A	NM_002107.3	190-290
3	HMBS	NM_000190.3	315-415
4	KARS	NM_005548.2	1885-1985
5	PSMB3	NM_002795.2	340-440
6	TUBB	NM_178014.2	320-420

Table II.4 - The final gene targets

comprising the Diagnostic Classifier (21 genes) and the MYC Activity Classifier (61 genes) are listed, together with the 'relative weight' (variable importance) of each gene in the classifier, as shown in Figures 3A, B and Figures 5A, B (see Supplementary methods). Eight genes (indicated in bold) are used in both classifiers. Housekeeping genes (6) used to normalize the datasets.

Diagnostic Classifier

	Gene Symbol	Accession Number	Target Region (bps)	Variable Importance (0-100)
1	STRBP	NM_001171137.1	1150-1250	98.2
2	PRKAR2B	NM_002736.2	1350-1450	92.9
3	E2F2	NM_004091.2	3605-3705	80.5
4	LZTS1	NM_021020.2	3970-4070	72.6
5	***CDC25A**	**NM_001789.2**	**690-790**	**72.6**
6	TCF3	NM_003200.2	4325-4425	69
7	***RANBP1**	**NM_002882.2**	**380-480**	**58.4**
8	***DLEU1**	**NR_002605.1**	**173-273**	**54.9**

9	***PAICS**	**NM_001079524.1**	**2604-2704**	**46.9**
10	DNMT3B	NM_175850.1	1950-2050	45.1
11	***PPAT**	**NM_002703.3**	**1210-1310**	**44.2**
12	***KIAA0101**	**NM_014736.4**	**65-165**	**43.4**
13	PYCR1	NM_006907.2	513-613	38.1
14	CD10	NM_000902.2	5059-5159	34.5
15	***NME1**	**NM_000269.2**	**500-600**	**17.7**
16	***FAM216A / C12ORF24**	**NM_013300.2**	**722-822**	**7.1**
17	BMP7	NM_001719.1	525-625	0
18	BCL2	NM_000657.2	5-105	46
19	CD44	NM_001001392.1	429-529	57.5
20	p50 (NFKB1)	NM_003998.2	1675-1775	72.6
21	BCL2A1	NM_004049.2	80-180	100

MYC Activity Classifier

	Gene Symbol	Accession Number	Target Region (bps)	Variable Importance (0-100)
1	MYC	NM_002467.3	1610-1710	100
2	SRM	NM_003132.2	254-354	77.8
3	AKAP1	NM_139275.1	2725-2825	73
4	***NME1**	**NM_000269.2**	**500-600**	**72.2**
5	FBL	NM_001436.3	883-983	71.4
6	RFC3	NM_002915.3	740-840	69.8
7	TCL1A	NM_001098725.1	867-967	66.7
8	POLD2	NM_006230.1	505-605	61.9
9	***RANBP1**	**NM_002882.2**	**380-480**	**61.9**
10	GEMIN4	NM_015721.2	1925-2025	60.3
11	MRPS34	NM_023936.1	719-819	60.3
12	DHX33	NM_001199699.1	2873-2973	59.5
13	PPRC1	NM_015062.3	4640-4740	59.5
14	***PPAT**	**NM_002703.3**	**1210-1310**	**57.9**
15	***FAM216A / C12ORF24**	**NM_013300.2**	**722-822**	**57.1**
16	***PAICS**	**NM_001079524.1**	**2604-2704**	**54.8**
17	UCHL3	NM_006002.3	375-475	53.2
18	NOLC1	NM_004741.3	3405-3505	52.4
19	KIAA0226L	NM_025113.2	1071-1171	50.8
20	PRMT1	NM_001536.4	444-544	50.8
21	LDHB	NM_001174097.1	1190-1290	49.2
22	TRAP1	NM_016292.2	1293-1393	47.6
23	AHCY	NM_000687.2	1805-1905	47.6
24	LRP8	NM_033300.2	1590-1690	45.2
25	EBNA1BP2	NM_006824.2	318-418	43.7
26	CDK4	NM_000075.2	1055-1155	42.1
27	ETFA	NM_001127716.1	630-730	41.3
28	UCK2	NM_012474.3	730-830	39.7
29	CTPS	NM_001905.2	2570-2670	39.7
30	GOT2	NM_002080.2	2145-2245	38.9
31	FAM211A / C17ORF76	NR_027164.1	214-314	36.5

	Gene Symbol	Accession Number	Target Region (bps)	Variable Importance (0-100)
32	TMEM97	NM_014573.2	2055-2155	36.5
33	RRS1	NM_015169.3	1247-1347	36.5
34	DDX21	NM_004728.2	685-785	34.9
35	PHB2	NM_007273.3	1210-1310	34.1
36	WDR3	NM_006784.2	90-190	33.3
37	***KIAA0101**	**NM_014736.4**	**65-165**	**31.7**
38	FASN	NM_004104.4	5387-5487	31.7
39	SAMD13	NM_001010971.2	672-772	31
40	***CDC25A**	**NM_001789.2**	**690-790**	**30.2**
41	LYAR	NM_001145725.1	230-330	30.2
42	CAD	NM_004341.3	2380-2480	26.2
43	APEX1	NM_001641.2	727-827	25.4
44	NOP2	NM_001033714.1	1800-1900	22.2
45	PHB	NM_002634.2	1270-1370	20.6
46	SSBP1	NM_003143.1	235-335	19.8
47	MRPS2	NR_051968.1	1512-1612	19
48	CIRH1A	NM_032830.2	84-184	17.5
49	SLC16A1	NM_003051.3	635-735	16.7
50	BUB1B	NM_001211.4	835-935	15.1
51	APITD1	NM_199294.2	950-1050	15.1
52	NCL	NM_005381.2	1492-1592	9.5
53	***DLEU1**	**NR_002605.1**	**173-273**	**7.9**
54	PCDH9	NM_020403.3	3580-3680	0
55	IGFBP2	NM_000597.2	675-775	8.7
56	TDO2	NM_005651.1	0-100	18.3
57	SLC12A8	NM_024628.5	770-870	30.2
58	P2RY12	NM_022788.3	230-330	40.5
59	TMEM119	NM_181724.2	1490-1590	53.2
60	SHISA8	NM_001207020.1	1111-1211	67.5
61	SLAMF1	NM_003037.2	580-680	69.8

****Genes indicated in bold type (n=8) are included in both the Diagnostic and MYC Transcriptional Activity Classifiers**

<u>Housekeeping Genes</u>

	Gene Symbol	Accession Number	Target Region (bps)
1	AAMP	NM_001087.3	1646-1746
2	H3F3A	NM_002107.3	190-290
3	HMBS	NM_000190.3	315-415
4	KARS	NM_005548.2	1885-1985
5	PSMB3	NM_002795.2	340-440
6	TUBB	NM_178014.2	320-420

APPENDIX III

Table III.1: Final CCC signature derived from both Affymetrix discovery sets.

This list contains the final 83 genes that were used for Nanostring profiling both as Ensembl gene identifiers and gene symbols. T-statistic log2 fold change and p values were derived using Limma and corrected for multiple testing using false discovery rate (FDR).

ensembl_gene_id	hgnc_symbol	ensembl_gene_id	hgnc_symbol
ENSG00000000971	CFH	ENSG00000140740	UQCRC2
ENSG00000004779	NDUFAB1	ENSG00000143933	CALM2
ENSG00000005075	POLR2J	ENSG00000147669	POLR2K
ENSG00000005844	ITGAL	ENSG00000149131	SERPING1
ENSG00000010810	FYN	ENSG00000149532	CPSF7
ENSG00000011376	LARS2	ENSG00000153563	CD8A
ENSG00000038427	VCAN	ENSG00000154518	ATP5G3
ENSG00000054983	GALC	ENSG00000155465	SLC7A7
ENSG00000065526	SPEN	ENSG00000156467	UQCRB
ENSG00000065911	MTHFD2	ENSG00000156482	RPL30
ENSG00000072110	ACTN1	ENSG00000157456	CCNB2
ENSG00000072506	HSD17B10	ENSG00000159403	C1R
ENSG00000072864	NDE1	ENSG00000160255	ITGB2
ENSG00000077312	SNRPA	ENSG00000160299	PCNT
ENSG00000078668	VDAC3	ENSG00000163541	SUCLG1
ENSG00000086102	NFX1	ENSG00000163599	CTLA4
ENSG00000088827	SIGLEC1	ENSG00000164258	NDUFS4
ENSG00000089280	FUS	ENSG00000164305	CASP3
ENSG00000095585	BLNK	ENSG00000164405	UQCRQ
ENSG00000096433	ITPR3	ENSG00000164733	CTSB
ENSG00000099783	HNRNPM	ENSG00000165025	SYK
ENSG00000099795	NDUFB7	ENSG00000165264	NDUFB6
ENSG00000099995	SF3A1	ENSG00000165629	ATP5C1
ENSG00000100316	RPL3	ENSG00000166260	COX11
ENSG00000100385	IL2RB	ENSG00000166340	TPP1
ENSG00000100416	TRMU	ENSG00000166483	WEE1
ENSG00000100554	ATP6V1D	ENSG00000167283	ATP5L
ENSG00000100600	LGMN	ENSG00000168040	FADD
ENSG00000102265	TIMP1	ENSG00000168827	GFM1
ENSG00000103653	CSK	ENSG00000171860	C3AR1
ENSG00000104852	SNRNP70	ENSG00000173369	C1QB
ENSG00000104897	SF3A2	ENSG00000173372	C1QA
ENSG00000105323	HNRNPUL1	ENSG00000173482	PTPRM
ENSG00000105568	PPP2R1A	ENSG00000173638	SLC19A1
ENSG00000105974	CAV1	ENSG00000174231	PRPF8
ENSG00000106366	SERPINE1	ENSG00000174748	RPL15
ENSG00000108821	COL1A1	ENSG00000175110	MRPS22

ENSG00000109390	NDUFC1	ENSG00000175216	CKAP5
ENSG00000109861	CTSC	ENSG00000175899	A2M
ENSG00000111537	IFNG	ENSG00000177733	HNRNPA0
ENSG00000112695	COX7A2	ENSG00000182180	MRPS16
ENSG00000115415	STAT1	ENSG00000182199	SHMT2
ENSG00000116288	PARK7	ENSG00000182326	C1S
ENSG00000116459	ATP5F1	ENSG00000182899	RPL35A
ENSG00000116478	HDAC1	ENSG00000183648	NDUFB1
ENSG00000116824	CD2	ENSG00000184076	UQCR10
ENSG00000119013	NDUFB3	ENSG00000184983	NDUFA6
ENSG00000120742	SERP1	ENSG00000186340	THBS2
ENSG00000122406	RPL5	ENSG00000189043	NDUFA4
ENSG00000125356	NDUFA1	ENSG00000189091	SF3B3
ENSG00000125730	C3	ENSG00000196230	TUBB
ENSG00000126267	COX6B1	ENSG00000196235	SUPT5H
ENSG00000127184	COX7C	ENSG00000197081	IGF2R
ENSG00000127540	UQCR11	ENSG00000197249	SERPINA1
ENSG00000127564	PKMYT1	ENSG00000197746	PSAP
ENSG00000129128	SPCS3	ENSG00000197766	CFD
ENSG00000131462	TUBG1	ENSG00000197943	PLCG2
ENSG00000133226	SRRM1	ENSG00000198833	UBE2J1
ENSG00000134470	IL15RA	ENSG00000204843	DCTN1
ENSG00000134575	ACP2	ENSG00000205937	RNPS1
ENSG00000135677	GNS	ENSG00000213619	NDUFS3
ENSG00000135940	COX5B	ENSG00000256043	CTSO
ENSG00000136143	SUCLA2	ENSG00000259494	MRPL46
ENSG00000136875	PRPF4	ENSG00000108883	EFTUD2
ENSG00000137462	TLR2	ENSG00000120217	CD274
ENSG00000137822	TUBGCP4	ENSG00000131368	MRPS25
ENSG00000138777	PPA2	ENSG00000135972	MRPS9
ENSG00000139131	YARS2	ENSG00000146282	RARS2
ENSG00000140374	ETFA	ENSG00000159189	C1QC
ENSG00000140612	SEC11A	ENSG00000160593	AMICA1
		ENSG00000184752	NDUFA12

Table III.2: Differential expression of gene signature in Nanostring dataset.

This spreadsheet shows the differential expression of the 83 HR genes in both Nanostring datasets.

Nanostring Frozen				Nanostring FFPE		
	logFC	FDR			logFC	FDR
A2M	0.908357	0.005403		A2M	0.591607	0.022232
ACP2	0.755339	0.002481		ACP2	0.400946	0.104
ACTN1	1.145464	0.000314		ACTN1	0.797946	0.018809
AMICA1	1.187875	0.003407		AMICA1	0.540411	0.182022
ATP5F1	-0.30696	0.07919		ATP5F1	-0.20964	0.334589
ATP5G3	-0.3892	0.037463		ATP5G3	-0.26839	0.180524
ATP5L	-0.44107	0.031704		ATP5L	-0.27223	0.231515
BID	0.271902	0.250435		BID	0.294223	0.334589
BLNK	-0.16061	0.598784		BLNK	-0.76929	0.04292
C1QA	1.355955	0.016833		C1QA	0.993107	0.034991

C1QB	1.439402	0.044264		C1QB	1.25775	0.032882
C1QC	1.268696	0.053057		C1QC	1.0455	0.040313
C1R	1.201143	0.000162		C1R	0.58975	0.055828
C3	0.607152	0.190195		C3	0.408018	0.404672
C3AR1	1.267321	0.00172		C3AR1	0.693393	0.047457
CASP3	-0.00957	0.963439		CASP3	0.198911	0.473
CAV1	0.476268	0.185395		CAV1	0.248062	0.488445
CCNB2	-0.93114	0.001473		CCNB2	-1.47576	0.018809
CD19	-0.51646	0.190195		CD19	-0.74324	0.164546
CD2	1.816045	0.001473		CD2	1.396259	0.018809
CD274	1.100348	0.010224		CD274	0.522875	0.190777
CD8A	1.213598	0.010224		CD8A	0.820339	0.057118
CFD	0.990687	0.002887		CFD	1.129232	0.01222
CFH	1.005821	0.002312		CFH	0.645179	0.049496
COL1A2	1.083196	0.040141		COL1A2	0.172161	0.747032
COX11	-0.44562	0.018821		COX11	-0.58204	0.033507
COX6B1	-0.13705	0.387406		COX6B1	-0.08	0.645026
COX6C	-0.49304	0.011623		COX6C	-0.46795	0.018809
COX7C	-0.20839	0.190195		COX7C	-0.14652	0.441167
CPSF7	0.013089	0.951176		CPSF7	-0.19429	0.257635
CTLA4	1.311839	0.005403		CTLA4	0.767848	0.168356
CTSB	0.97	0.007995		CTSB	0.532411	0.127067
DLD	-0.135	0.387406		DLD	-0.06571	0.746532
ETFA	-0.4492	0.010224		ETFA	-0.53938	0.01222
FYN	1.133679	0.003407		FYN	0.834786	0.025334
GNS	0.976857	0.000314		GNS	0.719554	0.009491
HDAC1	-0.55268	0.025474		HDAC1	-0.59607	0.019567
HSD17B10	-0.4558	0.022463		HSD17B10	-0.57426	0.019567
IGF2R	0.739652	0.009209		IGF2R	0.387598	0.164546
IL15RA	0.88592	0.002312		IL15RA	0.557286	0.061349
IL2RB	1.588688	0.001473		IL2RB	0.993241	0.043884
ITGAL	1.179679	0.004718		ITGAL	0.819411	0.040313
ITGB2	1.076339	0.001473		ITGB2	0.450036	0.243538
LGMN	0.564929	0.169571		LGMN	0.254107	0.441167
MRPL46	-0.40517	0.036398		MRPL46	-0.4054	0.128521
MRPS16	-0.33533	0.023619		MRPS16	-0.29691	0.126689
MRPS25	-0.32737	0.087576		MRPS25	-0.52985	0.033507
MRPS7	-0.32205	0.036398		MRPS7	-0.32595	0.077412
MRPS9	-0.19809	0.250435		MRPS9	0.115616	0.807739
MTHFD2	-0.28107	0.190195		MTHFD2	-0.32691	0.286389
NDUFA12	-0.24393	0.169983		NDUFA12	-0.36149	0.132461
NDUFA2	-0.1223	0.542329		NDUFA2	-0.22754	0.231515
NDUFA4	-0.40402	0.036499		NDUFA4	-0.19464	0.334589
NDUFAB1	-0.40366	0.010121		NDUFAB1	-0.22902	0.292903
NDUFB3	-0.01393	0.951176		NDUFB3	-0.16218	0.643855
NDUFB6	0.022786	0.930471		NDUFB6	-0.06992	0.747032
NDUFS3	-0.21143	0.133207		NDUFS3	-0.17982	0.231515
PKMYT1	-0.70107	0.044264		PKMYT1	-1.12479	0.018809
PLCG2	-0.40446	0.144206		PLCG2	-0.6411	0.040313
POLR2J	-0.11136	0.602131		POLR2J	-0.24378	0.257635
POLR2K	-0.5008	0.010121		POLR2K	-0.43179	0.099726
PSAP	0.825982	0.002312		PSAP	0.719107	0.018809
PTPRM	1.16992	0.002336		PTPRM	0.782018	0.034043

