Data Integration of High-Throughput Proteomic and Transcriptomic Data based on Public Database Knowledge

Dissertation
for the award of the degree

Doctor rerum naturalium

of the Georg-August-Universität Göttingen

within the doctoral program *Molecular Biology of Cells*
of the Georg-August-University School of Science (GAUSS)

submitted by
Astrid Wachter
from
Ahlen (Westfalen), Germany

Göttingen, 2017

Thesis Committee:
Prof. Dr. Tim Beißbarth
Department of Medical Statistics,
University Medical Center Göttingen
Prof. Dr. Edgar Wingender
Department of Bioinformatics,
University Medical Center Göttingen
Prof. Dr. Christine Stadelmann-Nessler
Institute of Neuropathology,
University Medical Center Göttingen

Members of the Examination Board:
1st Referee: Prof. Dr. Tim Beißbarth
Department of Medical Statistics,
University Medical Center Göttingen
2nd Referee: Prof. Dr. Edgar Wingender
Department of Bioinformatics,
University Medical Center Göttingen

Further members of the Examination Board:
Prof. Dr. Christine Stadelmann-Nessler
Institute of Neuropathology,
University Medical Center Göttingen
Prof. Dr. Steven Johnsen
Clinic for General, Visceral and Pediatric Surgery,
University Medical Center Göttingen
Prof. Dr. Gregor Bucher
Department of Developmental Biology,
Georg August University Göttingen
Prof. Dr. Heidi Hahn
Department of Human Genetics,
University Medical Center Göttingen

Date of oral examination: 22nd of March 2017

Abstract

With the advance of high-throughput methods enabling deep characterization of the cell on different cellular layers, ideas to combine different data types for inference of regulatory processes have emerged. Such integration promises an improved molecular understanding of physiological and pathophysiological mechanisms, which aids in the identification of drug targets and in the design of therapies. Current integration approaches are based on the idea of reducing false negatives by reinforcing concordant information between datasets. In most cases optimized for a specific integration setting and data structure, these approaches are rarely accompanied by bioinformatic tools enabling researchers to work on their own datasets.

In this thesis I present the public knowledge guided integration of phosphoproteomic, transcriptomic and proteomic time series datasets on the basis of signaling pathways. This integration allows to follow signaling cascades, to identify feedback regulation mechanisms and to observe the coordination of molecular processes in the cell by monitoring temporal variation upon external perturbation. To extract these cellular characteristics the cellular layers on which the individual datasets have been generated are taken into consideration. Separate downstream and upstream analyses of phosphoproteome and transcriptome data, respectively, and subsequent intersection analysis are coupled with a combination of network reconstruction and inference methods. Graphical consensus networks and co-regulation patterns can be extracted by this cross-platform analysis. Moreover, it provides high flexibility in terms of high-throughput platforms used for data generation as analysis is based on preprocessed datasets.

On the examples of epidermal growth factor signaling and B cell receptor signaling we were able to show that the results gained by this integration method confirm known regulatory patterns but also point to interactions that were not described previously in these contexts. This is demonstrated by performing a response-specific analysis instead of the typical layer-specific analysis.

Limitations of the approach described here are linked to database-bias and -dependency, to the low temporal resolution of high-throughput measurements and to data standardization. While overcoming these issues constitutes a challenge for the whole systems biology community, the integration approach itself can be optimized in future by working with refined disease-specific and tissue-specific signaling pathway models and database entries. The presented integration method was implemented as R software package ‘pwOmics’ and made available to other researchers.

Acknowledgements

I would like to express my thanks to all the people that have supported me throughout the last years:

First of all, I would like to express my sincere gratitude to my scientific advisor Prof. Dr. Tim Beißbarth for his wide open doors whenever I faced a problem, his guidance, constructive feedback and time. Besides, I would like to thank specifically Dr. med. Annalen Bleckmann for her support, her guidance and her willingness to solve all upcoming problems. Both helped me substantially throughout the projects, provided a great work environment and also demanded and encouraged personal development.
Many thanks go to my thesis committee members, Prof. Dr. Edgar Wingender and Prof. Dr. Stadelmann-Nessler for their helpful and constructive feedback and their commitment and support.

I would like to acknowledge all collaborators for their project contributions and their scientific enthusiasm, which I consider essential to foster good outcomes. In particular, I would like to express many thanks to the collaborators who provided the valuable data sets which I could integrate as part of this thesis.

In addition, I would like to thank all the members of the Department of Medical Statistics, especially Prof. Dr. Tim Friede and Dorit Meyer.
Many thanks go to my colleagues, who provided a great scientific and interpersonal environment: Silvia, Frank, Andreas, Klaus, Stephan, Manuel, Alex, Jochen, Michaela, Xenia, Julia, Maren, Florian, Saynab. Apart from fruitful discussions, good scientific advices and motivating feedback, it was simply a pleasure to work with them. Special thanks go to Annalen, Silvia, Michaela and Julia for ladies' nights and their never-ending willingness to proofread, discuss problems and provide support in any direction.

Furthermore, I would like to thank all the participants and organizers of my mentoring programmes, my mentoring group and current mentor, for their open ears, feedback, time and constructive and helpful advice. Many thanks also go to all the workshop organizers of these programmes, as well as the GGNB team and the GGNB course organizers for their great personal commitment. All these structures provided additional assistance and help as they are brought to life by great people.

Last but not least I would like to thank my friends and family for their patience and permanent support.

Contents

List of Figures

List of Tables

Abbreviations

HMEC	-	human mammary epithelial cells
TFs	-	transcription factors
bp	-	base pairs
PDF	-	probability distribution function
EGF	-	epidermal growth factor
DNA	-	desoxyribonucleic acid
RNA	-	ribonucleic acid
cDNA	-	complementary desoxyribonucleic acid
cRNA	-	complementary ribonucleic acid
RNA-Seq	-	RNA sequencing
RPPA	-	reverse phase protein arrays
MS	-	mass spectrometry
mRNA	-	messenger ribonucleic acid
ODE	-	ordinary differential equation
DBN	-	Dynamic Bayesian network
BCR	-	B cell receptor
EGFR	-	epidermal growth factor receptor
Ig	-	immunoglobulin
BL	-	Burkitt lymphoma
MMPs	-	matrix metalloproteinases

1 Introduction

To enhance the understanding of diseases and advance therapy approaches, decoding individual molecular interactions is crucial. Disease or changed environmental conditions lead to complex cellular processes that take place on different molecular levels. The interplay of these different levels is finely balanced and any intervention should be considered carefully to prevent molecular imbalance.

Our understanding of cellular processes and molecular interactions grows continuously, starting with individual detailed experimental work, which is today supported in parallel by an increased usage of high-throughput technologies. The high amounts of produced data enable a very comprehensive analysis of the investigated cellular state and are a big step towards a better understanding of cellular molecular reactions (Reuter et al., 2015; Larance and Lamond, 2015). However, with the increased creation of large high-throughput data sets there is a high demand of analysis tools and analysis pipelines.

With technological advances driven forward on each of the molecular levels in the cell, the available options to link data of multiple data types grow alike. This 'omics space' is currently investigated actively, as combined high-throughput data sets from different regulatory levels consequentially provide more information with regard to the complexity of biological processes than a single data set from just one regulatory level.

The term 'data integration' itself is used in a very broad context since the emergence of systems biology and systems medicine, as it evokes questions to be addressed on different levels of data handling and analysis. The main two utilizations of this term in the context of high-throughput expression data include i) linking different data types and disparate data sources with a focus on infrastructure in combined repositories: This includes linking of query-interfaces, resolving semantic problems via ontology-based integration and cross-referencing and requires benchmark information from different data types. Such infrastructure is out of the scope of this thesis, but strongly needed for the implementation of the second utilization of the term 'data integration' referring to ii) understanding the biological principles: This includes interlinking of heterogeneous high-throughput/low-throughput data sets from different platforms and combination with further biological information, e.g. biological signaling pathways. In this thesis, the term 'data integration' is used in line with the second meaning.

Though the general idea of cross-platform integration is fairly straightforward and technical prerequisites are improving constantly, there are a number of challenges that need to be

overcome when working with diverse data types. Besides infrastructural issues like data discovery problems, standardization of the experimental design and preprocessing steps (data generation routines), experimental annotations are of high relevance for the development of data integration approaches. A significant step towards a clear data annotation standard was e.g. the proposal of a Minimum Information about a Microarray Experiment (MIAME) by Brazma et al. (2001) for microarray data or the Minimum Information about a Proteomics Experiment (MIAPE) (Taylor et al. (2007)). This minimum information includes e.g. the experimental design, the array design, sample preparation and labeling, hybridization procedures and parameters, measurement specifications and normalization control types, their values and their specification. Unfortunately, such a clear data information standard is not common practice yet. Further challenges are e.g. formatting differences between data types, expert terminology, missing data, data not properly entered, merging of data with ambiguity issues or the need of very different experimental and data analysis expertise.

According to Kristensen et al. (2014) the three general objectives of data integration approaches in terms of systems medicine are

1. Understanding molecular mechanisms, relationships between and within different types of molecular structures: Only a deeper, cross-linked information throughout the different molecular structures can provide a view as complete as possible on disease and normal phenotype. Even though it is debatable whether we might get a complete view on cells in future, the emerging challenge is clearly the high number of individual phenotypes and their corresponding characterization.

2. Therefore it is necessary to perform disease subtyping with a focus on personalized medicine. With an improved characterization of the subtypes on each molecular level and clinical annotations of the patients falling into particular classes, it is possible to optimize treatment options in terms of 'personalized medicine'.

3. Prediction of outcome or phenotype for prospective patients: The knowledge gathered in the previous point can be used to classify patients prospectively via risk scores (such as Sankt Gallen risk categories for breast cancer patients from Goldhirsch et al. (2007)). This enables a direct estimation of optimal therapy based on parameters known early on.

In this thesis, I will focus on the first point: Understanding molecular mechanisms by integration of high-throughput proteomic and transcriptomic data sets, as this data inherently contains information about the precisely coupled multi-layer regulations taking place in the cell. The scope of this thesis is to interlink time-resolved gene and protein expression data sets to generate a more detailed understanding of molecular signaling processes. With this aim I developed a methodology for pathway-based data integration and implemented this approach in an open-source software package. Furthermore, I analyzed and evaluated molecular interactions identified by the proposed method in a data set comprising time series proteomic and transcriptomic data of epidermal growth factor (EGF) signaling in

human mammary epithelial cells (HMEC). In a second data set on B cell receptor (BCR) stimulation I refined the approach with the aim to track individual signaling axes in the cell.

In this introduction, I first present the main characteristics of cellular signaling pathways and different high-throughput expression data sets (see Section 1.1 *Cellular signaling* and Section 1.2 *High-throughput expression data*). Afterwards, I present the motivation for cross-platform integration, the biological rationale as well as the challenges and approaches for integration of these different types of data (see Section 1.3 *Cross-platform integration of transcriptomic and proteomic data*). Furthermore, I address the dynamic aspect and its impact on identification of molecular mechanisms, shortly introducing time course data analyses concepts (see Section 1.4 *Exploring molecular dynamics via time series data*). Subsequently, I introduce biological databases as a means for cross-platform data integration (see Section 1.5 *Biological knowledge resources*). A review on the biological pathways adressed in this thesis I will give in Section 1.6 *Investigated signaling pathways*. Section 1.7 *Objectives and overview* will provide a summary on the aims and the structure of this thesis.

1.1 Cellular signaling

Environmental stimuli, e.g. temperature changes, hormones or antigenes typically induce a cellular reaction that is needed for adaptation processes. In case of extracellular stimulatory molecules, these are sensed by receptors which are integral transmembrane proteins. With the binding of ligand molecules to these receptors conformational changes are triggered and further signal propagation is initiated through signaling cascades. The signaling pathway itself triggers transcription factors (TFs) to enter the cell nucleus and bind to specific regions on the desoxyribonucleic acid (DNA). Thereby, the rate of transcription is changed. This process itself can be dependent on the recruitment of further factors to build up specific protein complexes. Figure 1.1 shows these cellular processes schematically in a simplified way.

These cascades require a complex and finely balanced network of enzymes, small molecules and second messenger molecules, which depends on various factors itself, e.g. gene expression. Many signaling pathways have been characterized in detail, especially those associated with specific diseases. This is due to the perspective that with increased knowledge of signaling transduction cascades chances to understand non-physiological signaling and treatment options are higher. A good example are kinase proteins, responsible for phosphorylation processes in the cell, which define the activity, reactivity and binding characteristics of molecules (Hunter, 1995). However, it is not clear whether the characterization of a pathway is ever complete or if there are still unknown pathway members, given that there is a pathway overlap in a considerable number of pathways and cross-pathway signaling. Therefore, also feedback loops have an intricate and fundamental influence on cellular systems.

There are commercial and open-source biological databases which form a resource of knowledge for the described processes. The ones used in this thesis are shortly described in Section 1.5 *Biological knowledge resources*.

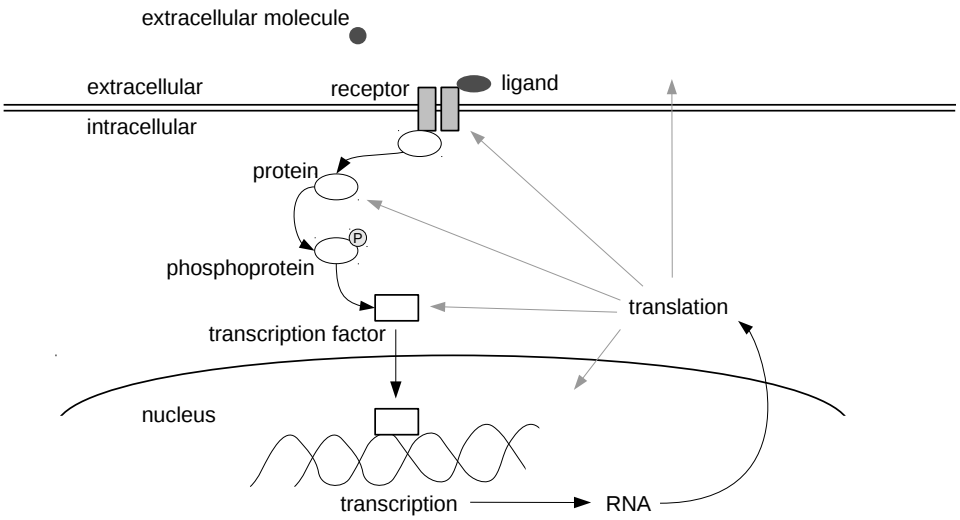


Figure 1.1. *Simplified schematic of cellular processes. Extracellular molecules can be ligands of membrane receptors, triggering a signaling cascade throughout the cell. By phosphorylation the three-dimensional structure of phosphoproteins can be changed, leading to a modification of their function. Via the signaling cascade transcription factors can be triggered to move into the nucleus and affect transcription. The transcribed RNA is needed for protein translation at ribosome sites in the cytoplasm, leading to a feedback on the signaling itself.*

The cellular systems mentioned above naturally depend on finely coordinated temporal and spatial processes, which makes it impossible to fully characterize them with just one single measurement. Time-resolved measurements can portray those processes considerably better by adding another dimension to the data collection. Section 1.4 *Exploring molecular dynamics via time series data* adresses analysis concepts for such data sets.

1.2 High-throughput expression data

Gene expression data and protein expression data are often used to characterize molecular differences between different biological settings. These data types provide information on different molecular levels of the cell: According to the central dogma of molecular biology (Crick, 1970) information transfer takes place in a sequential way from DNA to ribonucleic acid (RNA) and from RNA to protein, as depicted in Figure 1.2. While gene expression data, also commonly referred to as transcriptomics data, gives an idea about RNA abundance levels, protein expression data reflects the functional state of the cell by representing protein abundance levels.