RARS2	0.121598	0.486074		RARS2	0.082054	0.746532
RPL15	-0.32161	0.149345		RPL15	-0.185	0.504445
RPL3	-0.20795	0.382915		RPL3	-0.06768	0.816047
RPL35A	-0.46143	0.069957		RPL35A	-0.32518	0.286389
SERPINA1	1.553018	0.001473		SERPINA1	1.071214	0.040313
SERPINE1	0.967152	0.058529		SERPINE1	0.273509	0.58983
SERPING1	1.730714	0.000199		SERPING1	1.111804	0.009491
SF3A2	-0.17249	0.382915		SF3A2	-0.45021	0.038208
SHMT2	-0.71557	0.003637		SHMT2	-0.71779	0.019712
SIGLEC1	1.4725	0.037463		SIGLEC1	1.285429	0.028621
SLC19A1	-0.13056	0.657251		SLC19A1	-0.20425	0.441167
SLC7A7	0.933375	0.000353		SLC7A7	0.38567	0.190777
SPCS3	-0.13473	0.331157		SPCS3	-0.06017	0.747032
STAT1	0.958964	0.01514		STAT1	0.836946	0.034043
SYK	-0.29616	0.190195		SYK	-0.6144	0.040313
THBS2	0.816071	0.190195		THBS2	0.126393	0.827762
TIMP1	0.539429	0.160298		TIMP1	-0.2808	0.504445
TLR2	1.258804	0.001473		TLR2	0.683196	0.052328
TPP1	0.576768	0.010224		TPP1	0.249554	0.164546
TRMU	-0.14744	0.406838		TRMU	-0.28088	0.168356
TUBB	-0.35402	0.063963		TUBB	-0.4667	0.04292
TUBG1	-0.33619	0.086854		TUBG1	-0.42884	0.043884
UBE2J1	-0.32491	0.190195		UBE2J1	-0.10429	0.746532
UQCR10	-0.25105	0.113678		UQCR10	-0.20196	0.286389
UQCRB	-0.09622	0.662509		UQCRB	-0.56874	0.180538
UQCRQ	-0.35741	0.052394		UQCRQ	-0.27946	0.184003
WEE1	-0.60859	0.087576		WEE1	-0.93242	0.023991
YARS2	-0.20509	0.199798		YARS2	-0.30134	0.128521

Table III.3: Housekeeping genes.

This spreadsheet contains the differential expression statistics between all CCC and COO subtypes in the Discovery set I.

			Consensus Clustering Classification (CCC)				Cell-of-origin (COO)		
gene_symbol			class	asymp.fdr	fold.chg		COO class	Asymp fdr	Fold chg
GNAL	housekeeping 3-4		HR	0.9129	1.0019		ABC	0.741195	1.045
BHLHE22	housekeeping 3-4		BCR	0.9421	1.0031		ABC	0.832857	1.0051
BHMT2	housekeeping 4-5		OxPhos	0.9208	1.0099		ABC	0.829773	1.0548
EPHA4	housekeeping 4-5		BCR	0.9285	1.0632		GCB	0.982289	1.0725
EPHB2	housekeeping 5-6		HR	0.8624	1.0115		GCB	0.77195	1.0327
SERPINA3	housekeeping 5-6		OxPhos	0.9318	1.0371		ABC	0.875975	1.0101
TRPC4AP	housekeeping 6-7		BCR	0.946	1.0517		ABC	0.91335	1.0329
HAMP	housekeeping 6-7		OxPhos	0.8779	1.0582		GCB	0.945096	1.1809

SECISBP2	housekeeping 7-8		BCR	0.9495	1.0472		GCB	0.874192	1.0443
MEIS2	housekeeping 7-8		OxPhos	0.8723	1.0533		GCB	0.963452	1.0012
KXD1	housekeeping 8-9		BCR	0.8521	1.0593		GCB	0.577205	1.0269
PSMC5	housekeeping 9-10		BCR	0.7155	1.1459		ABC	0.359016	1.1071
SARS	housekeeping 9-10		BCR	0.7531	1.0979		GCB	0.641504	1.0359
EMC4	housekeeping 10-11		OxPhos	0.8897	1.069		GCB	0.905022	1.0053
KPNB1	housekeeping 10-11		BCR	0.6878	1.1089		GCB	0.910981	1.0102
CYBA	housekeeping 11-12		OxPhos	0.4594	1.0392		GCB	0.7798	1.0311

Table III.4: Final CCC signature derived from both Affymetrix discovery sets.

This list contains the final 141 genes that were used for Nanostring profiling both as Ensembl gene identifiers and gene symbols. In addition, we show the elastic net weights for every gene in all three CCC classes

ensembl_gene_id	hgnc symbol	BCR weights	HR weights	OxPhos weights	description
ENSG00000000971	CFH	-0.112	0.121	0.000	complement factor H
ENSG00000004779	NDUFAB1	0.000	0.000	0.114	NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1, 8kDa
ENSG00000005075	POLR2J	0.000	-0.003	0.000	polymerase (RNA) II (DNA directed) polypeptide J, 13.3kDa
ENSG00000005844	ITGAL	0.000	0.078	-0.071	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)
ENSG00000010810	FYN	-0.012	0.111	0.000	FYN proto-oncogene, Src family tyrosine kinase
ENSG00000011376	LARS2	0.000	0.000	-0.173	leucyl-tRNA synthetase 2, mitochondrial
ENSG00000038427	VCAN	-0.025	0.000	0.000	versican
ENSG00000054983	GALC	-0.043	0.000	0.000	galactosylceramidase
ENSG00000065526	SPEN	0.000	0.000	-0.106	spen family transcriptional repressor
ENSG00000065911	MTHFD2	0.000	-0.053	0.070	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
ENSG00000072110	ACTN1	0.000	0.126	-0.056	actinin, alpha 1
ENSG00000072506	HSD17B10	0.000	-0.032	0.000	hydroxysteroid (17-beta) dehydrogenase 10
ENSG00000072864	NDE1	0.000	0.000	-0.102	nudE neurodevelopment protein 1
ENSG00000077312	SNRPA	0.152	0.000	-0.121	small nuclear ribonucleoprotein polypeptide A
ENSG00000078668	VDAC3	-0.015	0.000	0.000	voltage-dependent anion channel 3
ENSG00000086102	NFX1	0.000	0.000	-0.062	nuclear transcription factor, X-box binding 1
ENSG00000088827	SIGLEC1	0.000	0.038	-0.111	sialic acid binding Ig-like lectin 1, sialoadhesin

ENSG00000089280	FUS	0.172	0.000	-0.135	FUS RNA binding protein
ENSG00000095585	BLNK	0.000	-0.007	0.000	B-cell linker
ENSG00000096433	ITPR3	0.154	0.000	-0.050	inositol 1,4,5-trisphosphate receptor, type 3
ENSG00000099783	HNRNPM	0.075	0.000	0.000	heterogeneous nuclear ribonucleoprotein M
ENSG00000099795	NDUFB7	-0.041	0.000	0.074	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 7, 18kDa
ENSG00000099995	SF3A1	0.079	0.000	-0.158	splicing factor 3a, subunit 1, 120kDa
ENSG00000100316	RPL3	0.000	-0.001	0.000	ribosomal protein L3
ENSG00000100385	IL2RB	0.000	0.069	0.000	interleukin 2 receptor, beta
ENSG00000100416	TRMU	0.455	0.000	0.000	tRNA 5-methylaminomethyl-2-thiouridylate methyltransferase
ENSG00000100554	ATP6V1D	-0.060	0.000	0.103	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D
ENSG00000100600	LGMN	0.000	0.015	0.000	legumain
ENSG00000102265	TIMP1	0.000	0.000	0.000	TIMP metallopeptidase inhibitor 1
ENSG00000103653	CSK	0.090	0.000	-0.045	c-src tyrosine kinase
ENSG00000104852	SNRNP70	0.000	0.000	-0.115	small nuclear ribonucleoprotein 70kDa (U1)
ENSG00000104897	SF3A2	0.007	0.000	-0.269	splicing factor 3a, subunit 2, 66kDa
ENSG00000105323	HNRNPUL1	0.135	0.000	-0.102	heterogeneous nuclear ribonucleoprotein U-like 1
ENSG00000105568	PPP2R1A	0.112	0.000	-0.080	protein phosphatase 2, regulatory subunit A, alpha
ENSG00000105974	CAV1	-0.141	0.047	0.000	caveolin 1, caveolae protein, 22kDa
ENSG00000106366	SERPINE1	0.000	0.000	0.000	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
ENSG00000108821	COL1A1	0.000	0.000	0.000	collagen, type I, alpha 1
ENSG00000109390	NDUFC1	0.000	0.000	0.051	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 1, 6kDa
ENSG00000109861	CTSC	-0.037	0.000	0.000	cathepsin C
ENSG00000111537	IFNG	-0.027	0.000	0.000	interferon, gamma
ENSG00000112695	COX7A2	0.000	0.000	0.031	cytochrome c oxidase subunit VIIa polypeptide 2 (liver)
ENSG00000115415	STAT1	-0.042	0.022	0.000	signal transducer and activator of transcription 1, 91kDa
ENSG00000116288	PARK7	0.000	0.000	0.063	parkinson protein 7
ENSG00000116459	ATP5F1	0.000	0.000	0.054	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit B1
ENSG00000116478	HDAC1	0.047	-0.035	0.000	histone deacetylase 1
ENSG00000116824	CD2	-0.072	0.118	0.000	CD2 molecule
ENSG00000119013	NDUFB3	0.000	0.000	0.076	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 3, 12kDa
ENSG00000120742	SERP1	0.000	0.000	0.073	stress-associated endoplasmic reticulum protein 1
ENSG00000122406	RPL5	0.000	0.000	0.053	ribosomal protein L5
ENSG00000125356	NDUFA1	0.000	0.000	0.048	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa

ENSG00000125730	C3	0.000	0.047	0.000	complement component 3
ENSG00000126267	COX6B1	0.000	-0.068	0.011	cytochrome c oxidase subunit VIb polypeptide 1 (ubiquitous)
ENSG00000127184	COX7C	0.000	0.000	0.013	cytochrome c oxidase subunit VIIc
ENSG00000127540	UQCR11	0.000	0.000	0.057	ubiquinol-cytochrome c reductase, complex III subunit XI
ENSG00000127564	PKMYT1	0.147	-0.066	0.000	protein kinase, membrane associated tyrosine/threonine 1
ENSG00000129128	SPCS3	0.000	-0.002	0.150	signal peptidase complex subunit 3 homolog (S. cerevisiae)
ENSG00000131462	TUBG1	0.040	0.000	0.000	tubulin, gamma 1
ENSG00000133226	SRRM1	0.000	0.000	-0.020	serine/arginine repetitive matrix 1
ENSG00000134470	IL15RA	0.000	0.191	0.000	interleukin 15 receptor, alpha
ENSG00000134575	ACP2	0.000	0.061	0.000	acid phosphatase 2, lysosomal
ENSG00000135677	GNS	0.000	0.161	0.000	glucosamine (N-acetyl)-6-sulfatase
ENSG00000135940	COX5B	0.000	0.000	0.021	cytochrome c oxidase subunit Vb
ENSG00000136143	SUCLA2	0.000	0.000	0.062	succinate-CoA ligase, ADP-forming, beta subunit
ENSG00000136875	PRPF4	0.120	0.000	0.000	pre-mRNA processing factor 4
ENSG00000137462	TLR2	-0.026	0.017	0.000	toll-like receptor 2
ENSG00000137822	TUBGCP4	0.017	0.000	0.000	tubulin, gamma complex associated protein 4
ENSG00000138777	PPA2	0.000	0.000	0.053	pyrophosphatase (inorganic) 2
ENSG00000139131	YARS2	0.000	-0.001	0.052	tyrosyl-tRNA synthetase 2, mitochondrial
ENSG00000140374	ETFA	0.000	-0.023	0.051	electron-transfer-flavoprotein, alpha polypeptide
ENSG00000140612	SEC11A	-0.049	0.000	0.067	SEC11 homolog A (S. cerevisiae)
ENSG00000140740	UQCRC2	-0.010	0.000	0.046	ubiquinol-cytochrome c reductase core protein II
ENSG00000143933	CALM2	0.000	0.000	0.026	calmodulin 2 (phosphorylase kinase, delta)
ENSG00000147669	POLR2K	0.000	-0.033	0.055	polymerase (RNA) II (DNA directed) polypeptide K, 7.0kDa
ENSG00000149131	SERPING1	0.000	0.078	-0.022	serpin peptidase inhibitor, clade G (C1 inhibitor), member 1
ENSG00000149532	CPSF7	0.000	0.000	-0.067	cleavage and polyadenylation specific factor 7, 59kDa
ENSG00000153563	CD8A	0.000	0.042	0.000	CD8a molecule
ENSG00000154518	ATP5G3	0.000	0.000	0.000	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C3 (subunit 9)
ENSG00000155465	SLC7A7	0.000	0.029	0.000	solute carrier family 7 (amino acid transporter light chain, y+L system), member 7
ENSG00000156467	UQCRB	0.000	0.000	0.017	ubiquinol-cytochrome c reductase binding protein
ENSG00000156482	RPL30	0.000	0.000	0.012	ribosomal protein L30
ENSG00000157456	CCNB2	0.051	-0.100	0.000	cyclin B2
ENSG00000159403	C1R	-0.066	0.087	0.000	complement component 1, r subcomponent
ENSG00000160255	ITGB2	0.000	0.064	-0.021	integrin, beta 2 (complement component 3

					receptor 3 and 4 subunit)
ENSG00000160299	PCNT	0.105	0.000	-0.016	pericentrin
ENSG00000163541	SUCLG1	-0.043	0.000	0.144	succinate-CoA ligase, alpha subunit
ENSG00000163599	CTLA4	0.000	0.018	0.000	cytotoxic T-lymphocyte-associated protein 4
ENSG00000164258	NDUFS4	0.000	0.000	0.041	NADH dehydrogenase (ubiquinone) Fe-S protein 4, 18kDa (NADH-coenzyme Q reductase)
ENSG00000164305	CASP3	0.000	-0.052	0.047	caspase 3, apoptosis-related cysteine peptidase
ENSG00000164405	UQCRQ	0.000	0.000	0.031	ubiquinol-cytochrome c reductase, complex III subunit VII, 9.5kDa
ENSG00000164733	CTSB	-0.004	0.028	0.000	cathepsin B
ENSG00000165025	SYK	0.087	0.000	0.000	spleen tyrosine kinase
ENSG00000165264	NDUFB6	0.000	0.000	0.010	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6, 17kDa
ENSG00000165629	ATP5C1	0.000	0.000	0.043	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1
ENSG00000166260	COX11	0.000	-0.005	0.044	COX11 cytochrome c oxidase copper chaperone
ENSG00000166340	TPP1	0.000	0.033	-0.158	tripeptidyl peptidase I
ENSG00000166483	WEE1	0.158	-0.139	0.000	WEE1 G2 checkpoint kinase
ENSG00000167283	ATP5L	0.000	0.000	0.021	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit G
ENSG00000168040	FADD	-0.033	0.000	0.111	Fas (TNFRSF6)-associated via death domain
ENSG00000168827	GFM1	0.000	0.000	0.000	G elongation factor, mitochondrial 1
ENSG00000171860	C3AR1	-0.032	0.062	0.000	complement component 3a receptor 1
ENSG00000173369	C1QB	0.000	0.000	0.000	complement component 1, q subcomponent, B chain
ENSG00000173372	C1QA	-0.004	0.000	0.000	complement component 1, q subcomponent, A chain
ENSG00000173482	PTPRM	0.000	0.143	0.000	protein tyrosine phosphatase, receptor type, M
ENSG00000173638	SLC19A1	0.116	-0.042	0.000	solute carrier family 19 (folate transporter), member 1
ENSG00000174231	PRPF8	0.002	0.000	-0.081	pre-mRNA processing factor 8
ENSG00000174748	RPL15	0.000	-0.046	0.027	ribosomal protein L15
ENSG00000175110	MRPS22	0.000	0.000	0.070	mitochondrial ribosomal protein S22
ENSG00000175216	CKAP5	0.211	0.000	-0.014	cytoskeleton associated protein 5
ENSG00000175899	A2M	0.000	0.105	-0.055	alpha-2-macroglobulin
ENSG00000177733	HNRNPA0	0.040	0.000	-0.049	heterogeneous nuclear ribonucleoprotein A0
ENSG00000182180	MRPS16	0.000	-0.149	0.109	mitochondrial ribosomal protein S16
ENSG00000182199	SHMT2	0.038	-0.120	0.000	serine hydroxymethyltransferase 2 (mitochondrial)
ENSG00000182326	C1S	-0.014	0.000	0.000	complement component 1, s subcomponent
ENSG00000182899	RPL35A	0.000	-0.083	0.038	ribosomal protein L35a