In a simplified representation the information flow thus starts with the information encoded on the exonic regions of the genes, which is then transcribed to RNA. The RNA leaves the cell nucleus in order to deliver information to the translational process, in which functional proteins are generated by ribosomes. However, when considering all possible regulatory influences on the different stages of the cellular machinery, this process is very

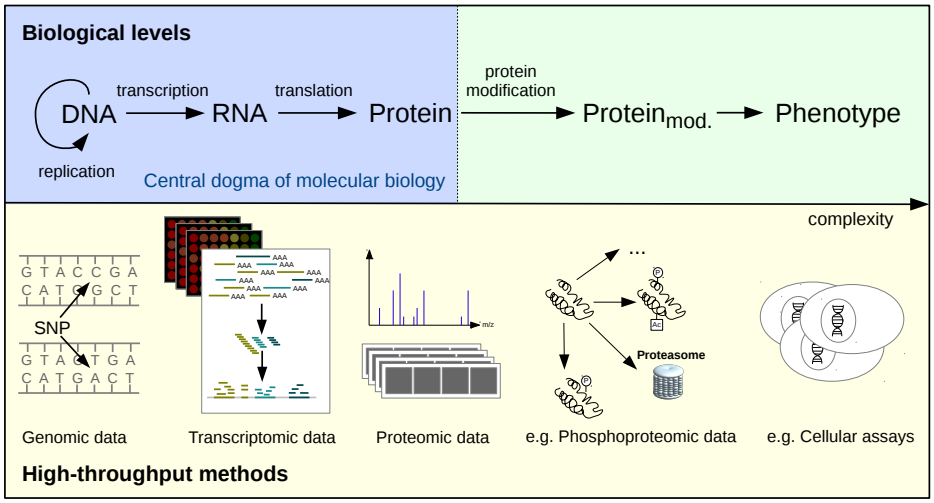


Figure 1.2. *Biological and high-throughput data generation levels. Shown on blue background are the different biological levels reflected in the central dogma of biology, being functionally connected in a linear way. Further biological levels are shown on green background. Following these levels implicates an increase in biological complexity. High-throughput methods for each of the biological levels are available, such that level-specific characteristics can be determined. However, integrating data generated by different high-throughput methods is still a challenge.*

complex. Yet the advance of very sensitive high-throughput techniques generating gene and protein expression data enables a deep characterization of the cellular states. To identify regulations on the different molecular levels a number of omics technologies have been developed within the last years, enabling identification of numerous interactions. Increasing demand for such data sets reduces the costs of data generation in turn.

In this thesis, I focus on the integration of transcriptomic and proteomic data as the corresponding molecular levels, RNA and proteins, are widely measured.

1.2.1 Transcriptomic data

Transcriptomic profiling has this far been possible mainly by DNA microarrays and now increasingly by RNA sequencing (RNA-Seq), since the high demand has driven forward next-generation sequencing technologies. The latter provides higher quality, enabling an unbiased detection of novel transcripts, offering a broader dynamic range, increased sensitivity and specificity and easier detection of low-abundance transcripts. Microarray data, though, is less expensive to generate, easier to process and less challenging in terms of storage (Wang et al., 2009; Zhao et al., 2014).

DNA microarrays for expression measurements contain a high number of fixed DNA spots of specific sequences, known as probes, attached to a glass slide. These hybridize specifically to usually fluorescently labeled complementary desoxyribonucleic acid (cDNA) or complementary ribonucleic acid (cRNA) which is prepared from a sample. This reaction is

detected and quantified to determine abundances of nucleic acid sequences in the sample and consecutively differential expression between different samples. Data preprocessing includes steps of background correction, summarization and log transformation, as well as quality control and normalization steps.

For RNA-Seq the RNA of a sample is converted to a cDNA fragment library, containing adaptors on one or both ends. During sequencing short sequences from one end or both ends are obtained (single-end sequencing vs. paired-end sequencing) by sequential hybridization readout. This results in read lengths of typically 30-400 bp. The reads are subsequently aligned to a reference genome or reference transcriptome, or assembled de novo in case no reference information is available. The higher the sequencing coverage, the better the detection of rare transcripts is. While data measured on a microarray is restricted to probes on the array, RNA-Seq provides an exhaustive view on the transcriptome present in the sample.

1.2.2 Proteomic data

Similar to transcriptomic techniques, proteomic high-throughput techniques have gone through an important development during the last years. Main techniques used in this field comprise antibody-based reverse phase protein arrays (RPPA) and mass spectrometry (MS). Unlike transcriptomic data, these data sets allow for functional profiling as they reflect the proteomic state in the cell.

RPPAs are protein arrays which constitute a reverse method compared to usual microarrays, as the samples, in this case cellular lysates, are directly spotted on nitrocellulose coated glass slides. For measuring the expression of multiple proteins a series of identical slides is spotted. The slides are incubated with antibodies which bind specifically to the proteins of interest. In a second round of incubation, another labelled antibody binds to the first antibody and thereby provides a means to measure the primary binding reaction. Detection can be based on chemiluminescence, fluorescence or colorimetric assays. The obtained data is preprocessed and used for quantification. Data quality is highly dependent on good antibody binding properties, which are assessed prior to incubation via western blot.

For MS the sample is ionized in order to retrieve charged fragments of the sample's molecules. These ions are ordered according to their mass-to-charge ratio by applying an electric and/or magnetic field. Usually, detection is performed by an electron multiplier or any device that can measure charged particles. Relative abundance of the detected ions can be displayed in so-called mass spectra as a function of the mass-to-charge ratio. Via database matching the measured spectra can be assigned to specific molecules. When used for protein expression measurements, the proteins of a sample are fragmented to peptides, which can be identified in the last step as part of specific proteins.

While RPPA provides a better throughput in terms of samples, MS can cover almost all proteins that are technically detectable via a sequence comparison with corresponding databases. Both techniques can provide additional information on protein phosphorylation

taking place post-translationally, yet they use different approaches. RPPA employ antibodies to detect the phosphorylation on the protein, while MS measures the change of mass-to-charge ratio to identify corresponding mass changes. As protein phosphorylation gives a lot of information on cellular activity, it is valuable information when interpreting cellular signaling.

While RPPA data is restricted to the selected antibodies, MS data can provide a whole view on the cellular proteome and is restricted only to technical sensitivity. However, MS has higher costs per sample when the aim is multiple sample profiling and there are still limitations in detecting proteins which are only present in low abundances.The presented proteomics techniques provide relative protein abundance values, compared to transcriptomics data measurements which provide absolute values such as read counts or fragments per kilobase of transcript per million mapped reads. However, this issue can be tackled e.g. by reference sample measurements when being addressed during the experimental design phase.

1.3 Cross-platform integration of transcriptomic and proteomic data

Cross-platform integration is a tempting approach when aiming to assess or to dispose - at least partly - of technical biases that are inherent to the different measurement techniques. Furthermore, high rates of false positives and false negatives can be addressed by reinforcing concordant information (Hwang et al., 2005). Data integration constitutes a very elegant way to not think in measurement systems, such as gene expression, protein-protein interaction assays or else, but to think in causal chains of effectors and effects, with these being measured by different means. Opening up a multi-dimensional space in terms of multiple data types and then reducing the dimensionality of information about the system of interest, it prevents from thinking horizontally on one level of measurement only, and thus allows for a deeper comprehension of systems biology. Promising examples are functional-linkage networks, protein function prediction from heterogeneous data or patient-specific data integration (Gligorijević and Pržulj, 2015).

Yet, these ideas are confronted with a lot of challenges when considering the practical implementation of cross-platform data integration. Further consideration of an optimal data integration methodology is needed when time-series data from different platforms are assessed. Their benefit for characterization of molecular processes is specified in Section 1.4 *Exploring molecular dynamics via time series data.*

1.3.1 Underlying biological rationale

Integration of proteomic and transcriptomic data poses an interesting question as it links the two 'product' layers of the central dogma of molecular biology. In addition, there is a large number of different regulatory mechanisms taking place on or between these layers. A non-exhaustive overview of these influences is depicted in Chapter 3 *Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration*

Approach for Proteomics and Transcriptomics Data, Figure 2. Such molecular regulation is physiologically occurring post-transcriptionally and might be deregulated in case of disease. It ranges from chromatin state dynamics over alternative splicing to post-translational modifications of proteins and effects reflected in cellular signaling cascades, such as feedback loops. On account of this, we do not observe perfect correlations when comparing protein expression with gene expression data.

As a reason for low correlation between protein and gene expression data, Haider and Pal (2013) discuss the following influences (to be considered as non-exhaustive):

- post-transcriptional modifications,
- translational efficiency (ribosome density, occupancy time in ribosomes),
- external factors (e.g. temperature),
- codon-bias (multiple number of codons translate the same amino-acid),
- variability of messenger ribonucleic acid (mRNA) expression levels during cell cycle,
- different half-lives of mRNA and proteins,
- experimental error.

The aim of integrating these two data types is therefore defined as identifying certain regulatory effect patterns. Certainly, a specific determination of a regulatory origin is hard to obtain at this stage of integration. Yet, extending the integration towards additional data types following the idea of systems integration might enable such specific assignments eventually.

1.3.2 Challenges for implementation

Challenges arising when addressing the implementation of data integration are various. Initially, it is of considerable importance to start with high quality data to prevent false assumptions downstream in the integration process: First, defining a significance threshold is problematic as this has to be dependent on the specific integration method of choice. Second, it needs to be decided at which level of information the integration should optimally be performed. Here, data reduction is an option in order not to stumble across limited statistical power when integrating higher data dimensions.

Another issue which needs to be tackled when data-driven methods in high-dimensionality problems are used is overfitting. With analyzing multiple different data sets the risk of trusting false positive results is increased. To reduce false discoveries in expression data, the gold standard is searching for replication of results in independent data sets. However, finding independent data sets analyzed in the same integrative manner is very challenging or even impossible.

Another issue needing attention even prior to comprehensive integration is confounding factors in the individual data sets. Though usually a number of additional variables is assessed, there might also be sources of signal due to unknown or unmeasured variables. This phenomenon is already a problem in well-designed studies (Leek and Storey, 2007), and its

effects accumulate during data integration. Therefore, disregarding this issue might also lead to misinterpretation.

A further point of consideration is how to biologically account for the different molecular layers the data is based on: Does biological variation in data from a certain molecular level has the same meaning as biological variation in another data type? Is a normalization step necessary? Does it biologically make sense to use the same structures during data reduction for different data types?

To summarize, there are many challenges that need to be addressed on the way to powerful integrative analysis. The integration method itself will still have to clearly depend on the ultimate goal of the analysis.

1.3.3 Integration approaches

Data integration approaches in general follow two different hypotheses (Ritchie et al., 2015), as depicted in Figure 1.3:

1. integration is performed reflecting variation hierarchically in a linear manner, i.e. from DNA to RNA to proteome to phenotype, or
2. integration considers the combination of variation across all omics levels leading to a specific phenotype.

The method of data reduction and the order of processing needs to be chosen accordingly. When different data types are integrated, three possible approaches have been described so far: concatenation-based, transformation-based and model-based approaches (Ritchie et al., 2015). Concatenation-based approaches link the different data types on a raw data level or a pre-processed data level, while in transformation-based approaches the data from different platforms undergoes an individual analysis and transformation process first. In model based integration approaches each data type builds the basis for an independent model before the integration process results in a combined model.

Table 1.1 gives a non-exhaustive overview on diverse data integration approaches and tools for transcriptomics and proteomics data integration.

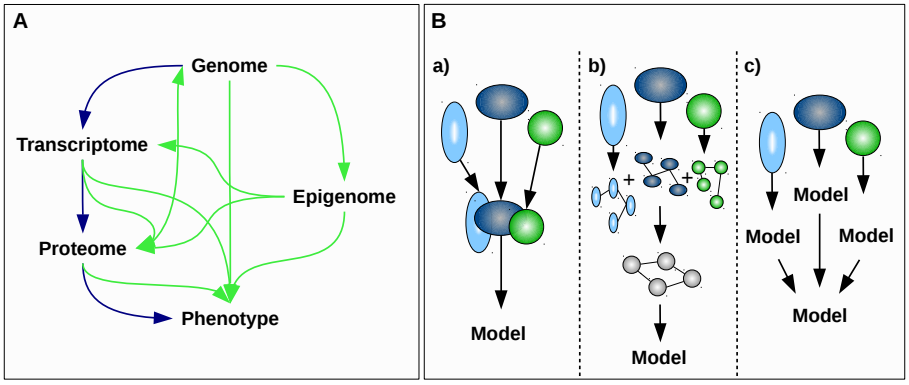


Figure 1.3. *Data integration approaches of biological high-throughput data. A: Data integration reflects the hierarchical variation in the data linearly (indicated with blue arrows) or is performed according to the combination of variation across all omics levels (indicated with green arrows). B: Data integration approaches described so far. a) Concatenation-based integration. b) Transformation-based integration. c) Model-based integration. Different colors represent different data types. Figure adopted from Ritchie et al. (2015).*

Integration approach [Class]	Hypothesis/ Scope	Tool	References [Data origin]
Simple union of transcriptomic and proteomic data resulting in a reference data set [C]	Limitations of one method (e.g. microarray probe set bias) are reduced	-	Delmotte et al. (2010) [B], Altenbach et al. (2010) [W], McRedmond et al. (2004) [H]
Extraction of common features or common functional context of features [T]	Same functional context (e.g. pathways) on transcriptomic and proteomic level	omicsNET IMPaLA (Kamburov et al., 2011) iPEAP (Sun et al., 2014)	Perco et al. (2010) [H], Com et al. (2012) [R]
Topological networks approach, e.g. over-connection analysis, hidden node analysis, rank aggregation, network analysis [M]	Find common regulators of different data types	3Omics (Kuo et al., 2013) SteinerNET (Tuncbag et al., 2012)	Piruzian et al. (2010) [H], Imielinski et al. (2012) [H]
Merging of datasets on individual levels (proteome, transcriptome), followed by correlation analysis [C]	Better correlation than in single data sets, also observed in subsets of merged data sets	-	Greenbaum et al. (2003) [Y]
Missing value estimation by non-linear optimization using relations between transcriptomic and proteomic data [M]	Predicting missing protein expression data	-	Torres-García et al. (2009) [B], Torres-García et al. (2011) [B]
Multiple regression analysis [M]	Predicting correlation of mRNA/proteins taking into consideration covariates such as multiple sequence features	- -	Nie et al. (2006b) [B], Nie et al. (2006a) [B]
Clustering approaches in proteome/transcriptome domain [T]	Cluster correlations	Matlab code available on request	Rogers et al. (2008) [H]
Dynamic modeling, e.g. boolean modeling, differential equations models, Bayesian networks [C]	Refining a model by combining heterogeneous data	-	Nariai et al. (2004) [Y], Werhli and Husmeier (2007) [Y], Zhang et al. (2007) [Y], Hamon et al. (2014) [H]
Generation of combined scores [T]	Identification of consistently changing proteins/genes over different data sets	-	Balbin et al. (2013) [H]
Multivariate regression [M]	Identification of information flow by modeling globally joint, locally joint and unique variation in data sets	-	Srivastava et al. (2013) [P]

Table 1.1. *Non-exhaustive overview of different data integration approaches and tools. Integration approaches are classified as C - concatenation-based approach, T - transformation-based approach, M - model-based approach. Species from which data was collected are annotated with B - bacteria, W - wheat, H - human, R - rat, Y - yeast, P - populus.*

1.4 Exploring molecular dynamics via time series data

Molecular regulatory mechanisms are due to their complexity not fully representable by a single measurement, even if their characterization is done in high-throughput. Signaling cascades, feedback mechanisms or pathway crosstalk are important examples that illustrate the necessity for time-resolved investigation. Therefore, time-series expression data is increasingly generated with the aim to monitor cyclic processes or the molecular reaction upon external perturbation (Bar-Joseph et al., 2012). From a systems biology perspective such data enables a deep characterization of the system dynamics with regard to the coordination of molecular processes, the relationship between individual molecules and the rate of changes observed. When data on coordinated processes is available, inference of causal regulatory links can be performed, leading to a better understanding of the finely orchestrated cellular reactions.