ENSG00000183648	NDUFB1	0.000	0.000	0.087	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 1, 7kDa
ENSG00000184076	UQCR10	0.000	-0.058	0.015	ubiquinol-cytochrome c reductase, complex III subunit X
ENSG00000184983	NDUFA6	-0.019	0.000	0.069	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa
ENSG00000186340	THBS2	0.000	0.023	0.000	thrombospondin 2
ENSG00000189043	NDUFA4	0.000	-0.006	0.000	NDUFA4, mitochondrial complex associated
ENSG00000189091	SF3B3	0.060	0.000	-0.127	splicing factor 3b, subunit 3, 130kDa
ENSG00000196230	TUBB	0.097	-0.032	0.000	tubulin, beta class I
ENSG00000196235	SUPT5H	0.141	0.000	-0.034	suppressor of Ty 5 homolog (S. cerevisiae)
ENSG00000197081	IGF2R	0.000	0.043	-0.041	insulin-like growth factor 2 receptor
ENSG00000197249	SERPINA1	0.000	0.028	0.000	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
ENSG00000197746	PSAP	0.000	0.102	-0.003	prosaposin
ENSG00000197766	CFD	-0.008	0.044	0.000	complement factor D (adipsin)
ENSG00000197943	PLCG2	0.201	-0.034	0.000	phospholipase C, gamma 2 (phosphatidylinositol-specific)
ENSG00000198833	UBE2J1	0.000	-0.044	0.014	ubiquitin-conjugating enzyme E2, J1
ENSG00000204843	DCTN1	0.000	0.000	-0.035	dynactin 1
ENSG00000205937	RNPS1	0.000	0.000	-0.020	RNA binding protein S1, serine-rich domain
ENSG00000213619	NDUFS3	0.000	0.000	0.060	NADH dehydrogenase (ubiquinone) Fe-S protein 3, 30kDa (NADH-coenzyme Q reductase)
ENSG00000256043	CTSO	-0.083	0.000	0.000	cathepsin O
ENSG00000259494	MRPL46	0.000	0.000	0.071	mitochondrial ribosomal protein L46
ENSG00000108883	EFTUD2	0.101	0.000	0.000	elongation factor Tu GTP binding domain containing 2
ENSG00000120217	CD274	-0.268	0.252	0.016	CD274 molecule
ENSG00000131368	MRPS25	0.138	0.000	0.000	mitochondrial ribosomal protein S25
ENSG00000135972	MRPS9	0.000	-0.049	0.000	mitochondrial ribosomal protein S9
ENSG00000146282	RARS2	0.000	0.000	0.000	arginyl-tRNA synthetase 2, mitochondrial
ENSG00000159189	C1QC	0.000	0.011	0.000	complement component 1, q subcomponent, C chain
ENSG00000160593	AMICA1	-0.103	0.146	0.000	adhesion molecule, interacts with CXADR antigen 1
ENSG00000184752	NDUFA12	0.000	-0.016	0.000	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 12

Table III.5: Comparison between different classification methods.

In this table we compared our elastic net classification model to two other state-of-the-art prediction model: Random Forest and Shrunken Centroids (PAM). We compared the three in both the discovery sets. The elastic net outperforms the other two in each dataset. (ACC: accuracy, SENS: sensitivity, SPEC: specificity)

	Elastic Net	Random Forest	Shrunken Centroid
	Discovery I - 10 fold CV		
ACC	0.965	0.922	0.971
SENS - BCR	0.980	0.940	0.98
SPEC - BCR	0.956	0.945	0.978
SENS - HR	0.905	0.810	0.929
SPEC - HR	1.000	0.980	1.000
SENS – OxP	1.000	1.000	1.000
SPEC - OxP	0.989	0.957	0.978
	Discovery II - 10 fold CV		
ACC	0.918	0.918	0.904
SENS - BCR	0.840	0.960	0.840
SPEC - BCR	0.979	0.958	1.000
SENS - HR	1.000	0.864	1.000
SPEC - HR	0.907	0.941	0.882
SENS – OxP	0.889	0.923	0.885
SPEC - OxP	0.982	0.979	0.979

Table III.6: Cross-validation within the discovery and validation sets.

For the discovery we used 10-fold cross-validation, while for the validation sets we used leave-out-one cross-validation (LOOCV). The first three columns show all the Affymetrix data, while the last two show the prediction performance of the Nanostring data. (ACC: accuracy, SENS: sensitivity, SPEC: specificity)

Measurement	Discovery I	Discovery II	Validation Affymetrix (44 replicates)	Validation Nanostring frozen	Validation Nanostring FFPE
Technology	Affymetrix			Nanostring	
Samples	141	72	44	44	44
ACC	0.965	0.918	0.818	0.818	0.591
SENS - BCR	0.980	0.840	0.545	0.786	0.929
SPEC - BCR	0.956	0.979	0.939	1.000	0.600
SENS - HR	0.905	1.000	0.947	0.875	0.688
SPEC - HR	1.000	0.907	0.840	0.821	0.893
SENS – OxP	1.000	0.889	0.857	0.786	0.143
SPEC - OxP	0.989	0.982	0.933	0.900	0.900

Table III.7: Comparison between different classification methods on HR.

In this table we compared our elastic net classification model to two other state-of-the-art prediction model: Random Forest and Shrunken Centroids (PAM). We compared the three in both the discovery sets. The elastic net outperforms the other two in each dataset. (ACC: accuracy, SENS: sensitivity, SPEC: specificity, PPV: positive predictive value, NPV: negative predictive value, FDR: false discovery rate, AUC: area under the receiver operating characteristic (ROC) curve)

	Elastic Net	Random Forest	Shrunken Centroid
	Discovery I - 10 fold CV		
ACC	97.87	95.75	94.33
SENS	100.00	97.62	88.10
SPEC	96.97	94.95	96.97
PPV	93.33	89.13	92.50
NPV	100.00	98.95	95.05
FDR	6.67	10.87	7.50
AUC	0.998	0.993	0.992
	Discovery II - 10 fold CV		
ACC	94.52	90.41	89.04
SENS	100.00	77.27	95.46
SPEC	92.16	96.08	86.28
PPV	84.62	89.47	75.00
NPV	100.00	90.74	97.78
FDR	15.39	10.53	25.00
AUC	0.986	0.975	0.985

Table III.8: Elastic net weights for every gene in the host response signature.

The first two columns indicate the ensemble gene identifier and the official human gene symbols.

ensembl_gene_id	hgnc symbol	Elastic net weights	ensembl_gene_id	hgnc symbol	Elastic net weights
ENSG00000134470	IL15RA	-0.436711628	ENSG00000165025	SYK	0.004823
ENSG00000135677	GNS	-0.329371765	ENSG00000091140	DLD	0.012767
ENSG00000173482	PTPRM	-0.308114947	ENSG00000095585	BLNK	0.018337
ENSG00000160593	AMICA1	-0.284119766	ENSG00000100316	RPL3	0.024199
ENSG00000000971	CFH	-0.266068087	ENSG00000154518	ATP5G3	0.027412
ENSG00000120217	CD274	-0.25226594	ENSG00000177455	CD19	0.030803
ENSG00000116824	CD2	-0.241012315	ENSG00000156467	UQCRB	0.032106
ENSG00000005844	ITGAL	-0.233924103	ENSG00000167283	ATP5L	0.03315
ENSG00000072110	ACTN1	-0.231583408	ENSG00000127184	COX7C	0.034568
ENSG00000175899	A2M	-0.230519533	ENSG00000131495	NDUFA2	0.040894
ENSG00000163599	CTLA4	-0.210241287	ENSG00000004779	NDUFAB1	0.041131

ENSG00000010810	FYN	-0.207799834	ENSG00000139131	YARS2	0.049355
ENSG00000100385	IL2RB	-0.199650683	ENSG00000119013	NDUFB3	0.049651
ENSG00000149532	CPSF7	-0.199345075	ENSG00000140374	ETFA	0.052395
ENSG00000197746	PSAP	-0.194025894	ENSG00000259494	MRPL46	0.053649
ENSG00000088827	SIGLEC1	-0.154998696	ENSG00000131462	TUBG1	0.055314
ENSG00000160255	ITGB2	-0.150766398	ENSG00000005075	POLR2J	0.063903
ENSG00000104897	SF3A2	-0.137783777	ENSG00000147669	POLR2K	0.084458
ENSG00000159403	C1R	-0.134652887	ENSG00000166260	COX11	0.08735
ENSG00000125730	C3	-0.129015599	ENSG00000116478	HDAC1	0.092286
ENSG00000171860	C3AR1	-0.128827905	ENSG00000189043	NDUFA4	0.107185
ENSG00000155465	SLC7A7	-0.123000336	ENSG00000146282	RARS2	0.126587
ENSG00000115415	STAT1	-0.12124145	ENSG00000125445	MRPS7	0.126703
ENSG00000166340	TPP1	-0.121183942	ENSG00000197943	PLCG2	0.129381
ENSG00000197081	IGF2R	-0.117023452	ENSG00000196230	TUBB	0.130272
ENSG00000149131	SERPING1	-0.11179474	ENSG00000015475	BID	0.130955
ENSG00000197766	CFD	-0.103825318	ENSG00000174748	RPL15	0.138678
ENSG00000105974	CAV1	-0.094964526	ENSG00000072506	HSD17B10	0.139958
ENSG00000164733	CTSB	-0.090364804	ENSG00000184752	NDUFA12	0.146401
ENSG00000153563	CD8A	-0.08277361	ENSG00000126267	COX6B1	0.163013
ENSG00000106366	SERPINE1	-0.081081423	ENSG00000198833	UBE2J1	0.169715
ENSG00000186340	THBS2	-0.071539643	ENSG00000065911	MTHFD2	0.170888
ENSG00000197249	SERPINA1	-0.07114056	ENSG00000182899	RPL35A	0.17527
ENSG00000159189	C1QC	-0.057051528	ENSG00000135972	MRPS9	0.175637
ENSG00000137462	TLR2	-0.053809724	ENSG00000184076	UQCR10	0.186319
ENSG00000102265	TIMP1	-0.052931467	ENSG00000164305	CASP3	0.188894
ENSG00000100600	LGMN	-0.0389156	ENSG00000129128	SPCS3	0.216158
ENSG00000134575	ACP2	-0.038184667	ENSG00000157456	CCNB2	0.22199
ENSG00000173369	C1QB	-0.022017497	ENSG00000100416	TRMU	0.254038
ENSG00000173372	C1QA	-0.016568392	ENSG00000182180	MRPS16	0.312914
ENSG00000116459	ATP5F1	0	ENSG00000131368	MRPS25	0.32306
ENSG00000164405	UQCRQ	0	ENSG00000182199	SHMT2	0.33646
ENSG00000164692	COL1A2	0	ENSG00000166483	WEE1	0.362628
ENSG00000164919	COX6C	0	ENSG00000127564	PKMYT1	0.517848
ENSG00000165264	NDUFB6	0	ENSG00000173638	SLC19A1	0.850869
ENSG00000213619	NDUFS3	0			

Table III.9: HR Cross-validation within the discovery and validation sets.

For the discovery we used 10-fold cross-validation, while for the validation sets we used leave-out-one cross-validation (LOOCV). The first three columns show all the Affymetrix data, while the last two show the prediction performance of the Nanostring data. (ACC: accuracy, SENS: sensitivity, SPEC: specificity, PPV: positive predictive value, NPV: negative predictive value, FDR: false discovery rate, AUC: area under the receiver operating characteristic (ROC) curve)

Measurement	Discovery I	Discovery II	Validation Affymetrix (44 replicates)	Validation Nanostring frozen	Validation Nanostring FFPE
Technology	Affymetrix			Nanostring	
Samples	141	72	44	44	44
ACC	97.87	94.52	90.91	84.09	84.09
SENS	100.00	100.00	93.75	93.75	81.25
SPEC	96.97	92.16	89.29	78.57	85.71
PPV	93.33	84.62	83.33	71.43	76.47
NPV	100.00	100.00	96.15	95.65	88.89
FDR	6.67	15.39	16.67	28.57	23.53
AUC	0.998	0.986	0.960	0.906	0.804

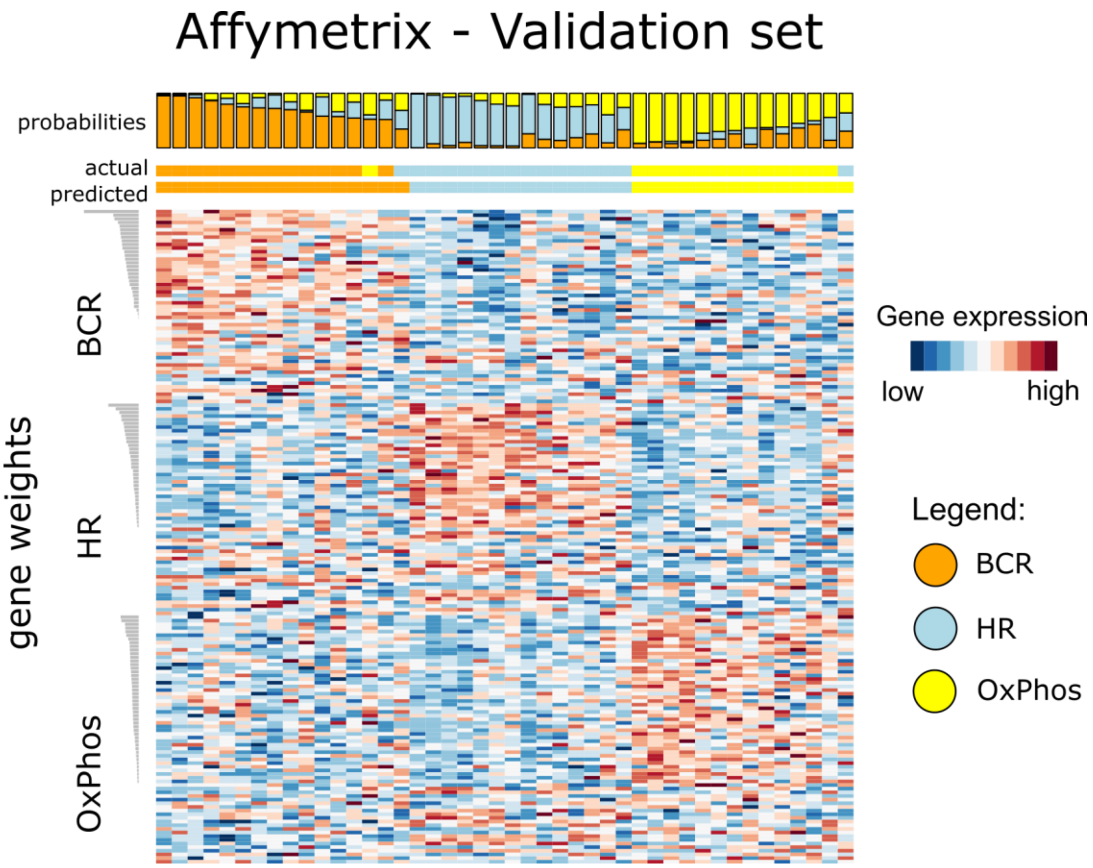


Figure III.1: Heatmap of the 44 replicates of the validation cohort in Affymetrix.

The samples are ordered by class probabilities, based on the predictions of an elastic net model trained on the discovery set I. The top barplot shows the single class probabilities of the classifier, the color-bars below shows the gold standard and predicted CCC subgroups. Each row corresponds to a gene in our CCC signature, which are grouped by class and weights within the elastic net model.

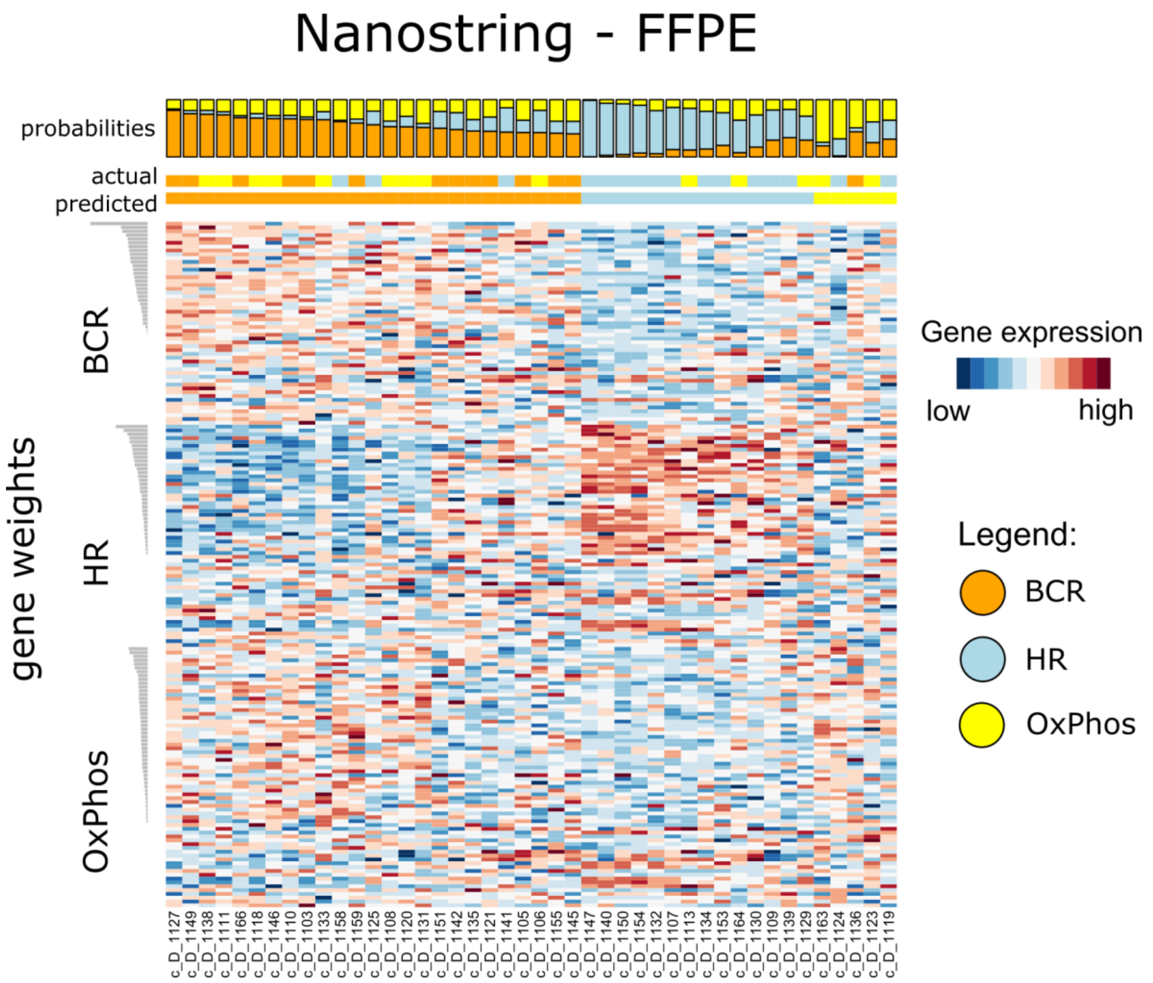


Figure III.2: Heatmap of the 44 replicates of the validation cohort in Nanostring FFPE.

The samples are ordered by class probabilities, based on the predictions of an elastic net model trained on the discovery set I. The top barplot shows the single class probabilities of the classifier, the color-bars below shows the gold standard and predicted CCC subgroups. Each row corresponds to a gene in our CCC signature, which are grouped by class and weights within the elastic net model.

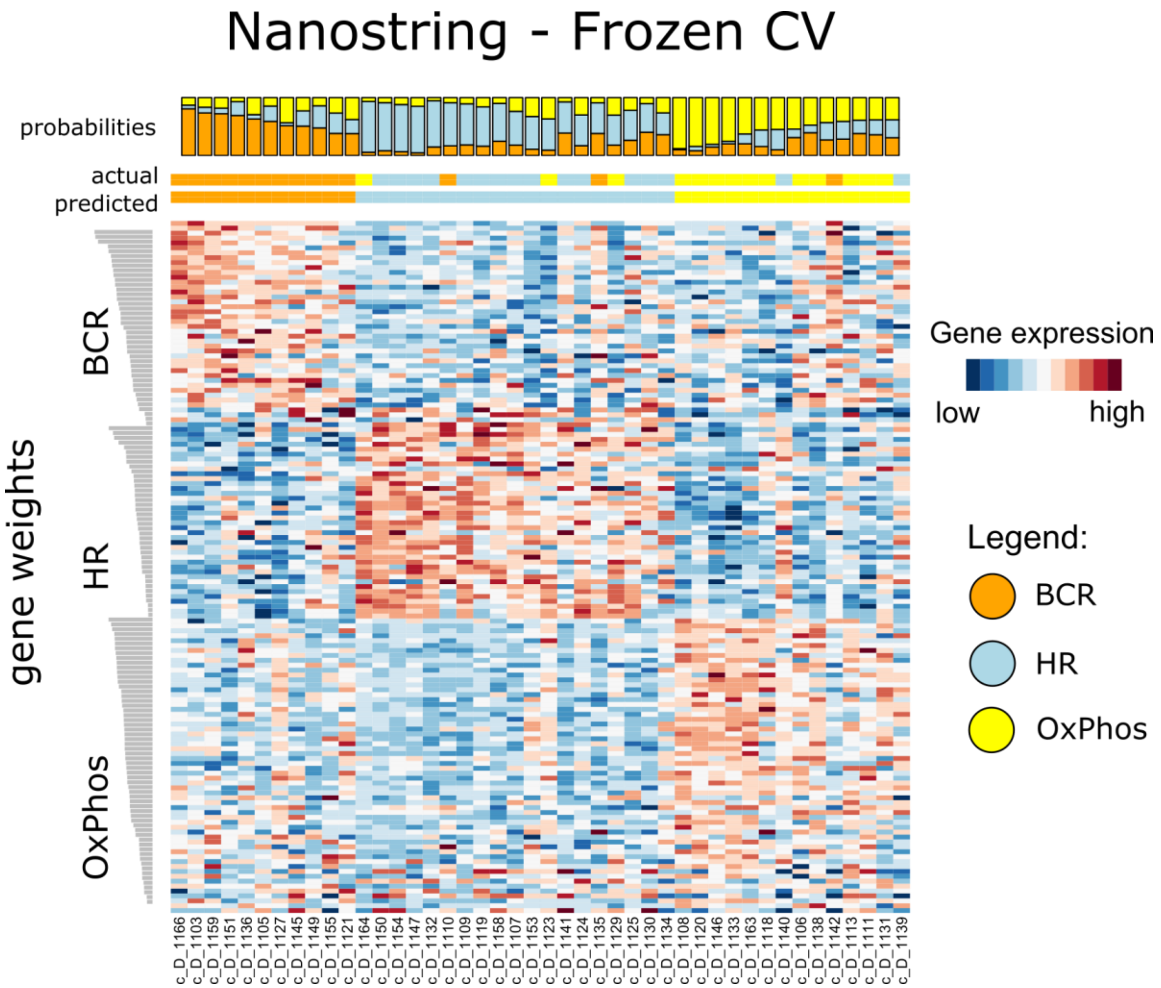


Figure III.3: Nanostring Frozen LOOCV.

The samples are ordered by class probabilities, based on the predictions of an elastic net model trained on the discovery set I. The top barplot shows the single class probabilities of the classifier, the color-bars below shows the gold standard and predicted CCC subgroups. Each row corresponds to a gene in our CCC signature, which are grouped by class and weights within the elastic net model.

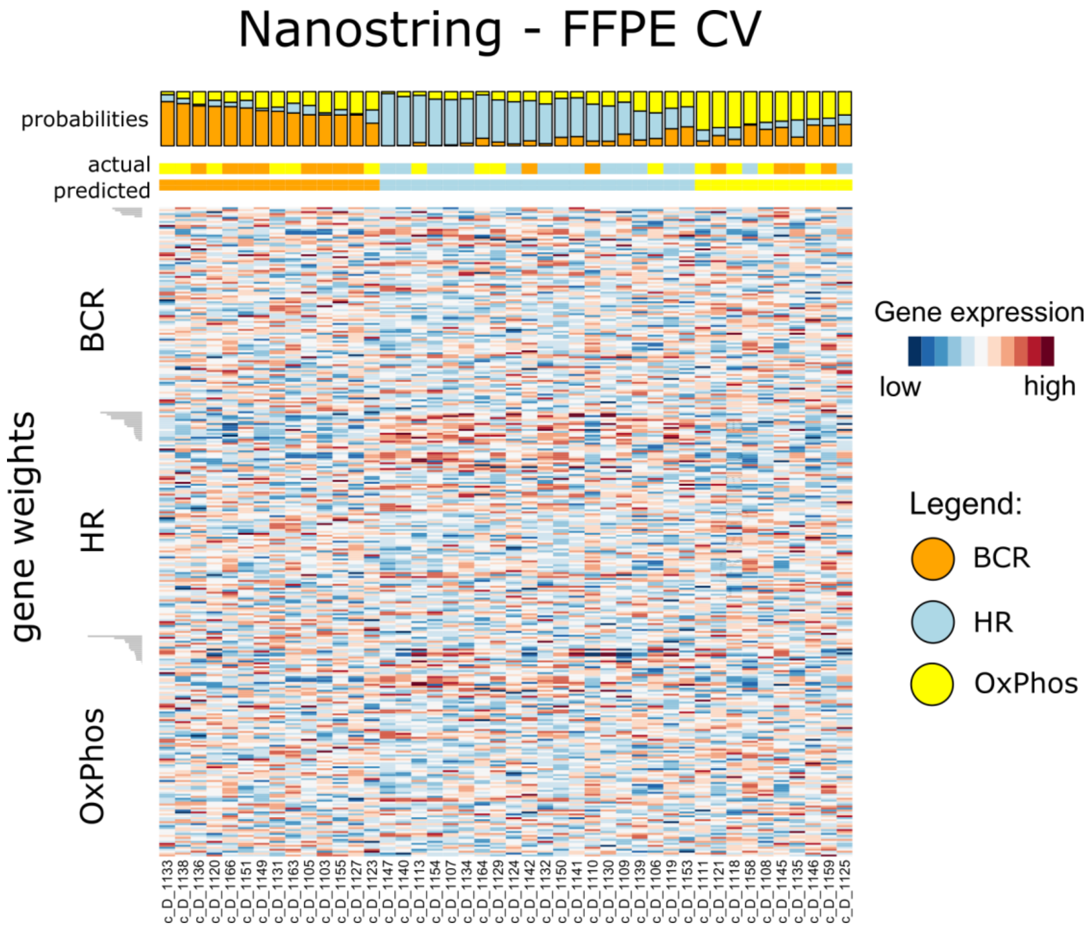


Figure III.4: Nanostring FFPE LOOCV.

The samples are ordered by class probabilities, based on the predictions of an elastic net model trained on the discovery set I. The top barplot shows the single class probabilities of the classifier, the color-bars below shows the gold standard and predicted CCC subgroups. Each row corresponds to a gene in our CCC signature, which are grouped by class and weights within the elastic net model.

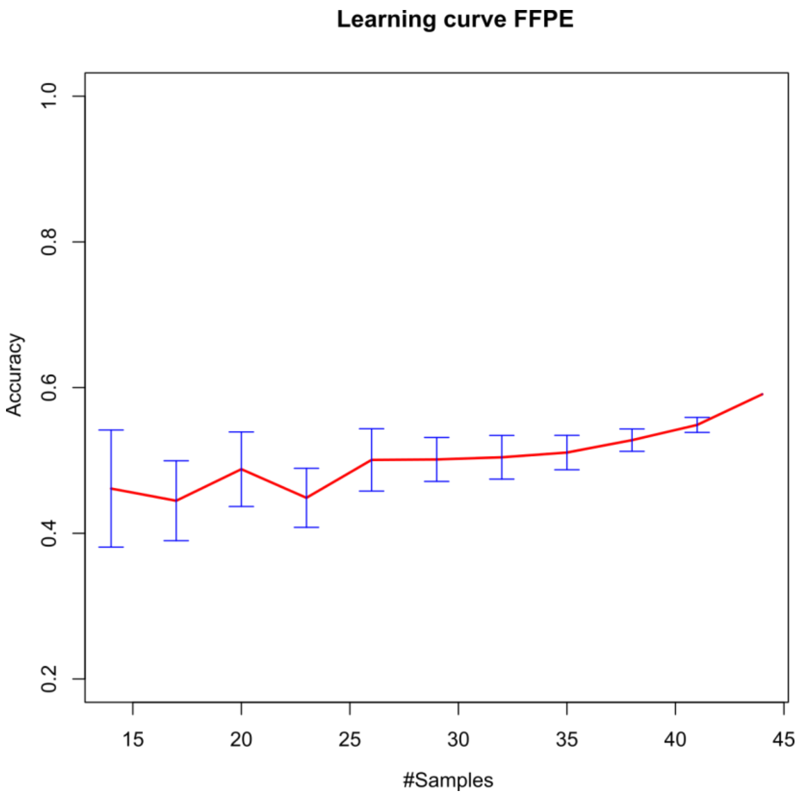


Figure III.5: Learning curves for the 44 sample Nanostring FFPE dataset.

Here we show how well classification works depending on differing sample sizes. For each increment we reran classification 50 times based on random sampled subsets. The red line shows the trend of classification performance, while the blue lines show the 95% confidence intervals based on the 50 reruns. There is a slight upward trend indicating that a larger sample size would result in a better classification performance.