1.4.1 Time series data - monitoring temporal variation

Time series data of cyclic processes, e.g. the cell cycle, have demonstrated that a deeper understanding of molecular dynamics is not obtainable by just measuring individual cellular states or 'snapshots'. This is due to the fact that transcriptional and translational processes do not only increase the complexity of the molecules' information content (as shown in Figure 1.2), but are also coupled dynamically.

With our linear understanding of how time passes, a molecular interaction is always dependent on previous interactions of the molecule itself and other interaction partners. Thus, both spatial and dynamic dependency is narrowing down options of the molecular interplay at a certain point in time. Given this dependency, there are fixed sets of possible further interaction steps for each molecule throughout transcriptional and translational processes. With the increasing knowledge of biological interactions, many of them being available in biological databases (see Section 1.5 *Biological knowledge resources*), the question arises whether it is feasible to define this set of possible interactions at certain points in time in the future.

However, with our current understanding of molecular processes, time-resolved data enables us to follow individual signaling axes over time, granted that different data types on different molecular levels are available. Therefore, upon an external stimulation of a cell, we expect a cellular response that starts with a signaling cascade involving phosphorylation processes and ending with transcription factor relocalization into the cellular nucleus. This process triggers transcriptional changes that are often dependent on other molecular partners of transcriptional complexes. The generation of new RNA then results in protein translation, which itself can affect the signaling pathway characteristics via changed protein expression levels to enable a long-term cellular response.

Such a cascade, as a matter of course, depends on molecular synthesis and degradation rates, as well as post-transcriptional and post-translational modifications. Only recently, we

have gained more precise knowledge about such 'molecular timing' in mammalian cells, often through single-cell techniques.

Transcription rates have been measured in mammalian cells with different techniques. Yunger et al. (2010) could observe rates between 0.3 and 0.8 kb min^{-1} in vivo. Maiuri et al. (2011) reported transcription rates of 10 and 35 kb min^{-1} for nascent RNAs from an integrated human immunodeficiency virus type 1-derived vector. Others reported values of 3.8 kb min^{-1} (Singh and Padgett, 2009) and 3.1 kb min^{-1} (Wada et al., 2009) from a bulk analysis of the first transcriptional wave.

After external stimulation, transcriptional bursting has been observed in mammalian cells, which was followed by silent periods. These bursts have been characterized further by Bahar Halpern et al. (2015), who investigated nuclear retention of mRNA as a buffer that dampens the linked gene expression noise. But it is the combination of burst fractions, transcription rates and mRNA stability that leads to the final level of cellular mRNA and can affect noise and response time (Rabani et al., 2011; Schwanhäusser et al., 2011).

mRNA stability is also dependent on its decay mechanisms, which either constitute a quality control step and/or mechanistically change the abundance of functional proteins by changing mRNA half-life. This depends on gene transcription itself, pre-mRNA splicing, pre-mRNA 3'-end formation and other post-transcriptional modification as well as mRNA export from nucleus to cytoplasm (Schoenberg and Maquat, 2012). Schwanhäusser et al. (2011) reported median mRNA half-lives of $\sim$ 9 hrs in a global quantification of mammalian gene expression control.

How much the dynamic changes in RNA levels are influenced by RNA stability has been under debate: The 'constant degradation hypothesis' has been opposed by the 'varying degradation hypothesis'. The former assumes a constant degradation per gene over time, the latter implies strong effects by RNA degradation rate, either by individual changes or by a continuous shift over time (Rabani et al., 2011). By combining metabolic labeling of RNA with advanced RNA quantification assays and computational modeling, these authors were able to show that for most genes (94 %) dynamic changes in degradation rates have very little impact on expression changes during the first 6 hrs of a cellular response. For the rest of the genes, they rejected the constant degradation model, indicating that either there is no constant but temporally changing degradation or that there are other intervening post-transcriptional events.

Schwanhäusser et al. (2011) also measured protein half-lives and observed them to be in the order of $\sim$ 50 hrs. Yet, high variation between proteins was observed. Boisvert et al. (2012) determined the average turnover rate in HeLa cell proteins to be $\sim$ 20 hrs in a quantitative proteomics analysis of protein turnover. Kristensen et al. (2013) observed that protein expression during cellular differentiation is largely controlled by synthesis rate changes, whereas the relative degradation rate shows only minor changes in the majority of proteins. Unstructured lower abundance proteins were reported to show very fast regulation of a large part of the signal transduction network, which was in line with findings by Lundberg

et al. (2010) which showed the disparities in different cell types to be largely dependent on lower abundance proteins.

Targeted protein degradation is crucial for regulation of signaling pathways. Large scale protein experiments have shown protein degradation to vary between a range of minutes and tens of hours. However, most proteins show half-lives similar to cell doubling times. Recently, protein degradation in different subcellular compartments of a human cell line has been reported for $\sim$ 5000 proteins (Larance et al., 2013).

Due to the aforementioned interdependencies, mRNA stability modulation has been suggested as a therapeutic approach (Eberhardt et al., 2007). However, individual molecule dynamics are diverse and also dependent on the availability of e.g. enzymes or co-factors. This generates a high number of combinatorial effects when trying to resolve the molecular relationships based on time series data. Still, time-resolved data is the only means that enables us to follow molecular generation or degeneration and molecular stability. It can provide us with links between dynamic signaling and functional specificity and enable us to answer questions e.g. regarding environmental influences on signaling. Accordingly, when interlinking both gene and protein time-series expression data sets, a more detailed understanding of the molecular interplay can be generated. Detailed time-resolved integration is part of the scope of this thesis.

1.4.2 Modeling molecular dynamics in systems biology

The most widely used systems biology bottom-up approach for modeling molecular dynamics is clearly ordinary differential equation (ODE) modeling. This approach, however, comes with the challenge, that usually a lot of individual parameters need to be known prior to modeling. Some of these parameters might not even be measurable, causing the need for a parameter estimation process. Additional consideration needs to be given to the degree of modeling complexity required in order to address the biological questions at hand.

The focus of systems biology top-down approaches in identifying and understanding molecular regulation is rather on the inference of causal molecular interactions. These usually require experimental validation in a second step. The two approaches mainly applied for analysis of time-series data in this context are Granger causality and dynamic Bayesian modeling.

Granger causality is based on the idea that if signal A causes signal B, then the past values of signal A should provide information for prediction of B, in addition to past values of B itself (Granger, 1969). Though the basic concept only gives information about linear features of signals, there are extensions to nonlinear cases.

Dynamic Bayesian networks (DBNs) are a probabilistic representation of a probability space. Based on a stochastic process probability distributions of random variables can be modeled. This stochastic process is presumed to satisfy the first order Markov property, i.e. future states of the process (conditional on both past and present states) depend only upon the present state. They have been proposed initially by Dagum et al. (1992) to extend linear

state space models and their theory will be introduced in 1.4.3 *Dynamic Bayesian Network inference*.

In a review of Xuan et al. (2012) different inference approaches were compared for the 10-gene networks released in the DREAM4 challenge. For the smaller systems investigated DBNs were competitive with non-parametric approaches in respect to computational time and accuracy and outperformed Granger causality-based methods and simple ODE models. Consequently, this thesis focusses on the application of DBN inference to elucidate molecular mechanisms.

1.4.3 Dynamic Bayesian Network inference

A Bayesian Network is a graphical model for representing conditional independencies between a set of random variables. It consists of

1. a directed acyclic graph $\mathbf{G} = (X, D)$ with $X = (X_i)_{i\in\{1,...,n\}}$ denoting the set of nodes and D denoting the set of edges between the nodes in $\mathbf{G}$. The nodes represent a set of random variables.

2. a set of local probability distributions $(P(X_i|Pa(X_i)))_{i\in\{1,...,n\}}$, defining the probability distribution of each node conditional only on the value of its parent variables $(Pa(X_i))_{i\in\{1,...,n\}}$.

The graph represents the qualitative dependence relationships, the local probability distribution function (PDF) represents quantitative information about the strength of those dependencies.

Bayes' rule states that the posterior probability of x_1 given x_2 ($P(x_1|x_2)$) can be computed given the prior $P(x_1)$ and the likelihood $P(x_2|x_1)$:

$$P(x_1|x_2) = \frac{P(x_2|x_1)P(x_1)}{P(x_2)} \tag{1.1}$$

where $P(x_2) \neq 0$.

Therefore, Bayes' rule enables updating our belief about a hypothesis x_1 based on new evidence x_2: While we might have direct information about $P(x_2|x_1)$ and prior information about $P(x_1)$, direct information about $P(x_1|x_2)$ might be difficult to obtain directly. The denominator represents a normalization term, ensuring that the posterior probability over all possible values adds up to 1. Given knowledge about conditional relationships between the variables, we can thus learn probability distributions of all parts of the system if evidence about the existence of certain entities (such as x_2 in the example) can be assessed.
Alternatively, Bayesian networks can be described as the product of conditional probabilities:

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|Pa(x_i)), \tag{1.2}$$

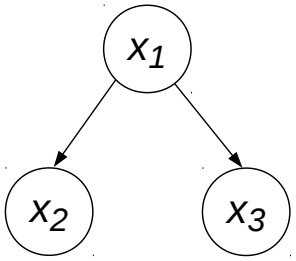


Figure 1.4. *Simple Bayesian Network. Given x_1, nodes x_2 and x_3 are conditionally independent.*

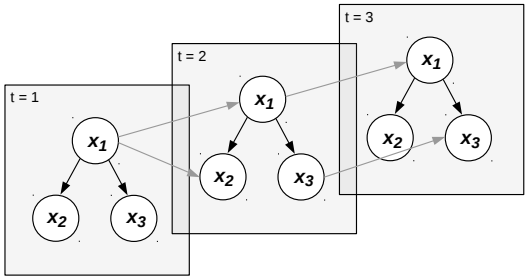


Figure 1.5. *Simple Dynamic Bayesian Network. Each time point is represented as a time slice. A node can only depend on a node in the previous time slice or on a parent node of the same time slice. Interslice edges are colored in grey, intraslice edges are depicted in black.*

with $Pa(x_i)$ being the parent node set of node x_i.

Given a directed acyclic graph, a Bayesian network (compare Figure 1.4) with respect to this graph is defined by initial specification of the conditional probability distributions of each node given its parents in this graph if the joint distribution satisfies Equation 1.2.

Bayesian networks can be used for three kinds of reasoning (Murphy and Mian, 1999):

- causal reasoning: from known causes to unknown effects,
- diagnostic reasoning: from known effects to unknown causes, or
- for any combination of these two,

depending on the degree of observability of the variables.

DBNs are an extension of Bayesian networks, which serve as models for systems which are dynamically evolving over time. They reflect a special case of singly connected Bayesian Networks, in which the connections are between discrete time 'slices' (Figure 1.5). The network's states fullfil the Marcovian condition in that any state of the network solely depends on its immediate precursor state.

As in a DBN not all states need to be observable, it can be described with a sequence of hidden-state variables $X = \{x_0, ..., x_{T-1}\}$ and a sequence of observed variables $Y = \{y_0, ..., y_{T-1}\}$ with T representing the time boundary:

$$P(X,Y) = \prod_{t=1}^{T-1} P(x_t|x_{t-1}) \prod_{t=0}^{T-1} P(y_t|x_t)P(x_0), \tag{1.3}$$

Hence, for a full specification of a DBN, we need definitions of

1. the state transition PDFs, giving the time dependencies between the states,
2. the observation PDFs, specifying dependencies of observation nodes from other nodes at the same time and
3. the initial state distribution $P(x_0)$.

This definition allows addressing the following issues:

- Inference: Estimation of unknown states on the basis of observed states and the initial probability distribution.
- Decoding: Identification of the most likely sequence of hidden variables given the observations.
- Learning: Estimation of DBN parameters that match the observed data to arrive at the best model for the system.
- Pruning: Removing nodes from the network structure which are of no relevance for inference.

In this thesis, network inference is performed with the state space model visualized in Figure 1.6, which constitutes a special case of a DBN.

Let t denote time points, let r denote replicates, let K be the dimension of hidden states, let P be the dimension of observations, let $x_{tr} = \{x_{tr1}, ..., x_{trK}\}$ denote the set of hidden states, let $y_{tr} = \{y_{tr1}, ..., y_{trP}\}$ denote the set of observed genes/proteins, let v be the gene/protein precisions vector and $A \in \mathbb{R}^{(K\times K)}$ be the state-to-state matrix, $B \in \mathbb{R}^{(K\times P)}$ be the observation-to-state matrix, $C \in \mathbb{R}^{(P\times K)}$ be the state-to-observation matrix, $D \in \mathbb{R}^{(P\times P)}$ be the observation-to-observation matrix. Further, let $w = (w_t)_{t\in\{0,...,T\}}$ and $z = (z_t)_{t\in\{0,...,T\}}$ denote collections of random variables with $w \sim MVN(0, I_{K\times K})$ and $z \sim MVN(0, diag(v)^{-1})$ where $I_{K\times K}$ denotes the K-dimensional identity matrix. Then the 'Empirical Bayes Dynamic Bayesian Network' as implemented in the R package *ebdbNet* of Rau et al. (2010) is defined by:

$$\begin{aligned} x_{tr} &= Ax_{t-1,r} + By_{t-1,r} + w_{tr} \\ y_{tr} &= Cx_{t,r} + Dy_{t-1,r} + z_{tr} \end{aligned} \tag{1.4}$$

This model was developed for inference of gene regulatory networks, but is employed in this thesis for the integrated data. Therefore, the observed states include not only gene expression

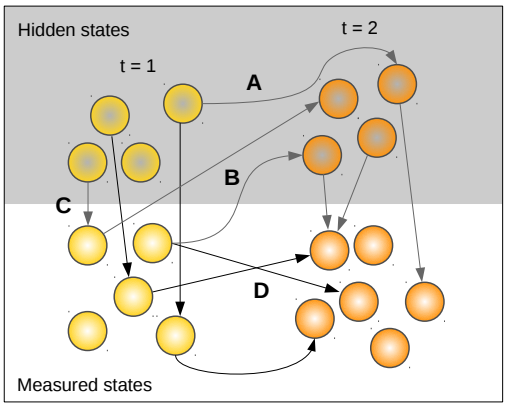


Figure 1.6. *ebdbNet state space model (modified from Rau et al. (2010)). Shown are two consecutive time points, t = 1 and t = 2, colored in yellow and orange. Hidden (non-observed) states are depicted on grey, measured (observed) states on white background. State matrix names, corresponding to A, B, C and D in Equations 1.4, are the state-to-state matrix, the observation-to-state matrix, the state-to-observation-matrix and the observation-to-observation matrix. Note, that the states can correspond to gene expression or phosphoprotein abundance levels.*

data, but as well phosphoprotein expression data. In addition, regulatory links between the different molecule types are enabled. In this work, the observation-to-observation matrix D, which provides the structure of the inferred network, is of ultimate interest. The hidden states dimension K is determined via the block-Hankel matrix of autocovariances of the observations. The latter is defined by the time lag between measurements and the estimated maximum relevant biological time lag between regulators and regulated molecules. Estimation of the hidden states dimension K is then performed by singular value decomposition of the block-Hankel matrix: The optimal value for K is found when a further singular value does not considerably increase the amount of explained variation anymore. A corresponding threshold value was determined by simulations (Rau et al., 2010). Based on the state matrices and the precisions vector v a Kalman filter and smoother is used to estimate the hidden states, given their dimension K.