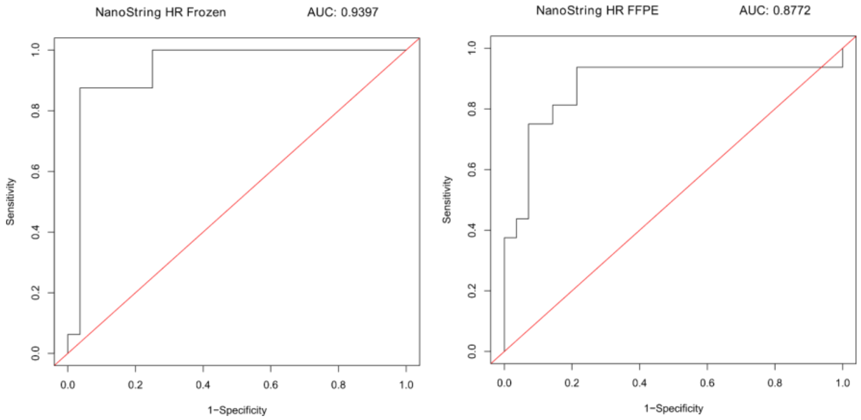


Figure III.6: HR ROC curves of HR classifier in the 44 validation cohort replicates in Nanostring.

The left shows the ROC curve within the Nanostring frozen cohort, while the right shows the results in the Nanostring FFPE. In red we show the expected performance of a random classifier.

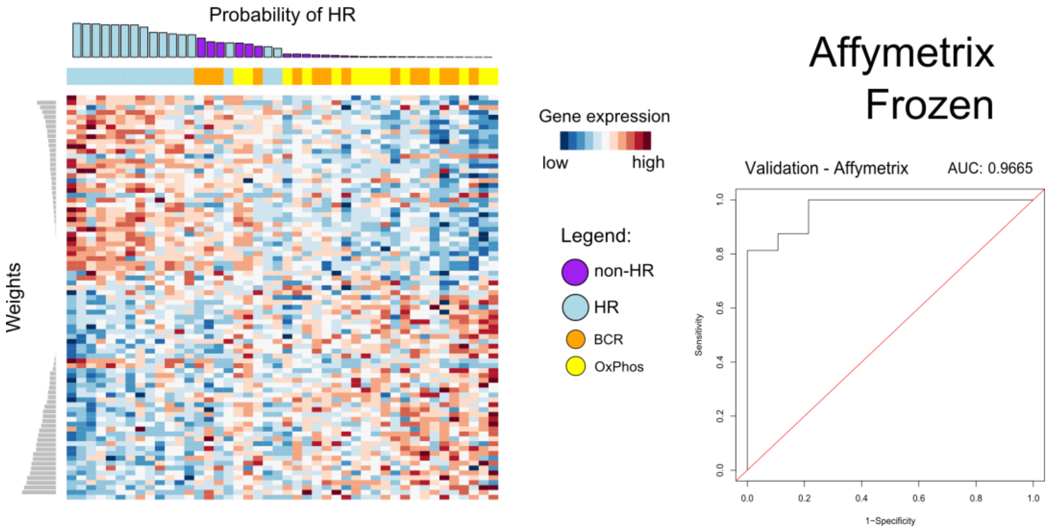


Figure III.7: HR Heatmap and ROC curve of the 44 Affymetrix replicates

Of the validation cohort in Affymetrix. On the left we show a heatmap of the Affymetrix fresh frozen samples, which are ordered by the HR class probabilities resulting from an elastic net model trained on the discovery set I. The rows show the 179 gene signature and the weights as assigned by the model.

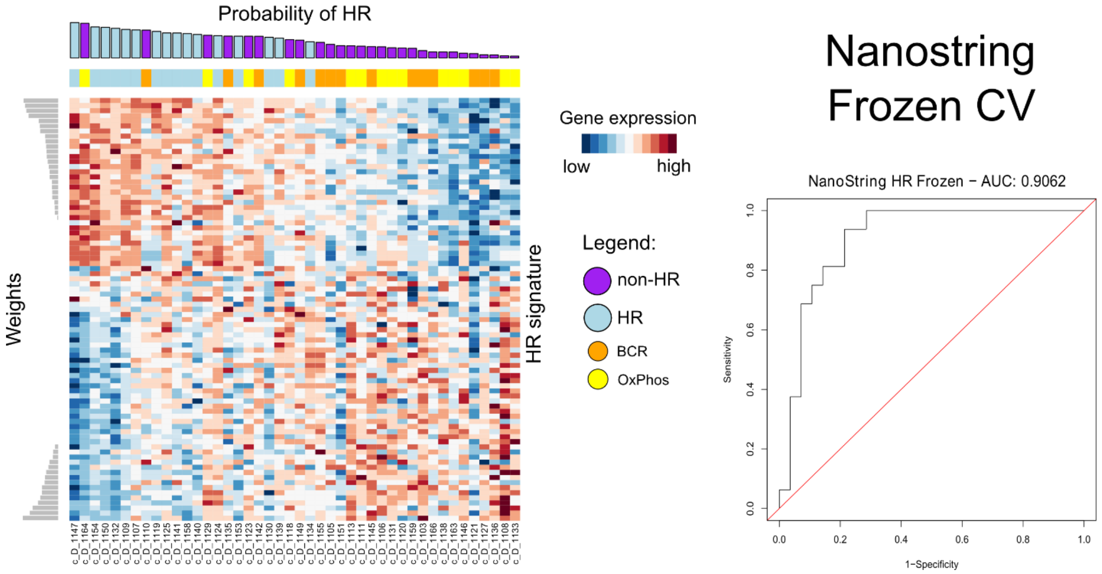


Figure III.8: HR Nanostring Frozen LOOCV.

On the left we show a heatmap of the Nanostring fresh frozen data. The samples are in the columns and are ordered by class probability derived from a leave-one-out cross-validation (LOOCV) using elastic net, while the genes in the signature are in the rows. On the right we show the receiver operating characteristic (ROC) curve of the same LOOCV results.

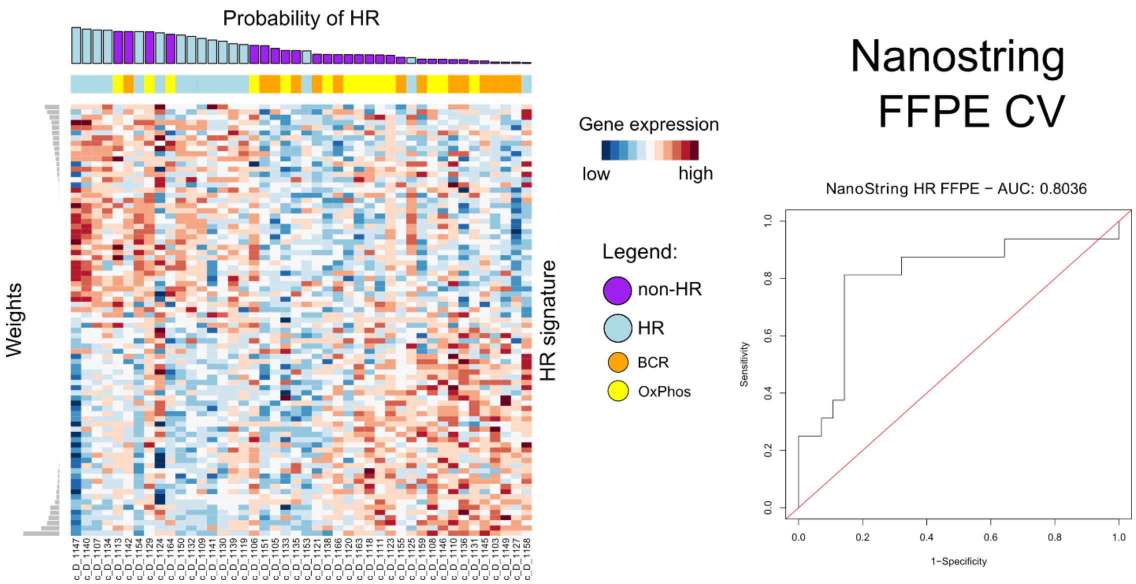


Figure III.9: HR Nanostring FFPE LOOCV.

On the left we show a heatmap of the Nanostring FFPE data. The samples are in the columns and are ordered by class probability derived from (LOOCV) using elastic net, while the genes in the signature are in the rows. On the right we show the (ROC) curve of the same LOOCV results.

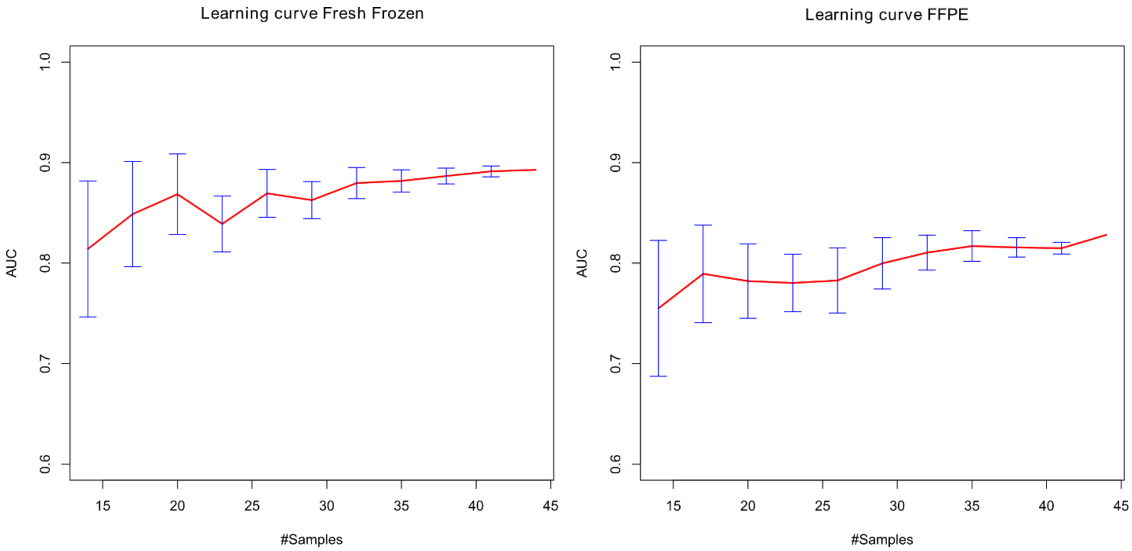


Figure III.10: HR Learning curves for Nanostring datasets.

Here we show how well classification works depending on differing sample sizes. For each increment we reran classification 50 times based on random sampled subsets. The red line shows the trend of classification performance, while the blue lines show the 95% confidence intervals based on the 50 reruns. The left shows the learning curve for the 44 Nanostring samples based on fresh frozen tissue, while the right shows the 44 replicates from FFPE tissue. In both cases, there is an upward trend indicating that a larger sample size would result in a better classification performance.

APPENDIX IV

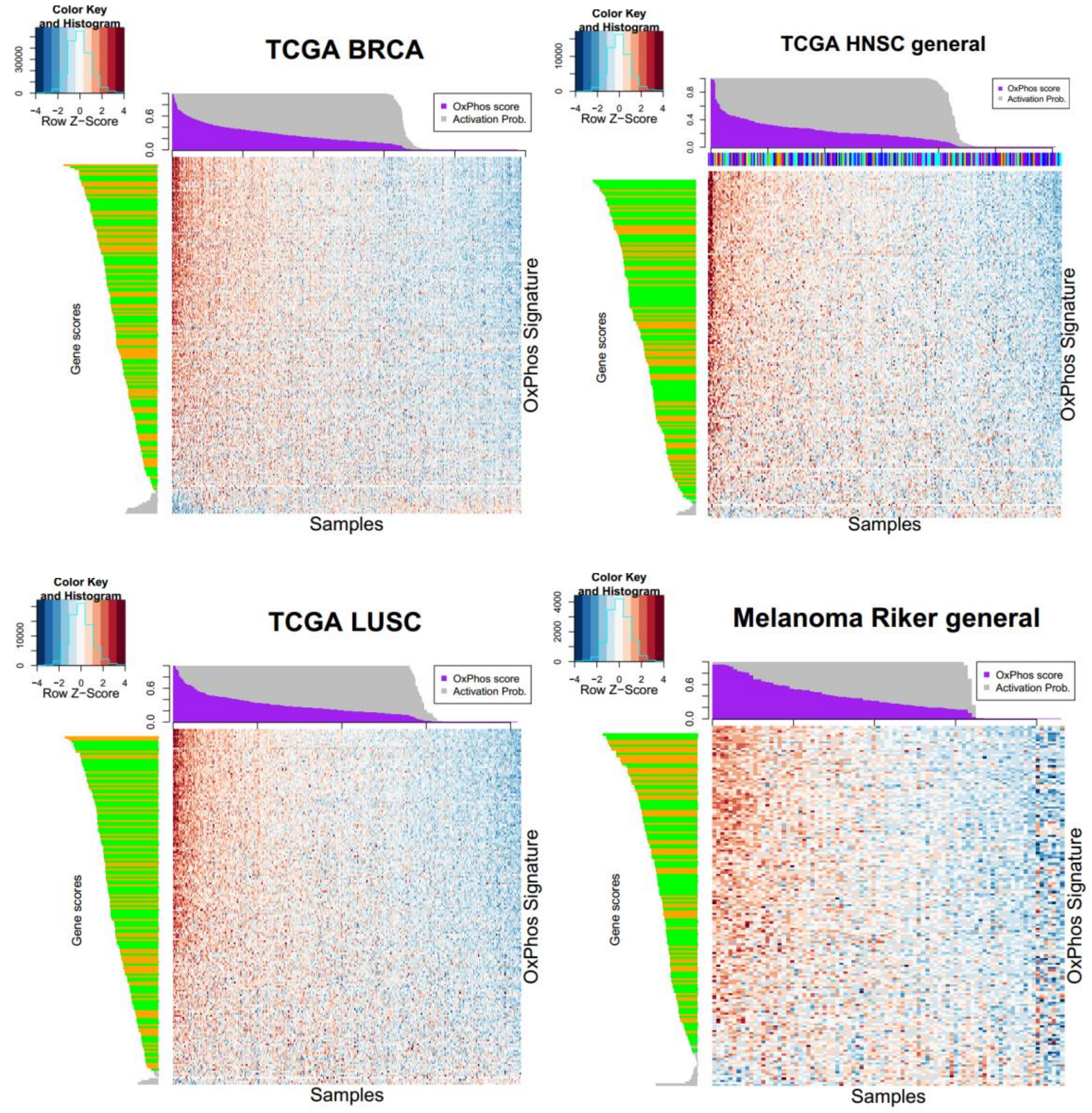


Figure IV.1: ASSIGN with generalized OxPhos signature across datasets

Red indicates up-regulated genes, whereas blue indicates down-regulated ones. The samples are ordered by OxPhos activity scores derived from an ASSIGN model, which is also shown on top in purple. On the left we show the gene weights from the ASSIGN model, in green the genes from the original DLBCL OxPhos signature, in orange the genes that were added from the generalized OxPhos gene signature and in grey the genes that have

either a negative weight or are insignificant. The top color bar on the HNSC set indicates the tumors were found.