Let $\mathbf{a}_{(j)}$, $\mathbf{b}_{(j)}$, $\mathbf{c}_{(j)}$ and $\mathbf{d}_{(j)}$ denote vectors of the jth rows of the matrices A, B, C and D, with $\alpha = \{\alpha_1, ..., \alpha_K\}$, $\beta = \{\beta_1, ..., \beta_P\}$, $\gamma = \{\gamma_1, ..., \beta_K\}$ and $\delta = \{\delta_1, ..., \delta_P\}$ building the set of hyperparameters $\psi = \{\alpha, \beta, \gamma, \delta\}$, v_i being the ith component of the precision vector v, $j \in \{1, ..., K\}$, $i \in \{1, ..., P\}$. Then the a priori precisions of the parameter set are described by the set of hyperparameters ψ and the set of parameters $\theta = \{A, B, C, D, v\}$ (Rau et al., 2010):

$$\begin{aligned} \mathbf{a}_{(j)}|\alpha &\sim N(\mathbf{0}, diag(\alpha)^{-1}) \\ \mathbf{b}_{(j)}|\beta &\sim N(\mathbf{0}, diag(\beta)^{-1}) \\ \mathbf{c}_{(i)}|\gamma, v_i &\sim N(\mathbf{0}, v_i^{-1} diag(\gamma)^{-1}) \\ \mathbf{d}_{(i)}|\delta, v_i &\sim N(\mathbf{0}, v_i^{-1} diag(\delta)^{-1}) \end{aligned} \tag{1.5}$$

The hyperparameters' point estimate is identified with an expectation maximization like algorithm, conditioned on the current estimates $\hat{\mathbf{x}}$ of the hidden states. Thus, the posterior means of A, B, C and D can be calculated. The final network is defined when global convergence of the parameters is reached. In this thesis, the convergence criteria tested in extensive simulation runs by Rau et al. (2010) have been used.

1.5 Biological knowledge resources

Numerous biological databases are available, many of them being commercial databases. However, also the number of open-access databases is large and constantly growing. Both enable a comparison of newly generated data with already known biological interactions or associations, which were gathered mostly in single experiments or with high-throughput methods over the last decades. With the growing use of high-throughput techniques, this comparison can be a valuable supplement to compare new results with database content and to check for contradictory findings (Glaab, 2015).

In this thesis, public biological knowledge from databases is employed to identify signaling axes ranging over different molecular levels. In this way, their potential as a means for cross-platform data integration approaches is exploited. Contents of the databases used include pathway models (from KEGG, Reactome, NCI and Biocarta databases), transcription factor target interactions (from Chea, Pazar and TRANSFAC databases/collections), protein-protein interactions (STRING database) and phosphorylation processes (PhosphoSitePlus database). Table 1.2 gives an overview on the databases used in this thesis, their sizes and versions, content, curation and references.

One of the drawbacks of exploiting database knowledge is the comparison with known biological interactions, therefore, no 'new knowledge' is generated. Another issue is the fact that knowledge stored in most databases is compiled over different experiments, often originating from different species, different cells and different experiments. Hence, interpretation of public knowledge based analyses needs to be performed with caution, yet it can also provide considerable insight into signaling links and relations that might not be clear and evident only based on the data.

Database	Content	Curation	Version/Size	References
Pathway model databases				
KEGG	molecular interaction/reaction network diagram	manually curated	KEGG public 2011 version	Kanehisa et al. (2016) Kanehisa and Goto (2000)
Reactome	database of pathways and reactions in human biology	free, open-source, curated and peer reviewed	Reactome 2011 version	Croft et al. (2014) Milacic et al. (2012)
PID	human molecular signaling and regulatory events and key cellular processes, now available via NDEx database, Pratt et al. (2015)	freely available collection of curated and peer-reviewed pathways	PID 2014 version	Schaefer et al. (2009)
Biocarta	diagrams of biological pathways and protein complexes	literature curation	Biocarta 2011 version	Nishimura (2001)
Transcription factor target relation databases/ collections				
ChEA	mammalian ChIP-X database	manually curated	no longer supported, ChEA 2015 version 92 transcription factors, 31,932 target genes; 189,933 interactions from 87 publications	Lachmann et al. (2010)
Pazar	software framework for the construction and maintenance of regulatory sequence data annotations	manually curated	Pazar 2015 version	Portales-Casamar et al. (2009)
TRANSFAC	data on eukaryotic transcription factors, their experimentally-proven binding sites, consensus binding sequences (positional weight matrices) and regulated genes	literature curation	TRANSFAC Professional version 2014.4/ 2016.1 48,000+ transcription factor binding site reports	Matys et al. (2006) Wingender (2008)
Protein-protein interaction database				
STRING	database of known and predicted protein interactions; direct (physical) and indirect (functional) associations derived from four sources : genomic context, high-throughput experiments, (conserved) coexpression, previous knowledge	literature curation, (i) text mining of scientific texts, (ii) interactions computed from genomic features, (iii) interactions transferred from model organisms based on orthology	STRING version 10 9,643,763 proteins from 2,031 organisms	Szklarczyk et al. (2015) Jensen et al. (2009)
Protein modification database				
PhosphoSitePlus	protein modification resource	gathered from published literature and other sources, curated and reviewed	PhosphoSitePlus 2015/2016 version 53,235 proteins with 485,815 post-translational modifications	Hornbeck et al. (2015)

Table 1.2. *Overview on biological databases used.*

1.6 Investigated signaling pathways

Knowledge about well characterized signaling pathways, such as the epidermal growth factor receptor (EGFR) signaling pathway or the BCR signaling pathway can be used as benchmark knowledge when new methodological integration methods are developed. In this thesis data sets covering these two pathways were investigated in detail. The data sets exhibit the considerable advantage that they cover most popular high-throughput methods both on (phospho-)proteomic and transcriptomic level. This includes antibody-based methods and MS for (phospho-)protein expression data and microarrays and RNA-Seq for gene expression data. Thus, assessment as well as comparison of the integration on different input data sets is facilitated.

1.6.1 Epidermal growth factor signaling

The EGFR signaling pathway, depicted in Figure 1.7, is crucial in the cellular response to growth factors. Upon ligand binding, EGFR undergoes a transition process by forming active homodimers. Additionally, pairing with other members of the ErbB family which resulted in active heterodimers, has been observed (Ward and Leahy, 2015). Ligand binding stimulates the intracellular tyrosine kinase activity of the receptor, resulting in autophosphorylation of tyrosine residues in the C-terminal domain of EGFR. This phosphorylation triggers downstream activation and initiates several signal transduction cascades, such as the MAPK, AKT, and JNK pathways. Their impacts are DNA synthesis and cellular proliferation, which explains why aberrant EGFR signaling is highly important in various diseases associated with cellular proliferation, such as cancer (Citri and Yarden, 2006).

In this thesis, the focus for EGFR signaling analysis was set on pathway crosstalk and feedback signaling. As time-course data was available on four time-points measured in parallel both on phosphoprotein and transcriptome level, inference of causal links between the different data sets was considered to be of biological interest. As pathway coverage of these phosphoproteins was low, the focus was not on a systematic characterization of signaling axes.

1.6.2 B cell receptor signaling

The activation of the BCR signaling pathway, depicted in Figure 1.8, is highly important in the adaptive immune response. The receptor is responsible for recognizing B cell encounters with antigens. In this case B cells are needed to proliferate and differentiate in order to generate high-affinity antibody secreting plasma B cells and long-lived memory B cells. Therefore, the BCR has a twofold function, i) receptor oligomerization for further signal transduction processes, ii) mediating internalization and further processing of the antigen as well as presentation of its peptides to helper T cells (Yuseff et al., 2013).
The receptor itself is comprised of membrane immunoglobulin (Ig) heavy and light chains, associated with an intracellular Ig$\alpha\beta$ heterodimer. Phosphorylation of its immunoreceptor

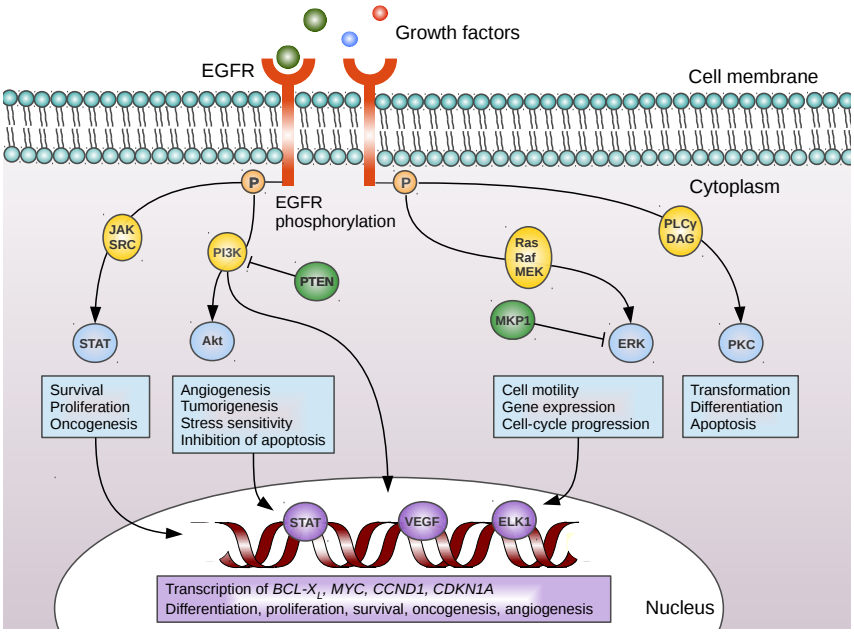


Figure 1.7. *EGFR signaling and downstream signaling effects (modified from Nyati et al. (2006)).*

tyrosine activation motif by Src-family kinases leads to activation of SYK. Downstream, various intracellular signaling molecules are assembled, leading to different cellular processes, such as gene expression, reorganization of the cytoskeleton, and BCR-mediated internalization of antigen complexes. These are processed in endosomal compartments and presented on the extracellular surface bound to the major histocompatibility complex II in order to recruit T helper cells (Harwood and Batista, 2008). Important signaling axes include MAPK signaling, NFAT signaling, AKT/mTOR signaling and NF-κB signaling.

In this work, a systematic characterization of BCR signaling in a Burkitt's lymphoma cell line model was performed throughout the different molecular levels of phosphoproteome and transcriptome data over time. Integration of these two levels was used to identify cross-platform derived consensus molecule sets regulated at certain time points after receptor stimulation. In addition, tracking of individual signaling axes based on the two data sets was established.

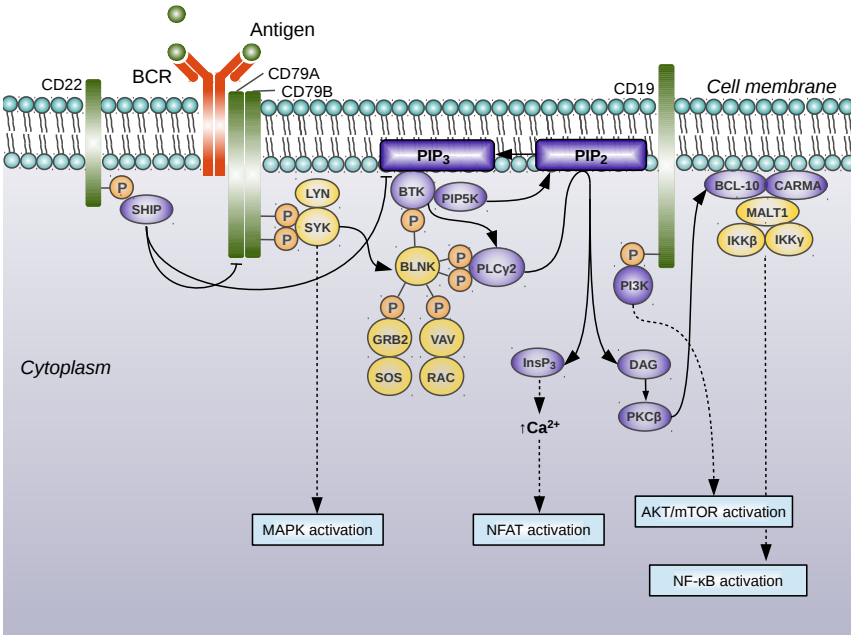


Figure 1.8. *BCR signaling and downstream signaling effects (modified from Young and Staudt (2013); Scharenberg et al. (2007); Monroe (2006)).*

1.7 Objectives and overview

With increasing numbers of different high-throughput data types generated based on an individual experiment, the challenge of i) integrating the different data types methodologically and ii) providing software making these methods available to the public is of particular importance.

The objectives of this thesis were to

- Develop a methodology to integrate time-series phosphoproteomic, transcriptomic and proteomic high-throughput data by taking into consideration the different molecular levels of measurement
- Implement an open source software package offering the developed methodology to other researchers
- Demonstrate the method's strengths on structurally different data sets
  - EGFR signaling data set: phosphoproteome and gene expression data sets are limited due to technical reasons,
  - BCR signaling data set: phosphoproteome and gene expression data sets were generated with up-to-date sensitivity methods.

In this thesis, the first publication Chapter 2 *pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge* introduces a new

approach for the integration of phosphoproteomic, transcriptomic and proteomic data in form of the R software package 'pwOmics' developed with the above-mentioned objective. It can not only be used with single parallel measurements of the different data types, but combines the integration with a focus on time-series data. This focus enables a more detailed characterization of possible causal links in mechanistic regulation processes. As open-source package it is available to a wide range of researchers working on data integration.
The second publication Chapter 3 *Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data* applies the new integration approach on a data set that measured epidermal growth factor signaling in human mammary epithelial cells. Assessment of the different molecular layers of measurement enables a detailed decoding of signaling processes over time. Data analysis with this new integration approach could identify regulatory patterns already known in EGF signaling, but also hint to other mechanisms not described yet in literature, thus allowing for hypothesis generation of biological processes that can be experimentally cross-checked. Such hypotheses enable the focused and pre-informed selection of experiments for an identification of e.g. signaling pathway crosstalks or feedback loops and can reduce resources and time during the characterization of a cellular response.
The third publication Chapter 4 *Integration of phosphoproteome and transcriptome data to link B cell receptor activation with gene expression dynamics* uses the presented integration approach to systematically identify individual signaling axes in B cell receptor signaling in human DG75 cells. While in this lymphoma cell line pathophysiological signaling is mainly attributed to tonic signaling (Corso et al., 2016), the focus in this work is on activated signaling. A dissection of the cellular response in regard to activated signaling pathways and affected transcription is possible when integrating the data sets with biological database knowledge.
Hence, the listed publications demonstrate applicability of the presented data integration approach on a technically limited data set as well as on a broader data set not biased by selection of measured phosphoproteins/-sites. In Chapter 5 *Discussion*, the overall results of the publications included in this thesis are discussed.