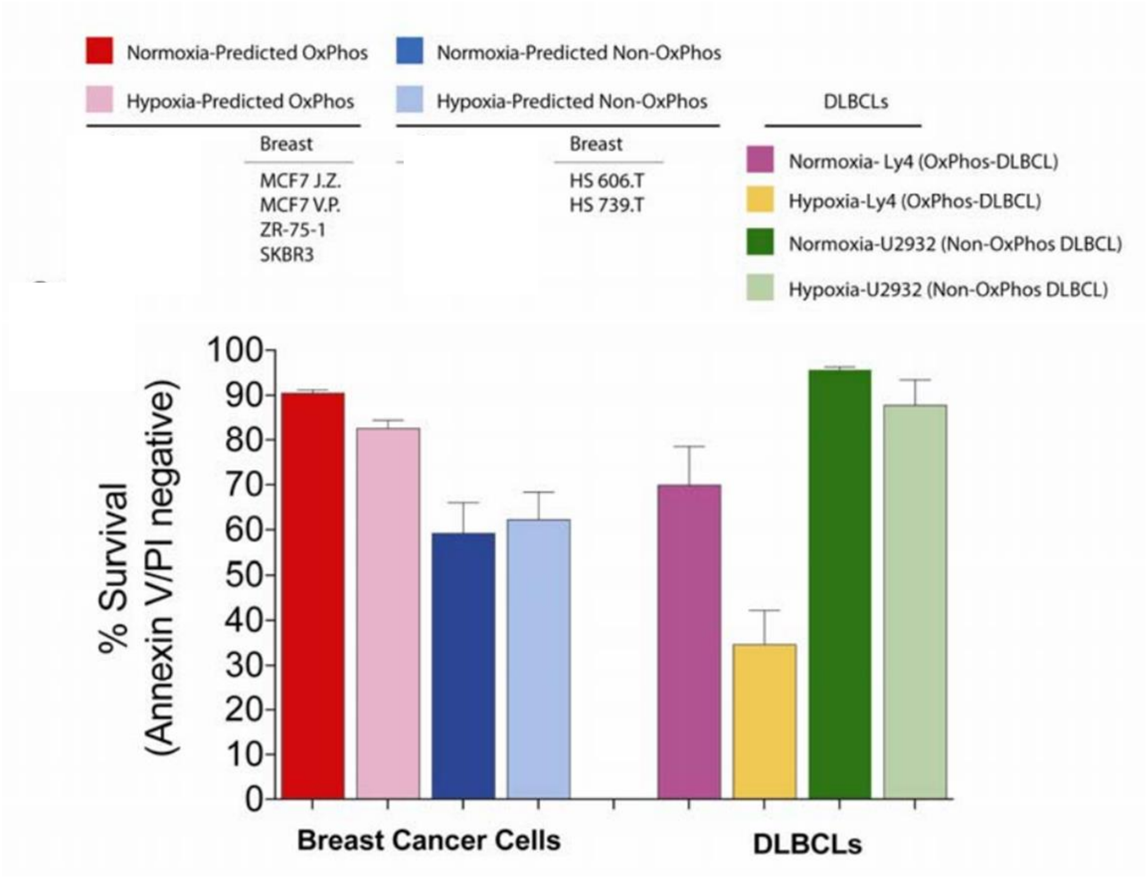


Figure IV.2: OxPhos dependency in breast cancer cell-lines

In this barplot we show the survival rate of lung and DLBCL cell-lines in dependence on oxygen availability. On the right we show the gold standard DLBCL cell-lines. Ly4 a OxPhos dependent cell-line shows a strong decrease in survival rate when comparing normal oxygen levels (purple) and hypoxia conditions (yellow), while U2932, which is not OxPhos dependent does not show a similar drop in survival rates (dark vs. light green) On the left we show the comparison between normoxia and hypoxia conditions for 6 breast cancer cell-lines that were predicted as OxPhos dependent (red) and cell-lines that were predicted as nonOxPhos (blue).

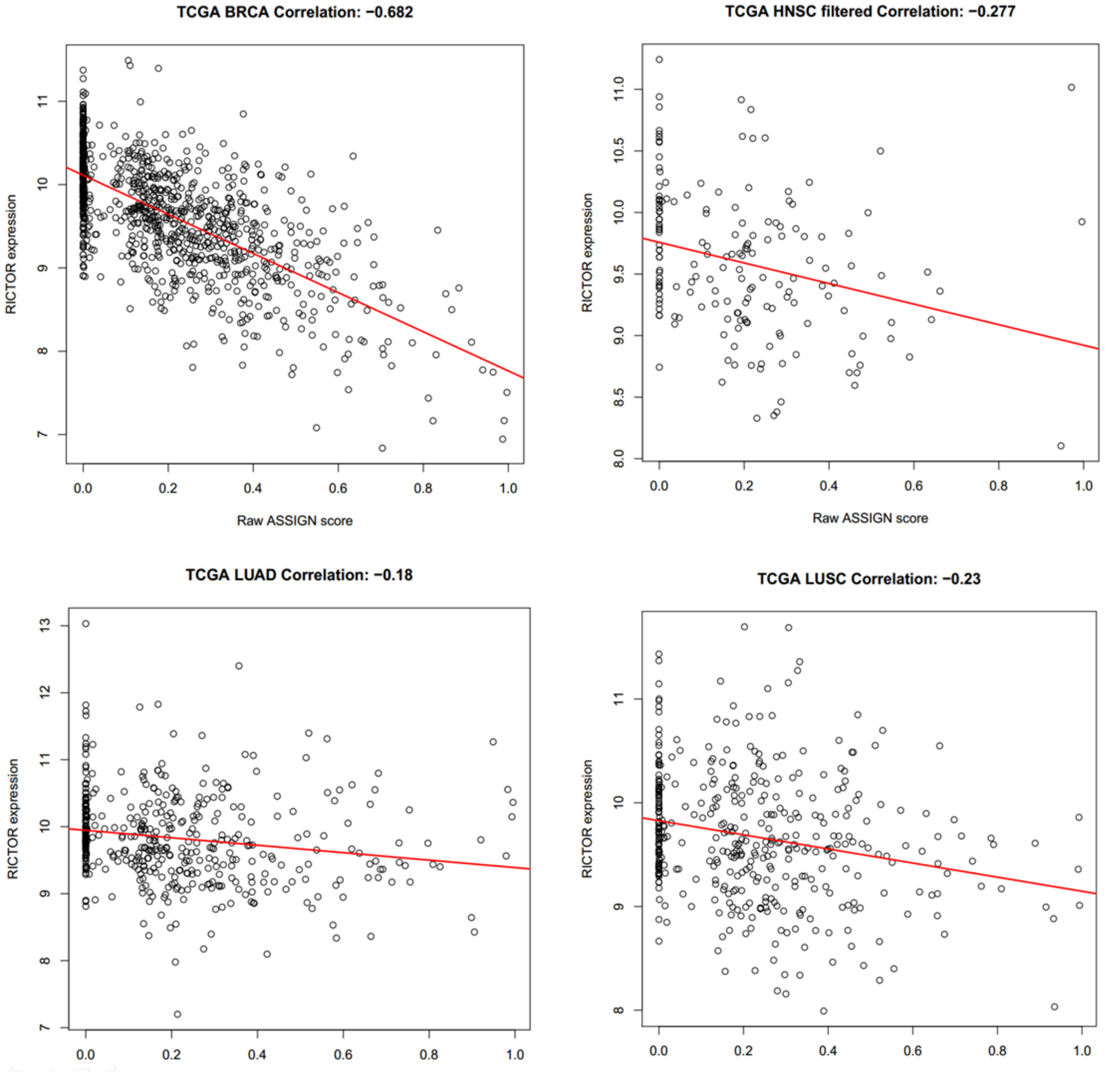


Figure IV.3: Expression of RICTOR vs. OxPhos activity across primary tumor datasets

We show the gene expression of RICTOR in comparison to OxPhos activity (Raw ASSIGN score) in TCGA BRCA (breast cancer), the oral cavity samples in TCGA HNSC (head and neck squamous carcinoma), TGCA LUAD (lung adenocarcinoma) and TCGA LUSC (lung squamous carcinoma). Red shows the trend-line based on a linear model.

Table IV.1: B-Cell receptor signaling (BCR) subtype gene signature

ensemble gene ID	gene symbol	ensemble gene ID	gene symbol
ENSG00000095585	BLNK	ENSG00000160299	PCNT
ENSG00000012048	BRCA1	ENSG00000171608	PIK3CD
ENSG00000134057	CCNB1	ENSG00000127564	PKMYT1
ENSG00000157456	CCNB2	ENSG00000197943	PLCG2
ENSG00000177455	CD19	ENSG00000099817	POLR2E
ENSG00000175216	CKAP5	ENSG00000105568	PPP2R1A
ENSG00000149532	CPSF7	ENSG00000136875	PRPF4
ENSG00000103653	CSK	ENSG00000174231	PRPF8
ENSG00000141551	CSNK1D	ENSG00000011304	PTBP1
ENSG00000204843	DCTN1	ENSG00000205937	RNPS1
ENSG00000197102	DYNC1H1	ENSG00000099995	SF3A1
ENSG00000108883	EFTUD2	ENSG00000104897	SF3A2
ENSG00000089280	FUS	ENSG00000087365	SF3B2
ENSG00000116478	HDAC1	ENSG00000189091	SF3B3
ENSG00000177733	HNRNPA0	ENSG00000104852	SNRNP70
ENSG00000122566	HNRNPA2B1	ENSG00000077312	SNRPA
ENSG00000138668	HNRNPD	ENSG00000125835	SNRPB
ENSG00000104824	HNRNPL	ENSG00000065526	SPEN
ENSG00000099783	HNRNPM	ENSG00000133226	SRRM1
ENSG00000105323	HNRNPUL1	ENSG00000149136	SSRP1
ENSG00000143772	ITPKB	ENSG00000092201	SUPT16H
ENSG00000096433	ITPR3	ENSG00000196235	SUPT5H
ENSG00000125686	MED1	ENSG00000165025	SYK
ENSG00000081189	MEF2C	ENSG00000127824	TUBA4A
ENSG00000072864	NDE1	ENSG00000196230	TUBB
ENSG00000086102	NFX1	ENSG00000131462	TUBG1
ENSG00000073969	NSF	ENSG00000137822	TUBGCP4
ENSG00000100836	PABPN1	ENSG00000166483	WEE1
ENSG00000169564	PCBP1		

Table IV.2: Oxidative Phosphorylation (OxPhos) subtype gene signature

ensemble gene ID	gene symbol	ensemble gene ID	gene symbol
ENSG00000145020	AMT	ENSG00000090266	NDUFB2
ENSG00000165629	ATP5C1	ENSG00000119013	NDUFB3
ENSG00000116459	ATP5F1	ENSG00000065518	NDUFB4
ENSG00000159199	ATP5G1	ENSG00000136521	NDUFB5
ENSG00000154518	ATP5G3	ENSG00000165264	NDUFB6
ENSG00000169020	ATP5I	ENSG00000099795	NDUFB7
ENSG00000154723	ATP5J	ENSG00000109390	NDUFC1
ENSG00000167283	ATP5L	ENSG00000023228	NDUFS1
ENSG00000100554	ATP6V1D	ENSG00000213619	NDUFS3
ENSG00000131100	ATP6V1E1	ENSG00000164258	NDUFS4
ENSG00000136888	ATP6V1G1	ENSG00000168653	NDUFS5
ENSG00000002330	BAD	ENSG00000116288	PARK7
ENSG00000015475	BID	ENSG00000168291	PDHB
ENSG00000143933	CALM2	ENSG00000168002	POLR2G
ENSG00000164305	CASP3	ENSG00000005075	POLR2J
ENSG00000166260	COX11	ENSG00000147669	POLR2K
ENSG00000131143	COX4I1	ENSG00000180817	PPA1
ENSG00000135940	COX5B	ENSG00000138777	PPA2
ENSG00000126267	COX6B1	ENSG00000177192	PUS1

ENSG00000164919	COX6C	ENSG00000146282	RARS2
ENSG00000112695	COX7A2	ENSG00000142676	RPL11
ENSG00000115944	COX7A2L	ENSG00000188846	RPL14
ENSG00000127184	COX7C	ENSG00000174748	RPL15
ENSG00000172115	CYCS	ENSG00000125691	RPL23
ENSG00000117593	DARS2	ENSG00000100316	RPL3
ENSG00000091140	DLD	ENSG00000156482	RPL30
ENSG00000140374	ETFA	ENSG00000182899	RPL35A
ENSG00000168040	FADD	ENSG00000130255	RPL36
ENSG00000168827	GFM1	ENSG00000122406	RPL5
ENSG00000178445	GLDC	ENSG00000171858	RPS21
ENSG00000196591	HDAC2	ENSG00000138326	RPS24
ENSG00000072506	HSD17B10	ENSG00000137154	RPS6
ENSG00000115317	HTRA2	ENSG00000170889	RPS9
ENSG00000114446	IFT57	ENSG00000117118	SDHB
ENSG00000011376	LARS2	ENSG00000140612	SEC11A
ENSG00000259494	MRPL46	ENSG00000120742	SERP1
ENSG00000182180	MRPS16	ENSG00000182199	SHMT2
ENSG00000175110	MRPS22	ENSG00000173638	SLC19A1
ENSG00000131368	MRPS25	ENSG00000142168	SOD1
ENSG00000144029	MRPS5	ENSG00000129128	SPCS3
ENSG00000125445	MRPS7	ENSG00000136143	SUCLA2
ENSG00000135972	MRPS9	ENSG00000163541	SUCLG1
ENSG00000103707	MTFMT	ENSG00000028839	TBPL1
ENSG00000120254	MTHFD1L	ENSG00000100416	TRMU
ENSG00000065911	MTHFD2	ENSG00000178952	TUFM
ENSG00000125356	NDUFA1	ENSG00000176890	TYMS
ENSG00000184752	NDUFA12	ENSG00000198833	UBE2J1
ENSG00000131495	NDUFA2	ENSG00000184076	UQCR10
ENSG00000189043	NDUFA4	ENSG00000127540	UQCR11
ENSG00000184983	NDUFA6	ENSG00000156467	UQCRB
ENSG00000119421	NDUFA8	ENSG00000140740	UQCRC2
ENSG00000004779	NDUFAB1	ENSG00000164405	UQCRQ
ENSG00000183648	NDUFB1	ENSG00000078668	VDAC3
ENSG00000147123	NDUFB11	ENSG00000139131	YARS2

Table IV.3: GSEA results of OxPhos vs. Non-OxPhos in Melanoma

NAME	SIZE	ES	NOM p-val	FDR q-val
KEGG GLYOXYLATE AND DICARBOXYLATE METABOLISM	15	0.695993	0.003846154	0.083679
KEGG OXIDATIVE PHOSPHORYLATION	95	0.617474	0.012345679	0.07788
REACTOME RNA POL III TRANSCRIPTION INITIATION FROM TYPE 2 PROMOTER	22	0.671559	0.001908397	0.071205
REACTOME MITOCHONDRIAL PROTEIN IMPORT	41	0.731001	0	0.066201
REACTOME TCA CYCLE AND RESPIRATORY ELECTRON TRANSPORT	95	0.663799	0.00409836	0.054375
KEGG HUNTINGTONS DISEASE	145	0.500905	0.003960396	0.052611
REACTOME GLUCOSE TRANSPORT	36	0.646394	0	0.052581
REACTOME RESPIRATORY ELECTRON TRANSPORT ATP SYNTHESIS BY CHEMIOSMOTIC COUPLING AND	62	0.666465	0.012371134	0.062736

HEAT PRODUCTION BY UNCOUPLING PROTEINS				
KEGG GALACTOSE METABOLISM	25	0.629911	0	0.058976
REACTOME RESPIRATORY ELECTRON TRANSPORT	48	0.704209	0.008163265	0.058041
KEGG PARKINSONS DISEASE	95	0.564994	0.016161617	0.053919
REACTOME TRNA AMINOACYLATION	40	0.643869	0.009433962	0.049648
REACTOME RNA POL III TRANSCRIPTION INITIATION FROM TYPE 3 PROMOTER	25	0.682143	0	0.046277
REACTOME RESOLUTION OF AP SITES VIA THE MULTIPLE NUCLEOTIDE PATCH REPLACEMENT PATHWAY	13	0.77249	0.001901141	0.046493
KEGG AMINOACYL TRNA BIOSYNTHESIS	40	0.62603	0.005628518	0.043575
KEGG PENTOSE PHOSPHATE PATHWAY	24	0.7426	0.00591716	0.061775
KEGG BASE EXCISION REPAIR	28	0.693078	0.00203252	0.069125
REACTOME RNA POL III CHAIN ELONGATION	16	0.704906	0.005725191	0.110117
REACTOME BASE EXCISION REPAIR	15	0.710969	0	0.129671
REACTOME MITOCHONDRIAL TRNA AMINOACYLATION	20	0.662973	0.013888889	0.125737
REACTOME MRNA SPLICING MINOR PATHWAY	38	0.673598	0.005802708	0.13239
REACTOME MICRORNA MIRNA BIOGENESIS	17	0.717141	0.003960396	0.129099
REACTOME INHIBITION OF REPLICATION INITIATION OF DAMAGED DNA BY RB1 E2F1	11	0.729477	0.003968254	0.123684
REACTOME RNA POL III TRANSCRIPTION	32	0.588913	0.015503876	0.121608
KEGG PURINE METABOLISM	144	0.45778	0.001937985	0.118796

Table IV.4: 85 PanOxPhos gene signature.