2 pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge

Reference
pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge Astrid Wachter; Tim Beissbarth Bioinformatics 2015; doi: 10.1093/bioinformatics/btv323.

Original Contribution
AW and TB conceived the data integration strategy. Development and implementation of the software package 'pwOmics' was performed by AW. This included the selection of the public databases used in the integration process, writing of the software documentation, as well as publication of the software package on the R platform 'Bioconductor'. AW generated the figures, prepared example results and wrote the manuscript.

BIOINFORMATICS

Pages 1–3

pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge

Astrid Wachter [1,*] and Tim Beissbarth [1]

[1]Department of Medical Statistics, Georg-August-University Göttingen, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Characterization of biological processes is progressively enabled with the increased generation of omics data on different signaling levels. Here we present a straightforward approach for the integrative analysis of data from different high-throughput technologies based on pathway and interaction models from public databases. *pwOmics* performs pathway-based level-specific data comparison of coupled human proteomic and genomic/transcriptomic data sets based on their log fold changes. Separate downstream and upstream analyses results on the functional levels of pathways, transcription factors and genes/transcripts are performed in the cross-platform consensus analysis. These provide a basis for the combined interpretation of regulatory effects over time. Via network reconstruction and inference methods (steiner tree, dynamic bayesian network inference) consensus graphical networks can be generated for further analyses and visualization.

Availability: The R package *pwOmics* is freely available on Bioconductor (http://www.bioconductor.org/).

Contact: astrid.wachter@med.uni-goettingen.de

1 INTRODUCTION

High-throughput technologies applied in systems biology research generate large amounts of molecular information nowadays. Interpretation of genome- and proteome-wide data is dependent on current analysis tools. As each technique shows a certain bias and has natural limitations in identifying full signaling responses (Yeger-Lotem *et al*., 2009), cross-platform analysis is an up-to-date approach in order to connect biological implications on different signaling levels. Usage of diverse data types provides a deeper understanding of global biological functions and the underlying processes (Kholodenko *et al.*, 2012). Thus, development of integrative software solutions for data from different high-throughput techniques is a current major challenge for bioinformatic analysis. Existing widely-used commercial software solutions such as QIAGEN's Ingenuity®Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) or MetaCore™ (GeneGo, Inc., St. Joseph, MI) and also open-source software, such as Cytoscape (Shannon *et al.*, 2003), often handle proteomic and genomic/transcriptomic data as if coming from the same functional level. More specific integration tools which are considering these levels include e.g. the web tool IMPaLA (Kamburov *et al.*, 2011), which provides knowledge based data integration on transcriptomics or proteomics data combined with metabolomics data, and the webserver SteinerNet (Tuncbag *et al.*, 2012), which enables integration of transcriptional, proteomic and interactome data utilizing Steiner trees. However, *pwOmics* combines these distinct omics levels of evidence in order to refine the understanding of molecular mechanisms including the biologically important time effect. Thereby, it joins tools used for network analysis (Kristensen *et al.*, 2014), but adds a level of complexity by attributing weight to the different functional levels of measurement in the first place and the dimension of time in the second place. We implemented *pwOmics* as open-source package for R, a free software environment for statistical computing commonly used for bioinformatic analyses.

*to whom correspondence should be addressed

2 APPROACH

pwOmics provides analyses functionalities and comparative integration features for coupled human proteome and genome/transcriptome data sets. The analysis workflow is adapted to account for the biological control mechanisms occurring on the different regulation levels such as transcriptional control on gene level, mRNA processing on transcript level and post-translational modifications on protein level, as illustrated in Figure 1. The two data sets are initially analyzed separately enabling a level-specific interpretation of up- and downstream changes of regulatory molecules. The protein based downstream analysis comprises the pathway-based identification of transcription factors (TF) of differentially abundant proteins and their target genes. The gene/transcription based upstream analysis identifies TFs and proteomic regulators based on differentially expressed transcripts or genes. As high-throughput data are increasingly used to follow time-dependent biological regulation after pertubation, the main benefit of *pwOmics* is the cross-platform time series analysis functionality, but consensus analysis can be performed also on single time point measurements.

3 PACKAGE FEATURES

3.1 Databases

Existing knowledge stored in public databases is a key element for data integration in the approach outlined above (Kramer *et al.*, 2014). Databases used here are pathway databases, TF-target databases and a protein-protein interaction database. Pathway databases can be selected individually or as combination of KEGG (Kanehisa *et al.*, 2014), Reactome (Croft *et al.*, 2014), Pathway Interaction Database (Schaefer *et al.*, 2009) and Biocarta (Nishimura, 2001). The information is used as gene sets in the downstream analysis and combined with topological information in

upstream analysis. Prior knowledge for network reconstruction is based on the connected graph from protein-protein-interaction (PPI) database STRING (Franceschini *et al.*, 2013). For TF-target gene identification processes the user can choose from databases ChEA (Lachmann *et al.*, 2010) and/or Pazar (Portales-Casamar *et al.*, 2009) or specify an own file e.g. containing commercial database information.

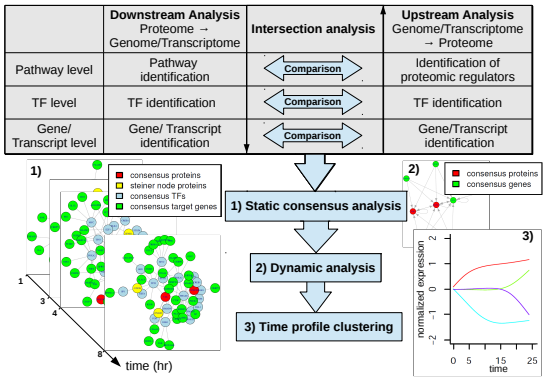


Fig. 1. *pwOmics* downstream and upstream analysis. Exemplarily shown are results of a static consensus analysis, a dynamic analysis and the time profile clustering.

3.2 Individual comparative analysis

In the individual analysis database information is used to identify signaling molecules of the different functional levels for a level-specific comparison. Identification of pathways containing differentially abundant proteins is performed via a Biopax model generated by the R package *rBiopaxParser* (Kramer *et al.*, 2013) on basis of the selected pathway databases. Enrichment of pathways in downstream analysis and TFs in upstream analysis is optional. Upstream regulators of TFs are identified via their pathways, but only those pathways are considered further which contain a user-specified number of TFs. Overlapping proteins found as neighbors of a certain order of those TFs are assumed to be proteomic regulators. Easy access to the individual level results is provided.

3.3 Consensus analysis

In the consensus analysis the intersection of signaling molecules on each functional level is identified and used for building consensus nets. For each matching time point a Steiner tree (Sadeghi and Fröhlich, 2013) is generated (implemented via the shortest paths based approximation algorithm) on the basis of intersecting proteins and TFs from up- and downstream analysis and the connected PPI STRING network. For this network reconstruction method intersecting molecules regarded as 'terminal nodes' are mapped to the PPI-network and those pathway components on shortest interconnecting paths are included which provide the shortest length of the overall network. Subsequently intersecting TF-target relations are included to contribute to the static consensus graphs for each matching time point. The dynamic consensus analysis additionally considers signaling changes over time by applying dynamic bayesian network inference via the R package *ebdbNet* (Rau *et al.*, 2010). Nodes considered in this step are those identified in all static consensus graphs. With smoothing splines an appropriate number of time points are generated under the simplifying assumption of a gradual change of signaling over time. This longitudinal data set is then used for the inference step. The result allows a significance level-based visualization of the dynamic bayesian network.

3.4 Time profile clustering

To identify similar co-regulation patterns over time *pwOmics* provides an integrated time profile clustering, based on the soft clustering fuzzy c-means algorithm implemented in the R package *Mfuzz* (Kumar *et al.*, 2007). The soft-clustering approach has the advantage of assigning several clusters to one signaling molecule based on similarity of log-fold change dynamics to several clusters. Thus it enables an adequate clustering of complex expression time profiles, which are characterized by fine-tuned transcriptional mechanisms.

3.5 Data visualization

For easier biological interpretation users can visualize following results: 1) Static consensus nets - based on matching time point comparisons of the two datasets. 2) Dynamic consensus net - based on dynamic bayesian network inference. 3) Time profile clustering - based on softly clustered log-fold changes with a combined visualization of proteins and genes/transcripts.

4 SUMMARY

We developed an R package as integrative pathway-based level-specific tool for the analysis and interpretation of signaling measured in parallel on different platforms. The presented approach enables the reduction of results to a very reliable set of regulatory signaling components, time profile clustering and the interpretation of static and dynamic consensus results. Further details and examples are provided in the package documentation.

Funding: This work was supported by the German Federal Ministry of Education and Research via the projects MetastaSys [0316173A] and MMML-Demonstrators [031A428B].

REFERENCES

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472-D477.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808-D815.

Kamburov, A. , Cavill, R. , Ebbels, T.M.D., Herwig, R., Keun, H.C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* **27 (20)**, 2917-2918.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199-D205.

Kholodenko, B., Yaffe, M.B., and Kolch, W. (2012). Computational Approaches for Analyzing Information Flow in Biological Networks. *Sci. Signal* **5**, re1-re1.

Kramer, F., Bayerlová, M., Klemm, F., Bleckmann, A., and Beissbarth, T. (2013). rBiopaxParser–an R package to parse, modify and visualize BioPAX data. *Bioinformatics* **29**, 520-522.

Kramer, F., Bayerlová, M., and Beissbarth, T. (2014). R-Based Software for the Integration of Pathway Data into Bioinformatic Algorithms. *Biology* **3**, 85-100.

Kristensen, V.B., Lingjrde, O.C., Russnes, H.G., Vollan, H.K.M, Frigessi, A., Brresen-Dale, A. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* **14 (5)**, 299-313.

Kumar, L., and E. Futschik, M. (2007). Mfuzz: A software package for soft clustering of microarray data. *Bioinformation* **2**, 5-7.

Lachmann, A., Xu, H., Krishnan, J., Berger, S.I., Mazloom, A.R., and Maayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438-2444.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report* **2**, 117-120.

Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M.I., Jiang, S., McCallum, A., Kirov, S., and Wasserman, W.W. (2009). The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucl. Acids Res.* **37**, D54-D60.

Rau, A., Jaffrzic, F., Foulley, J.-L., and Doerge, R.W. (2010). An Empirical Bayesian Method for Estimating Biological Networks from Temporal Microarray Data. *Statistical Applications in Genetics and Molecular Biology* **9**.

Sadeghi, A., and Fröhlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics* **14**, 144.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674-D679.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, T.W., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504.

Tuncbag, N., McCallum, S., Huang, S.C., Fraenkel, E. (2012), SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. *Nucl. Acids Res* **40** W505-W509.

Wang, X. and Zhang, B. (2013). customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29 (24)**, 3235-3237.

Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* **41**, 316-323.

Yosef, N., and Regev, A. (2011). Impulse control: Temporal dynamics in gene transcription. *Cell* **144**, 886-896.

3 Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data

Reference
Astrid Wachter, Tim Beissbarth: Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data. Front. Genet 2016, http://dx.doi.org/10.3389/fgene.2015.00351.

Original Contribution
AW developed the method, performed data analysis and wrote the manuscript. TB conceived the design, envisioned the project and revised the manuscript.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 January 2016
doi: 10.3389/fgene.2015.00351

Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data

Astrid Wachter and Tim Beißbarth*

Department of Medical Statistics, University Medical Center, Göttingen, Germany

OPEN ACCESS

Edited by:
Ekaterina Shelest, Hans-Knoell Institute, Germany

Reviewed by:
Frank Emmert-Streib, Tampere University of Technology, Finland
Lorenz Adlung, German Cancer Research Center (DKFZ), Germany

****Correspondence:***
Astrid Wachter
astrid.wachter@med.uni-goettingen.de

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *07 October 2015*
Accepted: *03 December 2015*
Published: *07 January 2016*

Citation:
Wachter A and Beißbarth T (2016) Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data. Front. Genet. 6:351. doi: 10.3389/fgene.2015.00351

Identification of dynamic signaling mechanisms on different cellular layers is now facilitated as the increased usage of various high-throughput techniques goes along with decreasing costs for individual experiments. A lot of these signaling mechanisms are known to be coordinated by their dynamics, turning time-course data sets into valuable information sources for inference of regulatory mechanisms. However, the combined analysis of parallel time-course measurements from different high-throughput platforms still constitutes a major challenge requiring sophisticated bioinformatic tools in order to ease biological interpretation. We developed a new pathway-based integration approach for the analysis of coupled omics time-series data, which we implemented in the R package *pwOmics*. Unlike many other approaches, our approach acknowledges the role of the different cellular layers of measurement and infers consensus profiles and time profile clusters for further biological interpretation. We investigated a time-course data set on epidermal growth factor stimulation of human mammary epithelial cells generated on the two layers of RNA and proteins. The data was analyzed using our new approach with a focus on feedback signaling and pathway crosstalk. We could confirm known regulatory patterns relevant in the physiological cellular response to epidermal growth factor stimulation as well as identify interesting new interactions in this signaling context, such as the regulatory influence of the connective tissue growth factor on transferrin receptor or the influence of growth arrest and DNA-damage-inducible alpha on the connective tissue growth factor. Thus, we show that integrated cross-platform analysis provides a deeper understanding of regulatory signaling mechanisms. Combined with time-course information it enables the characterization of dynamic signaling processes and leads to the identification of important regulatory interactions which might be dysregulated in disease with adverse effects.

Keywords: omics, data integration, high-throughput, time-series, EGF signaling

INTRODUCTION

Omics data integration is a conclusive concept for a systemic understanding of biological signaling mechanisms, both in healthy conditions and disease (Kristensen et al., 2014; Ritchie et al., 2015). The combination of different types of omics data can provide a more comprehensive and complete picture of individual cellular mechanisms. Furthermore, a cross-platform analysis represents a measure to overcome individual platform biases and technical limitations (Yeger-Lotem et al., 2009).

An even more informative approach is to analyze time-course data sets from different omics levels, as a lot of cellular signaling information is encoded in signaling dynamics (Purvis and Lahav, 2013). This type of data provides more than only a single "snapshot" of the underlying biological processes, thus it can augment the knowledge we have about cellular signaling events considerably. With these data feedback signaling loops, molecular interactions and pathway crosstalk can be tracked over time. Thus, combining different types of omics data with time course information enables a comprehensive characterization of cellular responses upon stimulation and also a detection of regulatory mechanisms initiated by specific perturbations. In **Figure 1** a selection of dynamic regulatory signaling mechanisms on protein and gene layer is depicted. These effects become directly apparent in such omics data sets, so the "dynamic knowledge" we can collect may also provide us with an idea of modifications responsible for pathologic signaling and signaling dynamics, thus forming a basis for an improvement of treatment strategies.

Of course, such parallel time-course data sets are even more challenging to analyze and interpret as they include an additional dimension and require a meaningful cross-platform integration method. Hence, there is a demand for bioinformatic tools that can deal with the diverse data types and combine them in such a way that their output enables a straightforward biological interpretation of the data. Although a lot of individual data integration methods have been developed so far, they mostly address very specific integration questions (Balbin et al., 2013; Hamon et al., 2014), are not implemented as tools which can be freely used by other biologists and bioinformaticians [e.g., QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City[1])] or do not acknowledge the different nature of different omics data types (Ding et al., 2012; Sun et al., 2014). Very few tools also include the biologically very interesting aspect of time-course data analysis (Rogers et al., 2008), although these types of data sets are expected to be generated more often in the near future (Bar-Joseph et al., 2012) in order to address systems biology questions.

We developed a pathway-based data integration approach for the analysis of coupled high-throughput time-course measurements on the cellular layers of proteins, transcripts and genes. We implemented this approach as R package *pwOmics,* that we presented earlier (Wachter and Beissbarth, 2015). In brief, *pwOmics* joins the tools of network analysis: It uses public signaling pathway knowledge to map molecular network interactions, thereby identifying activated and inactivated genes and proteins in cellular signaling upon perturbation. Thus, the cellular layers on which the data is collected are acknowledged during data analysis while simultaneously considering the

[1]www.quiagen.com/ingenuity.

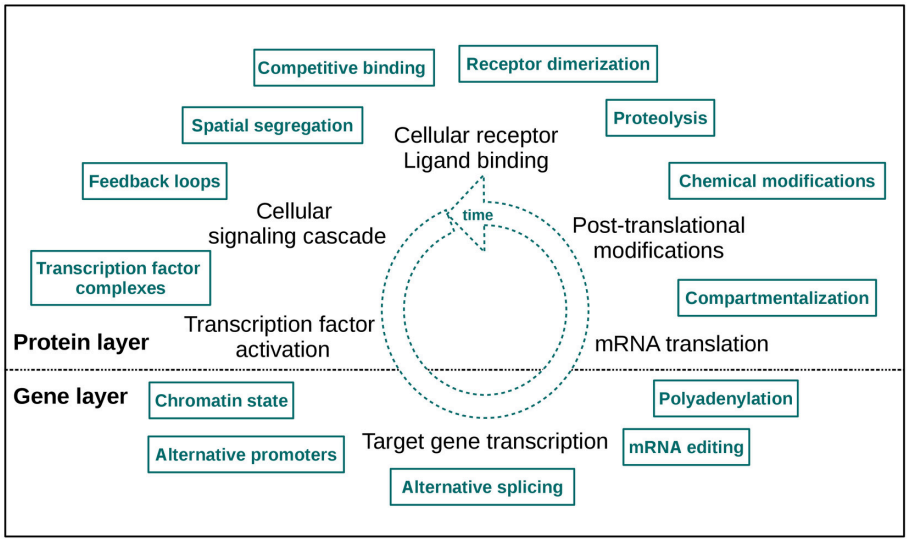


FIGURE 1 | A selection of cellular layer specific regulatory signaling mechanisms. The two layers of measurement are indicated as "protein" and "gene layer." The high number of effectors illustrates the mechanistic fine-tuning of signaling. Note that this fine-tuning also takes place in the dimension of time.

dynamics. Here we describe and test the utility of our method in more detail.

Epidermal growth factor (EGF) signaling has already been studied comprehensively in comparison to other signaling pathways as dysregulation is associated with poor prognosis in many human malignancies (Lurje and Lenz, 2009). As various high-throughput and low-throughput omics data sets are available and a lot of knowledge is already acquired on the basis of which methodical evaluation can be performed, it constitutes an adequate example for investigation of new approaches. The data set analyzed here measures the mitogenic response of human mammary epithelial cells (HMEC) to EGF on the proteomic and the transcriptomic layer over time (Waters et al., 2012), thereby representing physiological signaling conditions. **Figure 2** depicts the experimental design used in the study. EGF stimulation is associated with cellular proliferation, differentiation and survival (Herbst, 2004) and directly affects signaling pathways such as the MAPK signaling pathway, the ERBB signaling pathway and the RAS signaling pathway.

We chose the comparably well characterized example of EGF signaling in order to map the results of our new pathway-based integration approach to known experimental results for methodical evaluation and to reveal new dynamically relevant mechanisms in EGF signaling on the different functional layers. We focus on feedback signaling and pathway crosstalk, both complex regulatory mechanisms that have been under intensive biological investigation in individual experiments in physiological and pathological conditions (Avraham and Yarden, 2011; Wang et al., 2011).

METHODS

Data Set

The data set investigated with the new pathway-based integration approach was generated in a study on network analysis of EGF signaling. The experimental design used is illustrated in **Figure 2**, the measurements included transcriptomic, proteomic and phosphoproteomic data generation. Further details as well as the preprocessing steps performed on both microarray raw data and proteomic raw data are described in Waters et al. (2012). The raw microarray data files are available via the Gene Expression Omnibus database, GSE15668 (Waters et al., 2012). The corresponding proteomic data is also publicly available[2].

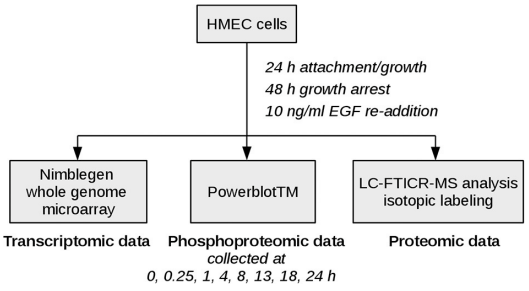


FIGURE 2 | Experimental design. HMEC cells were seeded and allowed to attach and grow for 24 h. After 48 h of growth arrest with medium lacking serum, EGF and other growth factors, EGF was added again to monitor the mitogenic response of the cells. Samples for high-throughput genomic and proteomic measurements were taken at time points 0, 0.25, 1, 4, 8, 13, 18, 24 h after EGF stimulation. The 0.25 h time point was excluded from the microarray data set due to quality issues, therefore the coupled data set on which our analysis is based includes time points 0, 1, 4, 8, 13, 18, and 24 h after EGF stimulation.

Shortly, biological samples were hybridized against NimbleGen microarrays. A quality check revealed that time point 0.25 h failed to hybridize, therefore the coupled data set analyzed here includes only time points 0, 1, 4, 8, 13, 18, and 24 h after EGF stimulation. Proteome analysis was performed MS-based, while phosphoproteome data were collected as part of a parallel western blot analysis. For each time point differentially expressed transcripts or differentially abundant phosphoproteins/proteins compared to time point 0 h were determined. Raw microarray data was quantile normalized before performing a pairwise analysis of variance with a 5% false discovery rate to determine differentially expressed transcripts. Proteome and phosphoproteome levels were considered significant when passing specific quality checks and showing a fold change $\geq$1.5.

Databases

Pathway information used for the pathway-based integration approach were taken from KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2014), Reactome (Croft et al., 2014), Pathway Interaction Database (Schaefer et al., 2009), and Biocarta (Nishimura, 2001). This information was used as gene sets in the analysis of the phosphoproteome data and combined with its topological information in the transcriptome data analysis. It was downloaded via the AnnotationHub R package[3] from Bioconductor (Huber et al., 2015) as BioPAX level 2 files and then processed further with the rBiopaxParser R package (Kramer et al., 2013). The transcription factor (TF)—target gene interaction information from the TRANSFAC® database (Biobase version 2014.4; Matys et al., 2006) was used. Network reconstruction was based on the connected protein-protein interaction (PPI) network of the STRING database (Franceschini et al., 2013).

Analyses

All analysis steps described here are based on pre-processed transcriptome, proteome and phosphoproteome data, as described in Waters et al. (2012). Main analyses steps were performed with the R package *pwOmics* (Wachter and Beissbarth, 2015). Our methodical framework is depicted in **Figures 3**, **4**.

Data Processing

First, individual analyses of the omics data sets were performed during phosphoprotein data based downstream and transcript based upstream analysis (**Figure 3**). For the downstream analysis an identification of the pathways, which include differentially

[2]http://omics.pnl.gov.

[3]Morgan, M., Carlson, M., Tenenbaum, D., and Arora, S. *AnnotationHub: Client to Access AnnotationHub Resources.* R package version 2.0.0.

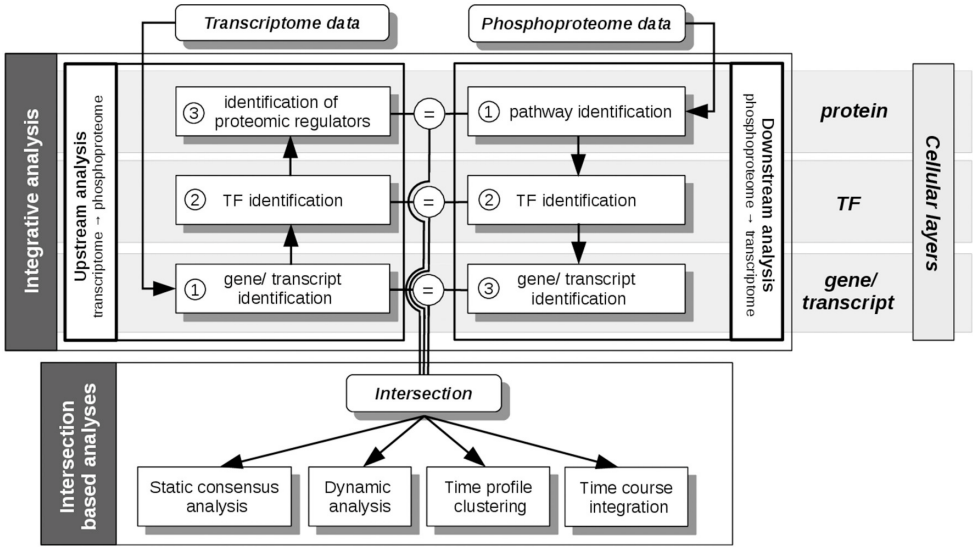


FIGURE 3 | pwOmics analyses steps. In the initial integrative analysis a downstream analysis of the phosphoproteome data and an upstream analysis of the transcriptome data is performed. The former includes the identification of pathways that include differentially abundant phosphoproteins, the identification of the TFs in these pathways and the determination of downstream target genes. In the upstream analysis the differentially expressed transcripts are identified, as well as their upstream TFs. By determining the pathways of these TFs also potential proteomic regulators can be identified. The intersection of the molecules on each cellular layer (protein, TF and gene/transcript) is determined before the intersection based analyses are performed. These include a static consensus analysis that can be performed for each measured time point, the consensus-based dynamic analysis that enables the generation of a probabilistic network exploiting the time-course information of those molecules that are part of the consensus analysis result. Furthermore, in a time profile clustering co-regulation patterns can be identified. Eventually, the time course integration allows to map downstream consensus transcripts with differentially abundant proteins. The "=" sign depicts the molecular overlap on each cellular layer, corresponding to the layer-specific consensus molecules.

abundant phosphoproteins, was performed. The transcription factors of these pathways were then found by matching the gene sets of the pathways against the transcription factors listed in the transcription factor—target gene database. Downstream target genes were identified, equivalently. The downstream analysis is based in general on the assumption of downstream regulation upon protein phosphorylation. Upstream analysis identified the upstream TFs of significantly differentially regulated transcripts. Subsequently, pathways including these TFs were identified in order to find possible upstream proteomic regulators of differentially expressed transcripts. The parameters chosen here corresponded to at least one TF per pathway for pathway identification and 10 orders of neighbors identified upstream of the TF for potential proteomic regulators. The results of each functional layer of signaling (pathway layer, TF layer, and gene/transcript layer) of downstream and upstream analysis were compared. These analyses steps were performed for each time point. Gene and protein ID matching was done by conversion of all IDs to HUGO gene symbols.

Static Consensus Analysis

In the static consensus analysis integrated signaling networks were constructed based on intersecting proteins, TFs, genes and transcripts on each functional layer (**Figure 4A**). The consensus proteins and TFs were mapped to the PPI STRING database and Steiner trees were generated via a shortest paths based approximation algorithm (Sadeghi and Fröhlich, 2013). The graphs were then completed by adding the corresponding TF—target interactions using TRANSFAC information. In case both consensus gene and consensus protein were part of the static consensus graph feedback loops were added.

Dynamic Consensus Analysis

In order to leverage the complete dynamic information from the data sets dynamic analysis was performed on basis of all consensus molecules (**Figure 4B**). The data associated with these nodes was used to fit cubic smoothing splines in order to generate a sufficiently dense data set for network inference via empirical Bayes estimation of a dynamic bayesian network with the R package ebdbNet (Rau et al., 2010). The generation of data points was based on the simplifying assumption of a gradual change of signaling over time. For further parameters default values were chosen. For visualization of the dynamic bayesian network a probability threshold was chosen which reflects a moderate number of regulatory interactions with a high probability in the network. The resulting threshold for plotting of the edges corresponded to a probability of an edge to be present by chance of 0.15.

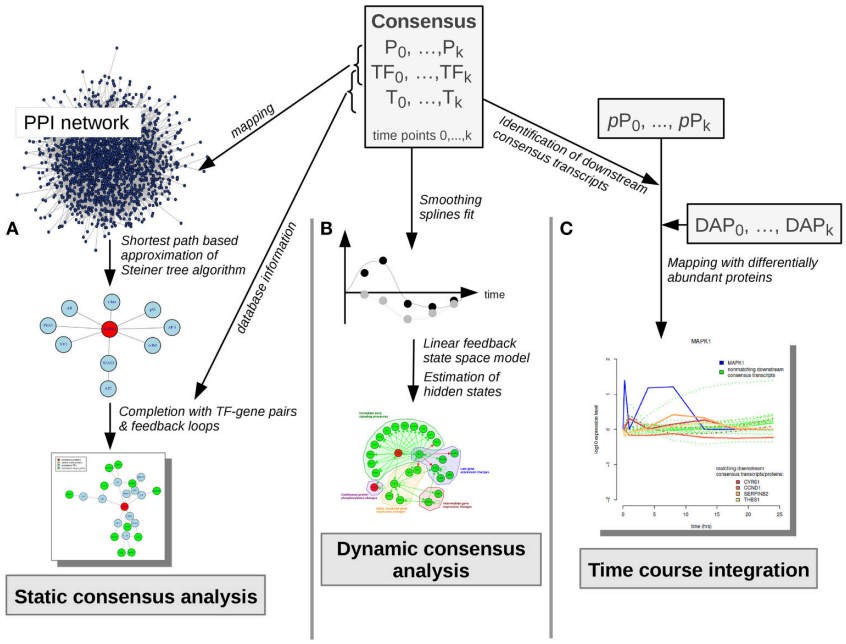


FIGURE 4 | Consensus molecule analyses. Consensus molecules on each cellular layer are used for the static consensus analysis, the dynamic consensus analysis and the time-course integration. **(A)** In the static consensus analysis static graphs are generated based on the PPI-mapped consensus proteins and transcription factors, an approximation of the Steiner tree algorithm is applied and the connected networks are complemented with TF-target interactions from TRANSFAC database. In case both consensus gene and corresponding consensus protein are part of the network, feedback loops are added. **(B)** In the dynamic consensus analysis smoothing splines are fitted to the time courses for all consensus molecules. Based on the higher density data set a linear feedback state space model is generated, hidden states are estimated and a probabilistic network is generated with dynamic Bayesian network inference (ebdbNet R package). **(C)** In the time course integration downstream consensus transcripts of the differentially abundant phosphoproteins are identified. These are mapped to the differentially abundant proteins. Time-courses of the downstream signaling players are visualized, subsequently. P, consensus proteins; TF, consensus transcription factors; T, consensus transcripts; pP, phosphoproteins; DAP, differentially abundant proteins.

Time Profile Clustering

Additionally, time profile clustering was performed in order to identify co-regulation patterns: Combining the described integration approach with a soft clustering implemented as fuzzy c-means algorithm (Kumar and Futschik, 2007) yielded an integrated time profile clustering based on the log-fold changes of consensus proteins and transcripts.

Time Course Integration

For further time course based integration with the proteome data set downstream consensus transcripts of the measured phosphoproteins were determined (**Figure 4C**). In a next step these were mapped to proteins, that were significantly differentially abundant at any time point (**Figure 2**, proteomic data).