This table contains the ensemble gene identifiers, gene symbols and descriptions the intersect list of all tissue specific OxPhos gene signatures as derived from ASSIGN.

Ensembl Gene ID	HGNC	Description
ENSG00000152234	ATP5A1	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle
ENSG00000110955	ATP5B	ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide
ENSG00000165629	ATP5C1	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1
ENSG00000116459	ATP5F1	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit B1
ENSG00000159199	ATP5G1	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C1 (subunit 9)
ENSG00000135390	ATP5G2	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C2 (subunit 9)
ENSG00000154518	ATP5G3	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C3 (subunit 9)
ENSG00000169020	ATP5I	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit E

ENSG00000154723	ATP5J	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit F6
ENSG00000167283	ATP5L	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit G
ENSG00000117410	ATP6V0B	ATPase, H+ transporting, lysosomal 21kDa, V0 subunit b
ENSG00000100554	ATP6V1D	ATPase, H+ transporting, lysosomal 34kDa, V1 subunit D
ENSG00000131100	ATP6V1E1	ATPase, H+ transporting, lysosomal 31kDa, V1 subunit E1
ENSG00000128524	ATP6V1F	ATPase, H+ transporting, lysosomal 14kDa, V1 subunit F
ENSG00000136888	ATP6V1G1	ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G1
ENSG00000002330	BAD	BCL2-associated agonist of cell death
ENSG00000135940	COX5B	cytochrome c oxidase subunit Vb
ENSG00000126267	COX6B1	cytochrome c oxidase subunit VIb polypeptide 1 (ubiquitous)
ENSG00000164919	COX6C	cytochrome c oxidase subunit VIc
ENSG00000112695	COX7A2	cytochrome c oxidase subunit VIIa polypeptide 2 (liver)
ENSG00000115944	COX7A2L	cytochrome c oxidase subunit VIIa polypeptide 2 like
ENSG00000127184	COX7C	cytochrome c oxidase subunit VIIc
ENSG00000179091	CYC1	cytochrome c-1
ENSG00000172115	CYCS	cytochrome c, somatic
ENSG00000130159	ECSIT	ECSIT signalling integrator
ENSG00000167136	ENDOG	endonuclease G
ENSG00000140374	ETFA	electron-transfer-flavoprotein, alpha polypeptide
ENSG00000168040	FADD	Fas (TNFRSF6)-associated via death domain
ENSG00000091483	FH	fumarate hydratase
ENSG00000072506	HSD17B10	hydroxysteroid (17-beta) dehydrogenase 10
ENSG00000115317	HTRA2	HtrA serine peptidase 2
ENSG00000101365	IDH3B	isocitrate dehydrogenase 3 (NAD+) beta
ENSG00000259494	MRPL46	mitochondrial ribosomal protein L46
ENSG00000182180	MRPS16	mitochondrial ribosomal protein S16
ENSG00000175110	MRPS22	mitochondrial ribosomal protein S22
ENSG00000144029	MRPS5	mitochondrial ribosomal protein S5
ENSG00000125445	MRPS7	mitochondrial ribosomal protein S7
ENSG00000103707	MTFMT	mitochondrial methionyl-tRNA formyltransferase
ENSG00000065911	MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
ENSG00000125356	NDUFA1	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa
ENSG00000184752	NDUFA12	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 12
ENSG00000131495	NDUFA2	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 2, 8kDa
ENSG00000170906	NDUFA3	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 3, 9kDa
ENSG00000189043	NDUFA4	NDUFA4, mitochondrial complex associated
ENSG00000184983	NDUFA6	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa
ENSG00000119421	NDUFA8	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 8, 19kDa
ENSG00000004779	NDUFAB1	NADH dehydrogenase (ubiquinone) 1, alpha/beta subcomplex, 1, 8kDa
ENSG00000183648	NDUFB1	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 1, 7kDa
ENSG00000147123	NDUFB11	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 11, 17.3kDa
ENSG00000090266	NDUFB2	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 2, 8kDa
ENSG00000119013	NDUFB3	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 3, 12kDa
ENSG00000065518	NDUFB4	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 4, 15kDa
ENSG00000136521	NDUFB5	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa
ENSG00000165264	NDUFB6	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6, 17kDa
ENSG00000099795	NDUFB7	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 7, 18kDa
ENSG00000147684	NDUFB9	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9, 22kDa
ENSG00000109390	NDUFC1	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 1, 6kDa
ENSG00000213619	NDUFS3	NADH dehydrogenase (ubiquinone) Fe-S protein 3, 30kDa (NADH-coenzyme Q reductase)
ENSG00000164258	NDUFS4	NADH dehydrogenase (ubiquinone) Fe-S protein 4, 18kDa (NADH-

		coenzyme Q reductase)
ENSG00000168653	NDUFS5	NADH dehydrogenase (ubiquinone) Fe-S protein 5, 15kDa (NADH-coenzyme Q reductase)
ENSG00000115286	NDUFS7	NADH dehydrogenase (ubiquinone) Fe-S protein 7, 20kDa (NADH-coenzyme Q reductase)
ENSG00000116288	PARK7	parkinson protein 7
ENSG00000168291	PDHB	pyruvate dehydrogenase (lipoamide) beta
ENSG00000168002	POLR2G	polymerase (RNA) II (DNA directed) polypeptide G
ENSG00000005075	POLR2J	polymerase (RNA) II (DNA directed) polypeptide J, 13.3kDa
ENSG00000147669	POLR2K	polymerase (RNA) II (DNA directed) polypeptide K, 7.0kDa
ENSG00000180817	PPA1	pyrophosphatase (inorganic) 1
ENSG00000138777	PPA2	pyrophosphatase (inorganic) 2
ENSG00000177192	PUS1	pseudouridylate synthase 1
ENSG00000188846	RPL14	ribosomal protein L14
ENSG00000130255	RPL36	ribosomal protein L36
ENSG00000137154	RPS6	ribosomal protein S6
ENSG00000170889	RPS9	ribosomal protein S9
ENSG00000117118	SDHB	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
ENSG00000140612	SEC11A	SEC11 homolog A (S. cerevisiae)
ENSG00000182199	SHMT2	serine hydroxymethyltransferase 2 (mitochondrial)
ENSG00000142168	SOD1	superoxide dismutase 1, soluble
ENSG00000163541	SUCLG1	succinate-CoA ligase, alpha subunit
ENSG00000100416	TRMU	tRNA 5-methylaminomethyl-2-thiouridylate methyltransferase
ENSG00000178952	TUFM	Tu translation elongation factor, mitochondrial
ENSG00000184076	UQCR10	ubiquinol-cytochrome c reductase, complex III subunit X
ENSG00000156467	UQCRB	ubiquinol-cytochrome c reductase binding protein
ENSG00000164405	UQCRQ	ubiquinol-cytochrome c reductase, complex III subunit VII, 9.5kDa
ENSG00000139131	YARS2	tyrosyl-tRNA synthetase 2, mitochondrial

List of Abbreviated Journal Titles

ACM TIST	*ACM Transactions on Intelligent Systems and Technology*
AJSP	*The American journal of surgical pathology*
Annals	*Annals of the New York Academy of Sciences*
AR	*Anticancer research*
BG	*BMC genomics*
BJM	*British Journal of Haematology*
BMG	*BMC Medical genomics*
BP	*Biomedicine & Pharmacotherapy*
CC	*Cancer cell*
CCR	*Clinical cancer research*
CM	*Cell metabolism*
CR	*Cancer research*
CRT	*Critical Reviews in Toxicology*
CTMI	*Current topics in microbiology and immunology*
DDS	*Digestive diseases and sciences*
DJBDC	*Database : the journal of biological databases and curation*
EHP	*Environmental Health Perspectives*
ETP	*Experimental and Toxicologic Pathology*

GB	*Genome Biology*
HOCNA	*Hematology/oncology clinics of North America*
IJMS	*International journal of molecular sciences*
JAMA	*JAMA: The Journal of the American Medical Association*
JCCS	*Journal of cell communication and signaling*
JCO	*Journal of clinical oncology*
JMD	*The Journal of molecular diagnostics*
JOB	*Journal of Biotechnology*
JRSS	*Journal of the Royal Statistical Society*
LL	*Leukemia & lymphoma*
LS	*Life sciences*
MC	*Molecular carcinogenesis*
MNFR	*Molecular nutrition & food research*
MR	*Mutation research*
MRSQ	*Medical reference services quarterly*
NAR	*Nucleic acids research*
NB	*Nature biotechnology*
NCI	National Cancer Institute
NEJM	*The New England journal of medicine*
NG	*Nature genetics*
NM	*Nature Methods*

NP	*Nature protocols*
NRC	*Nature reviews. Cancer*
NRDD	*Nature reviews. Drug discovery*
NRG	*Nature Reviews Genetics*
NRI	*Nature Reviews. Immunology*
PNAS	*Proceedings of the National Academy of Sciences*
TLO	*The Lancet Oncology*
TP	Toxicologic pathology
TS	*Toxicological sciences: an official journal of the Society of Toxicology*

BIBLIOGRAPHY

Alizadeh, A.A. et al., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769), pp.503–11.

Allen, D.G. et al., Prediction of rodent carcinogenesis: an evaluation of prechronic liver lesions as forecasters of liver tumors in NTP carcinogenicity studies. *Toxicologic pathology*, 32(4), pp.393–401.

Aukema, S.M. et al., 2014. Biological characterization of adult MYC-translocation-positive mature B-cell lymphomas other than molecular Burkitt lymphoma. *Haematologica*, 99(4), pp.726–35.

Barbie, D.A. et al., 2009. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269), pp.108–12.

Barrans, S. et al., 2010. Rearrangement of MYC is associated with poor prognosis in patients with diffuse large B-cell lymphoma treated in the era of rituximab. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(20), pp.3360–5.

Barretina, J. et al., 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391), pp.603–7.

Basso, K. & Dalla-Favera, R., 2015. Germinal centres and B cell lymphomagenesis. *Nature Reviews Immunology*, 15(3), pp.172–184.

Berger, J.O., 1985. *Analysis, Statistical decision theory and Bayesian* 2nd ed., Springer.

Biroschak, J.R. et al., 2013. Impact of Oncotype DX on treatment decisions in ER-positive, node-negative breast cancer with histologic correlation. *The breast journal*, 19(3), pp.269–75.

Boobis, A.R. et al., 2008. IPCS Framework for Analyzing the Relevance of a Cancer Mode of Action for Humans. *Critical Reviews in Toxicology*, 36(10), pp.781–792.

Breiman, L., 1998. Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*.

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24(2), pp.123–140.

Breiman, L., 2001. Random Forests. *Machine Learning*, 45(1), pp.5–32.

Bucher, J.R. & Portier, C., 2004. Human carcinogenic risk evaluation, Part V: The national toxicology program vision for assessing the human carcinogenic hazard of chemicals. *Toxicological sciences: an official journal of the Society of Toxicology*, 82(2), pp.363–6.

Cai, W. & Andres, D.A., 2014. mTORC2 is required for rit-mediated oxidative stress resistance. *PloS one*, 9(12), p.e115602.

Campo, E. et al., 2011. The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood*, 117(19), pp.5019–32.

Caro, P. et al., 2012. Metabolic signatures uncover distinct targets in molecular subsets of diffuse large B cell lymphoma. *Cancer cell*, 22(4), pp.547–60.

Chang, C.-C. & Lin, C.-J., 2011. LIBSVM. *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp.1–27.

Chapuy, B. et al., 2013. Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer cell*, 24(6), pp.777–90.

Chen, L. et al., 2013. SYK inhibition modulates distinct PI3K/AKT- dependent survival pathways and cholesterol biosynthesis in diffuse large B cell lymphomas. *Cancer cell*, 23(6), pp.826–38.

Chen, L. et al., 2008. SYK-dependent tonic B-cell receptor signaling is a rational treatment target in diffuse large B-cell lymphoma. *Blood*, 111(4), pp.2230–7.

Cohen, S.M., 2010. An enhanced 13-week bioassay: An alternative to the 2-year bioassay to screen for human carcinogenesis. *Experimental and Toxicologic Pathology*, 62(5), pp.497–502.

Collisson, E.A. et al., 2014. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511), pp.543–50.

Cook, J.R. et al., 2014. Clinical significance of MYC expression and/or “high-grade” morphology in non-Burkitt, diffuse aggressive B-cell lymphomas: a SWOG S9704 correlative study. *The American journal of surgical pathology*, 38(4), pp.494–501.

Crawford, L.J., Walker, B. & Irvine, A.E., 2011. Proteasome inhibitors in cancer therapy. *Journal of cell communication and signaling*, 5(2), pp.101–10.

Curtis, C. et al., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403), pp.346–52.

Dai, M. et al., 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic acids research*, 33(20), p.e175.

Danaei, G., 2012. Global burden of infection-related cancer revisited. *The Lancet Oncology*, 13(6), pp.564–565.

Dave, S.S. et al., 2006. Molecular diagnosis of Burkitt's lymphoma. *The New England journal of medicine*, 354(23), pp.2431–42.

Davis, A.P. et al., 2013. The Comparative Toxicogenomics Database: update 2013. *Nucleic acids research*, 41(Database issue), pp.D1104–14.

Diehn, M. et al., 2009. Association of reactive oxygen species levels and radioresistance in cancer stem cells. *Nature*, 458(7239), pp.780–3.

Ellinger-Ziegelbauer, H. et al., 2008. Prediction of a carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term in vivo studies. *Mutation research*, 637(1-2), pp.23–39.

Feldman, M.E. & Shokat, K.M., 2010. New inhibitors of the PI3K-Akt-mTOR pathway: insights into mTOR signaling from a new generation of Tor Kinase Domain Inhibitors (TORKinibs). *Current topics in microbiology and immunology*, 347, pp.241–62.

Fielden, M.R. et al., 2011. Development and evaluation of a genomic signature for the prediction and mechanistic assessment of nongenotoxic hepatocarcinogens in the rat. *Toxicological sciences : an official journal of the Society of Toxicology*, 124(1), pp.54–74.

Fielden, M.R., Brennan, R. & Gollub, J., 2007. A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. *Toxicological sciences : an official journal of the Society of Toxicology*, 99(1), pp.90–100.

Fitzpatrick, R.B., 2008. CPDB: Carcinogenic Potency Database. *Medical reference services quarterly*, 27(3), pp.303–11.

Friedberg, J.W., 2008. Diffuse large B-cell lymphoma. *Hematology/oncology clinics of North America*, 22(5), pp.941–52, ix.

Friedberg, J.W., 2012. Double-hit diffuse large B-cell lymphoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 30(28), pp.3439–43.

Friedberg, J.W. et al., 2010. Inhibition of Syk with fostamatinib disodium has significant clinical activity in non-Hodgkin lymphoma and chronic lymphocytic leukemia. *Blood*, 115(13), pp.2578–85.

Friedberg, J.W. & Fisher, R.I., 2008. Diffuse large B-cell lymphoma. *Hematology/oncology clinics of North America*, 22(5), pp.941–52, ix.

Ganter, B. et al., 2005. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *Journal of Biotechnology*, 119(3), pp.219–244.