RESULTS

Individual Downstream and Upstream Analyses

We performed individual downstream and upstream analyses of the phosphoproteome and microarray data sets taking into account the different functional layers of the cell the data originates from. The used pathway information exploits the signaling knowledge stored in public databases. **Figure 3** illustrates the steps of the individual analyses and further analysis steps explained in the next sections. **Table 1** shows the corresponding numbers of identified molecules and pathways on the different functional cellular layers in downstream and upstream analysis.

The data set for the phosphoproteome based downstream analysis is very small with only five phosphoprotein abundances investigated. However, as these were chosen thoroughly in the experiment we observe a considerable number of pathways that are influenced in downstream signaling. Altogether 121 pathways were identified when querying the four pathway databases used for the analysis. However, this set might include partly redundant pathways when originating from different databases, but describing the same signaling pathway. Pathways that are identified in every time point include e.g., the Biocarta "egf signaling" pathway, the NCI "EGF receptor (ErbB1) signaling pathway," the NCI pathway "EGFR-dependent Endothelin signaling events" or the NCI pathway "ErbB1 downstream signaling." Furthermore, a number of pathways are identified

TABLE 1 | Individual analysis.

Time after EGF stimulation [h]	0.25	1	4	8	13	18	24
DOWNSTREAM ANALYSIS							
No. of differentially abundant phosphoproteins	5	3	3	2	3	2	2
No. of pathways	121	68	98	90	81	79	79
No. of TFs	64	61	62	62	62	62	62
No. of potential target genes	1296	1293	1294	1294	1295	1295	1295
UPSTREAM ANALYSIS							
No. of differentially expressed transcripts	–	35	87	66	85	134	1551
No. of TFs	–	140	111	146	199	212	480
No. of pathways	–	163	154	169	200	200	230
No. of potential upstream proteomic regulators	–	871	950	897	920	976	1023

Downstream and upstream analyses characteristics over time. The expected bottleneck on the transcription factor layer can be observed. In the downstream analysis most pathways are overlapping, so we observe no large difference in the target gene numbers. The pre-processed proteomic data set comprises one time point of measurement more than the transcriptomic data set (0.25 h after EGF stimulation).

that are involved in cellular adhesion, STAT3 dependent signaling and PI3K signaling. Differential abundance of phopho-MAPK14 was only identified at time point 0.25 h after EGF stimulation. Corresponding pathways identified for that time point included e.g., the Biocarta "p38 mapk signaling pathway" and the Biocarta "mapkinase signaling pathway." According to the TF—target gene database the identified TFs activate the expression of a high number of genes as shown in **Table 1**.

In the transcriptome based upstream analysis an identification of upstream TFs was performed based on the differentially expressed transcripts. Corresponding numbers at each time point after EGF stimulation are displayed in **Table 1**. Identified upstream pathways included e.g., the "MAPK signaling pathway," the "EGF receptor (ErbB1) signaling pathway" and the "ErbB1 downstream signaling" pathway. The higher numbers of differentially expressed transcripts resulted likewise in the identification of more pathways. In those pathway sets the topological information enabled the identification of possible upstream proteomic regulators, subsequently.

The pathways identified in the downstream and upstream analyses at each measured time point after EGF stimulation are part of the Supplementary Material (**Tables S2**, **S3**).

Consensus Analysis

In the static consensus analysis we integrated the results of the different platforms for each time point on each functional layer. The aim was to reduce the individual downstream and upstream analyses results to molecule sets which include those molecules identified from both platforms and to reduce at the same time false positive molecules on the different functional layers. Exemplary, the consensus network of 1 h after EGF stimulation is shown in **Figure 5A**, later time point static consensus networks are part of the Supplementary Material (**Figures S2**–**S7**). These networks provide interaction and regulatory information on the consensus molecules. Yet, in our further analyses we focus on the static consensus profiles reflecting the presence of specific molecules in the consensus networks at each time point, as illustrated in **Figure 5B**. The static consensus profiles were used to explore the static consensus characteristics of certain molecules in order to evaluate the integration method. As dynamic signaling is especially interesting with regard to feedback signaling mechanisms and pathway crosstalk, we focus on these two signaling patterns in the following. **Figure 5B** shows the static consensus profiles of the members of the static consensus graph 1 h after EGF stimulation. A considerable number of genes being part of this consensus graph are exclusively found at this early time point. The profiles additionally show that both *PLAU,* the urokinase-type plasminogen activator, and *CTGF*, the connective tissue growth factor, comprise late regulatory changes. A figure with all static consensus profiles is part of the Supplementary Material (**Figure S1**). In these, 13 of 19 genes that are at least identified at two time points not including the 1 h time point after stimulation show a sustained pattern, indicative of a secondary cellular response. The genes without such a sustained pattern are *PLAU*, *CTGF* and *IL1A*, being already active 1 h after EGF stimulation or genes showing an intermediate activation.

Next, we investigated the pattern of proteins in the static consensus networks as well as the identified steiner nodes. The first group comprises the intersection of differentially abundant phosphoproteins in the proteomic data set and the potential upstream proteomic regulators of the differentially expressed genes. The second group is derived by generating Steiner trees after mapping the consensus molecules to the PPI network and might be functionally interesting, as its nodes are candidates for the regulation of the unconnected, mapped proteins. The static consensus profiles of the included proteins and the steiner node identified in this analysis are shown in **Figure 5C**. Transcription factor STAT3 is identified on the transcription factor layer at all-time points. MAPK1 is identified 4–8 h after EGF stimulation. PRKAR2B is identified later on (18–24 h after stimulation) on the protein layer. VAV1 is identified as a Steiner node in the static consensus graph 24 h after stimulation.

Additionally, we wanted to test in how far our integratory pathway-based approach is able to trace pathway crosstalk in the given data sets. In order to do so we chose a crosstalk mechanism which we expected to be reflected in the data set as it is not exclusively based on phosphorylation or ubiquitylation

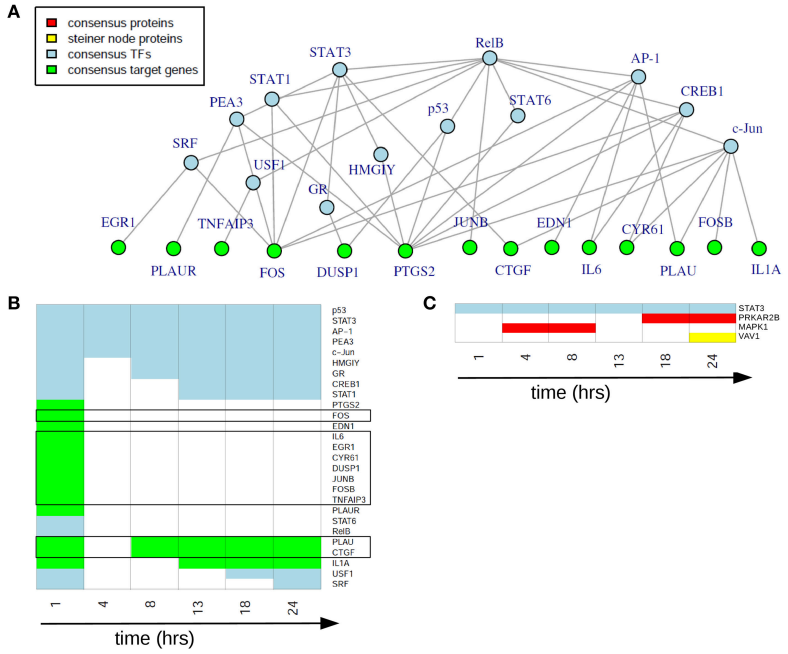


FIGURE 5 | Static consensus analysis results. (A) Static consensus graph for time point 1 h after EGF stimulation. **(B)** Static consensus profiles for members of the static consensus graph 1 h after EGF stimulation. Colors in the heatmap correspond to colors used in the consensus graphs, "white" boxes represent no membership in the consensus graph at that time point after EGF stimulation. Genes known to be IEGs (according to Tullai et al., 2007) are framed in black. **(C)** Static consensus profiles for selected proteins.

events. This mechanism is characterized by the activation of metalloproteinases (MMPs) by G-protein-coupled-receptors (GPCRs; Yarden and Sliwkowski, 2001). Upon activation MMPs cleave membrane-tethered ErbB ligands, which enables their binding to ErbB receptors, thereby positively regulating the ErbB signaling pathway. With EGFR being a receptor of the ErbB family our approach could identify a considerable number of the mentioned regulatory molecules in the consensus molecules (**Table 2**). Expression of different MMPs is observed starting at time point 4 h after EGF stimulation. Differentially expressed ErbB ligands for the different time points after EGF stimulation could be identified (such as self-induced EGF and AREG).

TABLE 2 | Consensus analysis.

Time after EGF stimulation [h]	1	4	8	13	18	24
MMPs	–	MMP1	MMP1	MMP1	MMP1	MMP1
			MMP1	MMP1	MMP1	MMP2
			MMP1	MMP1	MMP1	MMP10
ErbB ligands	–	–	–	EGF	AREG	AREG
					EGF	EGF

Regulatory molecules identified on the gene layer that are hypothesized to be involved in the signaling crosstalk via GPCRs and MMPs. GPCRs activate MMPs which then cleave the membrane-bound ErbB ligands leading to activated ErbB signaling (Yarden and Sliwkowski, 2001). Although differential expression is not direct evidence for the activity of these molecules, such regulatory mechanism can be hypothesized here.

Exploiting Dynamic Information of Coupled Time Course Data Sets

Our pathway-based approach additionally enables the utilization of the complete time-series for each molecule in order to generate a probabilistic network displaying those nodes of the network with a high posterior probability of interaction. The dynamic analysis is based on the simplifying assumption of a gradual change in signaling over time, as existing high-frequency components are not considered due to the small sampling rate. Each consensus molecule at any time point after EGF stimulation was taken into account. With this approach we obtained the probabilistic network displayed in **Figure 6**. This network is a reduced way to look at activating or inhibiting relationships between consensus proteins and genes. Here, we observe mainly activating relationships corresponding to an activation of the regulatory effect of EGF stimulation and not to upregulation directly. Likewise an inhibiting relationship in the network does not imply a downregulation, but the inhibition of the effects induced by EGF stimulation.

In total, we could identify five subgroups in the consensus-based dynamic network by mapping them to the times in which they are part of the consensus graphs (**Figure 6**): (1) immediate

early signaling processes, (2) early, but sustained gene expression changes, (3) intermediate gene expression changes, (4) late gene expression changes, and (5) continuous protein phosphorylation changes. In the group of the "immediate early signaling processes" most early response genes that were identified in the static consensus profiles are activated by the protein MAPK1 and the gene *IL1A*. This group reflects early phosphorylation induced transcriptional changes. The next group, consisting of five genes, is the group of "early, but sustained gene expression changes" upon EGF stimulation. It includes *CTGF*, a connective growth tissue factor. Its regulation is activated by MAPK1, *FKBP5*, *GADD45A* and also self-activation is observed. *CTGF* itself has activatory influence on gene members of its own group (*IGFBP3*, *FKBP5*), but also on members of the "intermediate gene expression changes" group and the "late gene expression changes" group. Two further members (*PLAU* and *ODC1*) are influenced by *IL1A*, a hub gene in the network, which we assigned to the "immediate early signaling processes" group and to the "late gene expression changes" group, as it shows immediate membership in the static consensus graphs, but also a late response profile. A small group showing intermediate gene expression changes comprises *TFRC* and *GADD45A*. We observe in the graph that *GADD45A* activates itself, but also *PCNA*, a gene of the "late gene expression changes" group. *PCNA* is additionally self-activated, as well as externally activated by the ErbB ligand *AREG* and *ASPH*, the aspartate beta-hydroxylase. *AREG* and *ASPH* are upregulated late after EGF stimulation. *IL1A* also activates *SLC3A2*, the solute carrier family 3 member 2, and inhibits *LAMA3*, a proliferating cell nuclear antigen, laminin alpha 3. The second protein being part of the network is the transcription factor STAT3. The changes in STAT3 phosphorylation are found in the consensus graphs over all time points, thus we assign it to the group of "continuous protein phosphorylation changes." Beside the activating influence of MAPK1 also autoregulation of STAT3 can be detected.

Time Profile Clustering

In order to identify co-regulation patterns in the signaling response after EGF stimulation we performed time profile clustering. We obtained four dynamic co-regulation patterns of which two exhibit positive regulation and two exhibit negative regulation. Both positive and negative clusters each comprise one cluster of immediate regulation and one of delayed regulation. The clusters are depicted in **Figure 7**. Corresponding molecule membership in the four different clusters is listed in the Supplementary Material (**Table S1**). Cluster 1 is immediately activated and thus contains various immediate early genes, but also the proteins MAPK1 and STAT3, which are part of the consensus-based dynamic analysis. Compared to the groups identified in the latter analysis this cluster constitutes the immediate early signaling processes together with early, but sustained gene expression changes. Cluster 2 is the biggest cluster with 52 members and is the delayed positively regulated cluster. Cluster 3 only comprises two members (RARRES3 and SLC3A2), both of which are showing a delayed negative dynamic co-regulation. Cluster 4 is the early negatively regulated cluster.

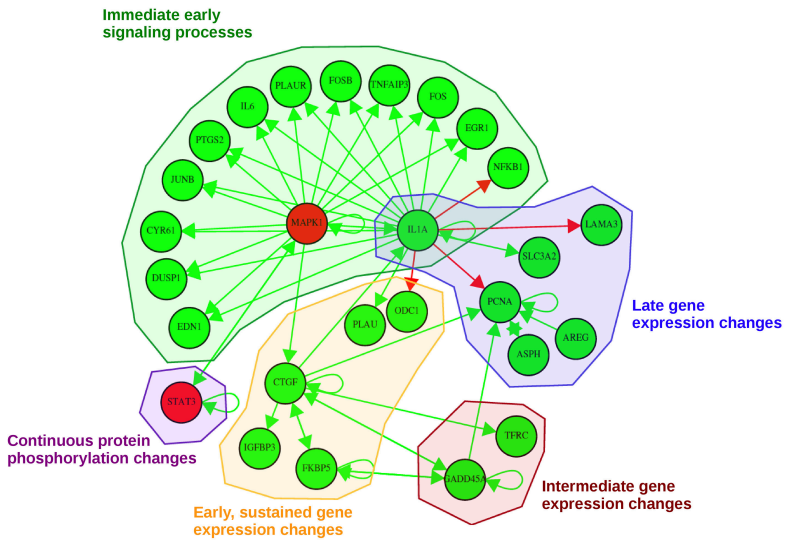


FIGURE 6 | Probabilistic network displaying result of the consensus-based dynamic analysis. For network inference all consensus genes and proteins at any time point were considered. For visualization of the dynamic bayesian network a probability threshold was chosen corresponding to a probability of an edge to be present by chance of 0.15. Five groups could be identified comprising direct immediate early signaling processes, continuous protein phosphorylation changes, late gene expression changes, intermediate gene expression changes and early, but sustained gene expression changes upon stimulation. Activating regulatory effects are depicted with green edges whereas inhibiting regulatory effects are depicted as red edges. Consensus protein nodes are colored in red, consensus transcript nodes in green. Activation/inhibition refers to changes in the regulatory effects initiated by EGF stimulation, not to activated or inhibited expression.