Gatenby, R.A. & Gillies, R.J., 2004. Why do cancers have high aerobic glycolysis? *Nature reviews. Cancer*, 4(11), pp.891–9.

Gebauer, N. et al., 2013. ID3 mutations are recurrent events in double-hit B-cell lymphomas. *Anticancer research*, 33(11), pp.4771–8.

Gebauer, N. et al., 2015. TP53 mutations are frequent events in double-hit B-cell lymphomas with MYC and BCL2 but not MYC and BCL6 translocations. *Leukemia & lymphoma*, 56(1), pp.179–85.

Geiss, G.K. et al., 2008. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3), pp.317–25.

Gentleman, R.C. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10), p.R80.

Gold, L.S. et al., 2005. Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997-1998. *Toxicological sciences : an official journal of the Society of Toxicology*, 85(2), pp.747–808.

Green, M.R. et al., 2010. Integrative analysis reveals selective 9p24.1 amplification, increased PD-1 ligand expression, and further induction via JAK2 in nodular sclerosing Hodgkin lymphoma and primary mediastinal large B-cell lymphoma. *Blood*, 116(17), pp.3268–77.

Green, T.M. et al., 2012. Immunohistochemical Double-Hit Score Is a Strong Predictor of Outcome in Patients With Diffuse Large B-Cell Lymphoma Treated With Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone. *Journal of Clinical Oncology*, 30(28), pp.3460–3467.

Greenbaum, D. et al., 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology*, 4(9), p.117.

Gusenleitner, D. et al., 2013. Rat Carcinogenome Portal. Available at: http://smonti.bumc.bu.edu/~montilab/Carcinogenome.

Habermann, T.M. et al., 2006. Rituximab-CHOP versus CHOP alone or with maintenance rituximab in older patients with diffuse large B-cell lymphoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 24(19), pp.3121–7.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–674.

Hastie, T. et al., 2011. pamr: Pam: prediction analysis for microarrays.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition*, Springer.

Hecht, J.L. & Aster, J.C., 2000. Molecular biology of Burkitt's lymphoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 18(21), pp.3707–21.

Vander Heiden, M.G., Cantley, L.C. & Thompson, C.B., 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (New York, N.Y.)*, 324(5930), pp.1029–33.

Heinrich Goehlmann & Talloen, W., 2009. *Gene Expression Studies Using Affymetrix Microarrays - CRC Press Book*, Taylor & Francis Online.

Herbst, R.S. et al., 2014. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*, 515(7528), pp.563–567.

Holsapple, M.P. et al., 2006. Mode of action in relevance of rodent liver tumors to human cancer risk. *Toxicological sciences : an official journal of the Society of Toxicology*, 89(1), pp.51–6.

Horn, H. et al., 2013. MYC status in concert with BCL2 and BCL6 expression predicts outcome in diffuse large B-cell lymphoma. *Blood*, 121(12), pp.2253–2263.

Howlader N, Noone AM, Krapcho M, Garshell J, Neyman N, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Lewis DR, Chen HS, Feuer EJ, C.K., 2013. *SEER Cancer Statistics Review, 1975-2010*, Bethesda, MD.

Hu, S. et al., 2013. MYC/BCL2 protein coexpression contributes to the inferior survival of activated B-cell subtype of diffuse large B-cell lymphoma and demonstrates high-risk gene expression signatures: a report from The International DLBCL Rituximab-CHOP Consortium Program. *Blood*, 121(20), pp.4021–31; quiz 4250.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1), pp.44–57.

Huff, J., Jacobson, M.F. & Davis, D.L.D.A.-J. 30 2008 D.O.-10. 1289/ehp. 1071., 2008. The Limits of Two-Year Bioassay Exposure Regimens for Identifying Chemical Carcinogens. *Environmental Health Perspectives*, 116(11), pp.1439–1442.

Hui Zou, T.H., 2005. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society*, (Series B), pp.301–320.

Hummel, M. et al., 2006. A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *The New England journal of medicine*, 354(23), pp.2419–30.

Ingolia, N.T. et al., 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature protocols*, 7(8), pp.1534–50.

Interagency Breast Cancer and Environmental Research Coordinating Committee (IBCERCC), 2013. *Breast Cancer and the Environment: Prioritizing Prevention*,

Irizarry, R.A., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2), pp.249–264.

Johnson, N.A. et al., 2012. Concurrent expression of MYC and BCL2 in diffuse large B-cell lymphoma treated with rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 30(28), pp.3452–9.

Johnson, N.A. et al., 2009. Lymphomas with concurrent BCL2 and MYC translocations: the critical factors associated with survival. *Blood*, 114(11), pp.2273–9.

De Jong, D. et al., 2007. Immunohistochemical prognostic markers in diffuse large B-cell lymphoma: validation of tissue microarray as a prerequisite for broad clinical applications--a study from the Lunenburg Lymphoma Biomarker Consortium. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 25(7), pp.805–12.

Kasprzyk, A., 2011. BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation*, 2011(0), p.bar049.

Kim, Y.H. et al., 2006. Combined microarray analysis of small cell lung cancer reveals altered apoptotic balance and distinct expression signatures of MYC family gene amplification. *Oncogene*, 25(1), pp.130–8.

Kluk, M.J. et al., 2012. Immunohistochemical detection of MYC-driven diffuse large B-cell lymphomas. *PloS one*, 7(4), p.e33813.

Krämer, A. et al., 2014. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics (Oxford, England)*, 30(4), pp.523–30.

Lamb, J., 2007. The Connectivity Map: a new tool for biomedical research. *Nature reviews. Cancer*, 7(1), pp.54–60.

Lamb, J. et al., 2006. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)*, 313(5795), pp.1929–35.

Lamond, N.W.D., Skedgel, C. & Younis, T., 2013. Is the 21-gene recurrence score a cost-effective assay in endocrine-sensitive node-negative breast cancer? *Expert review of pharmacoeconomics & outcomes research*, 13(2), pp.243–50.

Lawrence, M.S. et al., 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536), pp.576–582.

Lawrence, M.S. et al., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457), pp.214–8.

Lee Davis, D. et al., 2007. The need to develop centers for environmental oncology. *Biomedicine & Pharmacotherapy*, 61(10), pp.614–622.

Leffall, L.D. & Kripke, M.L., 2010. *President's Cancer Panel: Reducing Environmental Cancer Risk*, National Cancer Institute.

Lenz, G. & Staudt, L.M., 2010. Aggressive lymphomas. *The New England journal of medicine*, 362(15), pp.1417–29.

Liberzon, A. et al., 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England)*, 27(12), pp.1739–40.

Lichtenstein, P. et al., 2000. Environmental and Heritable Factors in the Causation of Cancer -- Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *N Engl J Med*, 343(2), pp.78–85.

Lichtenstein, P. et al., 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine*, 343(2), pp.78–85.

Linton, K. et al., Microarray gene expression analysis of fixed archival tissue permits molecular classification and identification of potential therapeutic targets in diffuse large B-cell lymphoma. *The Journal of molecular diagnostics : JMD*, 14(3), pp.223–32.

Liu, Q. et al., 2013. Characterization of Torin2, an ATP-competitive inhibitor of mTOR, ATM, and ATR. *Cancer research*, 73(8), pp.2574–86.

Locasale, J.W. & Cantley, L.C., 2011. Metabolic flux and the regulation of mammalian cell growth. *Cell metabolism*, 14(4), pp.443–51.

Lohr, J.G. et al., 2012. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 109(10), pp.3879–84.

Love, C. et al., 2012. The genetic landscape of mutations in Burkitt lymphoma. *Nature genetics*, 44(12), pp.1321–5.

Lu, C.-L. et al., 2015. Tumor Cells Switch to Mitochondrial Oxidative Phosphorylation under Radiation via mTOR-Mediated Hexokinase II Inhibition - A Warburg-Reversing Effect Y. J. LEE, ed. *PLOS ONE*, 10(3), p.e0121046.

Magrath, I. et al., 1996. Adults and children with small non-cleaved-cell lymphoma have a similar excellent outcome when treated with the same chemotherapy regimen.

Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 14(3), pp.925–34.

Maher, C.A. et al., 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234), pp.97–101.

Marx, V., 2013. Drilling into big cancer-genome data. *Nature Methods*, 10(4), pp.293–297.

Masqué-Soler, N. et al., 2013. Molecular classification of mature aggressive B-cell lymphoma using digital multiplexed gene expression on formalin-fixed paraffin-embedded biopsy specimens. *Blood*, 122(11), pp.1985–6.

McCall, M.N., Bolstad, B.M. & Irizarry, R.A., 2010. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)*, 11(2), pp.242–53.

Mermel, C.H. et al., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4), p.R41.

Meyer, D. et al., 2012. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.

Monti, S. et al., 2012. Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer cell*, 22(3), pp.359–72.

Monti, S. et al., 2005. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105(5), pp.1851–61.

Mootha, V.K. et al., 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3), pp.267–73.

Mori, S. et al., 2008. Utilization of pathway signatures to reveal distinct types of B lymphoma in the Emicro-myc model and human diffuse large B-cell lymphoma. *Cancer research*, 68(20), pp.8525–34.

Morin, R.D. et al., 2013. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood*, 122(7), pp.1256–65.

Newman, T.B., 1996. Carcinogenicity of Lipid-Lowering Drugs. *JAMA: The Journal of the American Medical Association*, 275(1), p.55.

Nie, A.Y. et al., 2006. Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity. *Molecular carcinogenesis*, 45(12), pp.914–33.

Patterson, A.D. et al., 2011. Aberrant lipid metabolism in hepatocellular carcinoma revealed by plasma metabolomics and lipid profiling. *Cancer research*, 71(21), pp.6590–600.

Perry, A.M. et al., 2014. MYC and BCL2 protein expression predicts survival in patients with diffuse large B-cell lymphoma treated with rituximab. *British journal of haematology*, 165(3), pp.382–91.

Polo, J.M. et al., 2007. Transcriptional signature with differential expression of BCL6 target genes accurately identifies BCL6-dependent diffuse large B cell lymphomas. *Proceedings of the National Academy of Sciences*, 104(9), pp.3207–3212.

Pyragius, C.E. et al., 2013. Aberrant lipid metabolism: an emerging diagnostic and therapeutic target in ovarian cancer. *International journal of molecular sciences*, 14(4), pp.7742–56.

R Core Team, 2012. *R: A Language and Environment for Statistical Computing*,

Riker, A.I. et al., 2008. The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC medical genomics*, 1, p.13.

Rimsza, L.M. et al., 2011. Accurate classification of diffuse large B-cell lymphoma into germinal center and activated B-cell subtypes using a nuclease protection assay on formalin-fixed, paraffin-embedded tissues. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 17(11), pp.3727–32.

Rummel, M., 2010. Reassessing the standard of care in indolent lymphoma: a clinical update to improve clinical practice. *Journal of the National Comprehensive Cancer Network : JNCCN*, 8 Suppl 6, pp.S1–14; quiz S15.

Salaverria, I. & Siebert, R., 2011. The gray zone between Burkitt's lymphoma and diffuse large B-cell lymphoma from a genetics perspective. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29(14), pp.1835–43.

Savage, K.J. et al., 2009. MYC gene rearrangements are associated with a poor prognosis in diffuse large B-cell lymphoma patients treated with R-CHOP chemotherapy. *Blood*, 114(17), pp.3533–7.

Schlosser, I. et al., 2005. Dissection of transcriptional programmes in response to serum and c-Myc in a human B-cell line. *Oncogene*, 24(3), pp.520–4.

Schmitz, R. et al., 2012. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature*, 490(7418), pp.116–20.

Schuhmacher, M. et al., 2001. The transcriptional program of a human B cell line in response to Myc. *Nucleic acids research*, 29(2), pp.397–406.

Scott, D.W. et al., 2014. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*, 123(8), pp.1214–7.

Shen, Y. et al., 2015. ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. *Bioinformatics (Oxford, England)*, 31(11), pp.1745–53.

Shi, L. et al., 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8), pp.827–38.

Simonetti, R.G. et al., 1991. Hepatocellular carcinoma. A worldwide problem and the major risk factors. *Digestive diseases and sciences*, 36(7), pp.962–72.

Siu, H. et al., 2011. Implication of next-generation sequencing on association studies. *BMC genomics*, 12, p.322.

Smaglo, B.G. et al., 2015. Comprehensive multiplatform biomarker analysis of 199 anal squamous cell carcinomas. *Oncotarget*.

Smyth, G., 2005. Limma: linear models for microarray data. , pp.397 – 420.

Sorensen, T.I.A. et al., 1988. Genetic and Environmental Influences on Premature Death in Adult Adoptees. *New England Journal of Medicine*, 318(12), pp.727–732.

Stewart, B.W. & Wild, C.P., 2014. World Cancer Report 2014. *IARC*, (224).

Su, Z. et al., 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9), pp.903–914.

Subramanian, A. et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp.15545–50.

Takashima, K. et al., 2006. Effect of the difference in vehicles on gene expression in the rat liver--analysis of the control data in the Toxicogenomics Project Database. *Life sciences*, 78(24), pp.2787–96.

TCGA, 2012a. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417), pp.519–25.

TCGA, 2012b. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp.61–70.

Tibshirani, R. et al., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), pp.6567–72.

Tibshirani, R., 1994. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, pp.267–288.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), pp.273–282.

Tillinghast, G.W., 2010. Microarrays in the clinic. *Nature biotechnology*, 28(8), pp.810–2.

Trachootham, D., Alexandre, J. & Huang, P., 2009. Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? *Nature reviews. Drug discovery*, 8(7), pp.579–91.

Uehara, T. et al., 2010. The Japanese toxicogenomics project: application of toxicogenomics. *Molecular nutrition & food research*, 54(2), pp.218–27.

Valera, A. et al., 2013. MYC protein expression and genetic alterations have prognostic impact in patients with diffuse large B-cell lymphoma treated with immunochemotherapy. *Haematologica*, 98(10), pp.1554–62.

Vazquez, F. et al., 2013. PGC1α expression defines a subset of human melanoma tumors with increased mitochondrial capacity and resistance to oxidative stress. *Cancer cell*, 23(3), pp.287–301.

Waggott, D. et al., 2012. NanoStringNorm: an extensible R package for the pre-processing of NanoString mRNA and miRNA data. *Bioinformatics (Oxford, England)*, 28(11), pp.1546–8.

Wang, Z., Gerstein, M. & Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), pp.57–63.

Warburg, O., 1956. On the origin of cancer cells. *Science (New York, N.Y.)*, 123(3191), pp.309–14.

Ward, P.S. & Thompson, C.B., 2012. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell*, 21(3), pp.297–308.

Waters, M.D., Jackson, M. & Lea, I., 2010. Characterizing and predicting carcinogenicity and mode of action using conventional and toxicogenomics methods. *Mutation Research/Reviews in Mutation Research*, 705(3), pp.184–200.

Weinberg, F. & Chandel, N.S., 2009. Mitochondrial Metabolism and Cancer. *Annals of the New York Academy of Sciences*, 1177(1), pp.66–73.

Weinberg, R.A., 2013. *The Biology of Cancer* 2nd ed., Garland Science.

Yu, D. et al., 2005. Functional validation of genes implicated in lymphomagenesis: an in vivo selection assay using a Myc-induced B-cell tumor. *Annals of the New York Academy of Sciences*, 1059, pp.145–59.

Zeller, K.I. et al., 2003. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome biology*, 4(10), p.R69.

Zhang, J. et al., 2013. Genetic heterogeneity of diffuse large B-cell lymphoma. *Proceedings of the National Academy of Sciences*, 110(4), pp.1398–1403.

Zhou, K. et al., 2014. C-MYC aberrations as prognostic factors in diffuse large B-cell lymphoma: a meta-analysis of epidemiological studies. *PloS one*, 9(4), p.e95020.

Zitvogel, L. & Kroemer, G., 2012. Targeting PD-1/PD-L1 interactions for cancer immunotherapy. *Oncoimmunology*, 1(8), pp.1223–1225.

Zou, H. & Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp.301–320.

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]

[redacted]