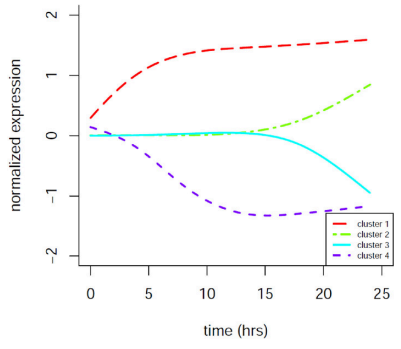


FIGURE 7 | Fuzzy c-means time profile clustering revealed 4 co-regulation clusters with distinct cluster sizes. Two of the clusters exhibit positive regulation and two exhibit negative regulation. Both positive and negative clusters each comprise one cluster of immediate regulation and one of delayed regulation. The clusters contain both protein and gene expression changes. Cluster membership is listed in the Supplementary Material.

Time Course Integration

The results of the time-course integration based on the consensus analysis results are displayed in **Figure 8** and in the Supplementary Material (**Figure S8**). Of the five phosphoproteins that were measured over time in the coupled data set we could identify four phosphoproteins with their downstream transcripts being part of our consensus analysis and mapping to differentially abundant proteins (MAPK1, STAT3, MAPK14, and PRKAR2B). MAPK1 downstream analysis revealed four transcripts (**Figure 8A**), which mapped to significantly differential proteins, CYR61—cysteine-rich angiogenic inducer 61, CCND1—cyclin D1, SERPINB2—serpin peptidase inhibitor, clade B, member 2, and THBS1—thrombospondin 1. MAPK1 itself shows increased phosphorylation levels in the very beginning after EGF stimulation and again between 1 and 13 h after EGF stimulation. In regard to temporal coordination CYR61 shows correlating temporal expression on the transcript and protein layer up to time point 4 h after EGF stimulation, but then a rather opposed pattern. CCND1 belongs to the group of cyclins and thus exhibits a specific expression and degradation pattern over the cell cycle, in this way contributing to the temporal coordination of mitotic events. Here we can observe an opposed temporal pattern of transcripts and proteins over the whole timespan measured: While on the mRNA layer, CCND1 shows higher expression levels after EGF stimulation, the corresponding proteins are found at lower levels over the whole time course. High mRNA-to-protein levels have already been reported by Waters et al. (2012). In the time-course SERPINB2 shows slowly rising levels of transcripts after EGF stimulation, whereas on the protein layer there is a direct decrease, an intermediate increase, and a second decrease again to the 0-level at 18 h after EGF stimulation. THBS1 protein levels are similar to that of SERPINB2, however, here we observe rather correlating transcript levels in the beginning and deviating ones after the 18 h time point.

STAT3 is the phosphoprotein showing the most downstream transcripts that match to significantly regulated proteins (**Figure 8B**). STAT3 itself shows sustained high expression levels over the whole time-course. All MAPK1 downstream transcripts that are part of the consensus analysis also belong to the downstream transcripts of STAT3. Further ones are *SLC3A2*, *FKBP5*, *PPP2CA*, *CD44,* and *ODC1*. All of these except for *ODC1* show anti-correlating patterns between transcripts and proteins until 4 h after EGF stimulation. For later time points most pairs exhibit correlating behavior. MAPK14 also has *CYR61*, *CCND1,* and *SERPINB2* as downstream targets with corresponding proteins being significantly differentially abundant, whereas for PRKAR2B only *CYR61* could be identified.

DISCUSSION

Pathway Layer Based Integration

In the downstream and upstream analyses the results indicate that pathway identification based on differentially abundant phosphoproteins and differentially expressed transcripts is effective. In both pathway sets those pathways known to be activated by EGF stimulation were identified reliably in the different databases, expectedly the "EGF signaling pathway" itself. This shows, that the two data sets are in concordance on the pathway layer even if they are measured on different cellular layers and analyzed individually. Based on these initial results a pathway-based integration was considered to be constructive. However, downstream and upstream analyses might also introduce false positive findings, which we aimed to reduce from further analysis steps by the subsequent intersection analysis. The small set of phosphoproteins measured over time gives a strong basis for the pathway layer based integration as they were selected carefully for the experiment and belong to key pathways in EGF signaling. However, a larger set of phosphoprotein data as obtained now e.g., from mass-spectrometry approaches could lead to more robust results.

Consensus Analysis Enables Identification of Regulatory Dynamics

In order to evaluate our methods it is important to first classify the data according to their temporal transcriptional domains. According to Avraham and Yarden (2011) feedback mechanisms in EGFR signaling can be assigned to two temporal domains, one of them being the immediate group which includes receptor endocytosis, secondary phosphorylation and further protein modifications, the other constituting the late group which includes newly synthesized adaptors, transcriptional repressors, RNA-binding proteins and phosphatases of the mitogen-activated protein kinase (MAPK) pathway. Especially the integrated data with parallel time points between 1 and 24 h after EGF stimulation thus reflects the late group capturing the transcriptional regulation with a wave-like regulation of

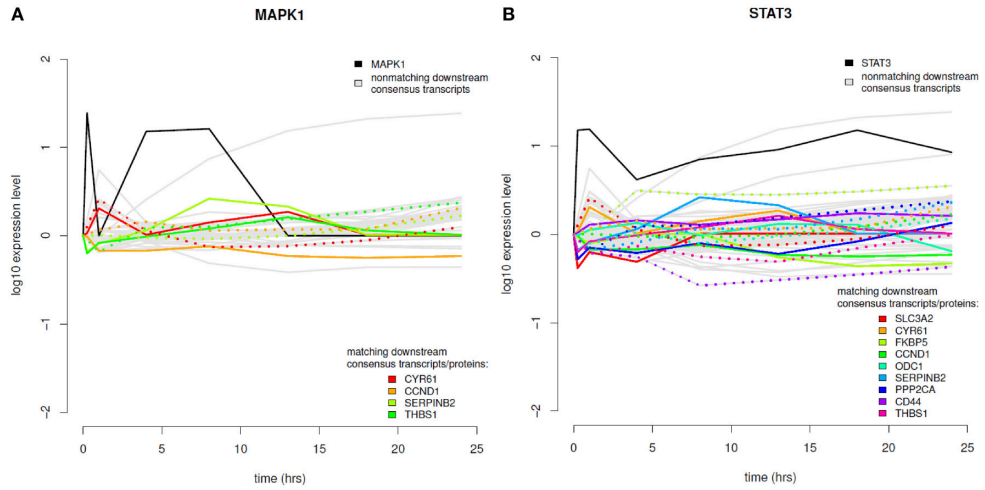


FIGURE 8 | Time course integration. (A) MAPK1 downstream consensus transcripts identified were mapped to differentially abundant proteins. **(B)** Time course integration for downstream consensus transcripts of STAT3. Note that the measurement range of the expression profiles across platforms can vary. Phosphoprotein time course data is shown in solid, black lines, non-matching transcript data in solid, gray lines and matching transcript and proteome data in rainbow color palette with proteins depicted as solid lines and transcripts depicted as dotted lines.

immediate early genes (IEGs), delayed early genes (DEGs), secondary response genes (SRGs; Avraham and Yarden, 2011) and their corresponding subsequent protein expression. IEGs are known to induce transcriptional changes of DEGs which then reduce the regulation of IEGs in a feedback subsequently, but initiate regulation of SRG expression. Based on this transcriptional regulation scheme the measured time points in the investigated data sets capture stimulation of both IEGs and DEGs 1 h after EGF stimulation while in subsequent time points we expect only regulation of SRGs, conferring the stable cellular phenotype.

We used the static consensus analysis in order to generate a static view on the integrated networks at each time point. Via static consensus profiles we can identify transcription factors with regulatory effects and their regulated consensus molecules on the gene layer at the 1 h time point. A large number of those genes were already reported to be IEGs in the cellular response to growth factor stimulation according to Tullai et al. (2007). PLAU and CTGF, regulated as well at later time points, apparently have an additional function in the definition of the phenotype. The two-phase regulation pattern indicates 2-fold tasks and can be interpreted to underly direct or indirect auto-feedback regulation.

The static consensus profiles of most SRGs, in contrast, are supposed to show a sustained activity. This is exactly what we find in our consensus graph analysis.

Due to the low number of differentially abundant phosphoproteins as a starting point the number of intersecting proteins from downstream and upstream analyses are low, as well. MAPK1 is involved in a variety of cellular growth processes such as proliferation and differentiation, thus its presence in the consensus graph corresponds well to the expected cellular response after EGF stimulation. As a regulatory subunit of the cAMP-dependent protein kinases PRKAR2B is involved in various cellular functions. With its late activity we suspect an involvement in the cellular reconstruction processes taking place for the final phenotype definition. The VAV proteins are guanine nucleotide exchange factors that activate pathways leading to cytoskeletal actin rearrangements and transcriptional alterations (Han et al., 1998). Thus, its functional association can be linked to cellular restructuring during proliferation.

In EGF signaling several pathways are involved which do not only process signals in a linear way but also enable cross-pathway regulatory influence on transcription. Oda et al. (2005) tried to compress all known signaling interactions into a comprehensive pathway map, resulting in a bow-tie architecture signaling pathway. As this network has to convey fine-tuned messages, it is deducable that slight dysregulation results in pathological transcriptional responses. Many crosstalk mechanisms have been investigated in more detail, most of them under pathological conditions. However, in order to understand the consequences of such dysregulation it is essential to also have a detailed understanding of physiological pathway crosstalk mechanisms. This is why we reviewed the consensus molecules in terms of their possible role in the crosstalk described by Yarden and Sliwkowski (2001). The large number of identified consensus molecules implicated in this crosstalk on the gene layer supports our hypothesis, that they are part of this signaling crosstalk mechanism.

As the described regulatory dynamic patterns are based on two independent data sets from different platforms we suppose that

this pattern is not identified due to measurement bias and thus has a biologically relevant function in the cellular response.

Identification of Regulatory Mechanisms by Exploiting Dynamic Information of Coupled Time Course Data Sets

In order to fully exploit the dynamic information of the time course data sets, we inferred a probabilistic network based on all consensus molecules. This network enables an identification of important players in the cellular response to EGF as well as the determination of inhibitory or activating regulation patterns.

The consensus proteins which are part of the dynamic network are MAPK1 and STAT3, both being part of the starting phosphoprotein data set. This indicates, that their important role in EGF signaling can be confirmed as such via the transcriptomic data set. STAT3 is a transcription factor, which is phosphorylated upon growth factor stimulation of the cell and builds homo- or heterodimers, which can then translocate to the nucleus and activate transcription (Park et al., 1996). It has multiple target genes with its protein products being involved in proliferative processes. MAPK1 is associated with cellular processes such as proliferation, differentiation and transcriptional regulation. Both show a self-activation as well as a mutual activation, which illustrates their functional relevance in EGF signaling. This regulatory interaction between MAPK1, also known as ERK2, and STAT3 is triggered via the activation of the MAPK/ERK cascade upon EGF stimulation, leading to MAPK1 phosphorylation by upstream kinases. STAT3 transcriptional activation by phosphorylation of STAT3 pS727 is then performed by the serine/threonine kinase ERK (Zhang and Liu, 2002), leading to activation of STAT3, which then acts as transcription factor and initiates the expression of downstream target genes. Target genes of STAT3 that might lead to further activation of MAPK1 are e.g., downstream transcription factors, multiplying indirectly the effective activation, or EGFR allowing for binding of more EGF. Furthermore, JAK2 is a target gene of STAT3, which can contribute to positive auto-feedback of STAT3 via the JAK-STAT pathway (Dauer et al., 2005).

Beside the already discussed early regulation processes and the protein phosphorylation changes of STAT3, the other identified groups are particularly interesting for further interpretation: The regulation of *CTGF*, the connective growth tissue factor, is activated by MAPK1, *FKBP5*, *GADD45A* and by itself. Interestingly, we observe auto-feedback regulation here, as already suspected from the static consensus profiles. *CTGF* is a hub gene in the consensus-based dynamic network, so the activation of its downregulation upon EGF stimulation is associated with downregulation of other genes in this cluster, such as *FKBP5*, or genes of the "intermediate gene expression changes" group. One of these is *GADD45A*, the growth arrest and DNA-damage-inducible alpha, which activates the regulation of PCNA. It is known to comprise increased transcript levels when cells are subjected to arrest conditions, treatment with DNA-damaging agents and environmental stresses (Hollander et al., 1993), thus we suspect the experimental design of the experiment with the chosen growth arrest time to be of no direct harm to the cells. PCNA, the proliferating cell nuclear antigen, is a cofactor of DNA polymerase delta and plays a central role during DNA replication. In DNA damage response it is positioned at the replication fork coordinating replication with DNA repair and DNA damage tolerance pathways (Cazzalini et al., 2014). Thus, its function is intensely needed in the phase of cellular remodeling and proliferation. The link between *GADD45A* and *PCNA*, that we determined with our integrative analysis, was previously reported (Chen et al., 1995).

AREG is upregulated in the "late gene expression changes" group as part of the regulatory pathway crosstalk loop via metalloproteinases described above and presumably provides an additional amplifying cellular way of an activation cascade after initial EGF stimulation. Also ASPH, which is thought to play an important role in calcium homeostasis (Treves et al., 2000), is part of this group. With its diverse roles e.g., as a messenger between cellular compartments calcium regulation is essential for proliferating cells.

IL1A, as another hub in the network, has immediate and late regulatory influence. In the "late gene expression changes" group it activates *SLC3A2*, solute carrier family 3 member 2, and inhibits *LAMA3*, proliferating cell nuclear antigen, laminin alpha 3. With their functions in regulating intracellular calcium levels, amino acid transport, formation and function of the basement membrane, cell migration and mechanical signal transduction and DNA replication, this part of the network rather shows the expression changes which represent the secondary (late) response of the cells.

In summary, we identified MAPK1, *IL1A* and *CTGF* as main players driving EGF stimulation response in the cell. Interestingly, we could detect the link between *GADD45A* and *PCNA* in two independent high-throughput time course data sets measured on different platforms using our pathway-based integration approach. As a matter of course, with a higher temporal resolution of the coupled time course measurements more accurate results can be identified by our approach, as less intermediate time points need to be estimated. To gain insight into the biological response after an external stimulation at least four time points after the stimulation time point are necessary, though there is a high information content in such coupled data sets on the different cellular layers. The chosen time points and the temporal resolution, however, need to be adjusted specifically to the cellular signaling dynamics and the stimulation of choice in order to reflect the crucial time points of regulation.

Time Profile Clustering Identifies Four Dynamic Co-Regulation Patterns Ruling EGF Signaling

With our time profile clustering approach we could identify four co-regulation patterns with distinct functions in the cellular response to EGF signaling. Cluster 1 contains many of the directly upregulated immediate early genes. Most of these are in fact downregulated again after their early response, which is not reflected by this cluster, as it contains also a considerable number of genes that are secondary response genes and are only upregulated at later time points (such as MMP1 or MMP10)

or immediate early genes which are upregulated again at later time points (PLAU or IL1A). Our hypothesis, that cluster 2 includes mainly genes upregulated as secondary response genes, responsible for the phenotype definition, holds true, when having a closer look to the members: We observe *CCND1*, the cyclin family protein, *ANXA1* and *ASPH*, *LAMA3* and *AREG*, which were identified in the consensus-based dynamic analysis in the group of late gene expression changes, *VEGFC*, a vascular endothelial growth factor promoting angiogenesis, *CCND2*—cyclin D2, *NME1*—nucleoside diphosphate kinase 1, which has been associated with high tumor metastatic potential based on different studies (MacDonald et al., 1996) and many more genes which act during cellular proliferation and migration. As cell cycle inhibitory protein coding genes we can observe the membership of *CDKN1A*, the cyclin-dependent kinase inhibitor 1A, which is tightly controlled by transcription factor p53 (He et al., 2005). Its membership in cluster 2 might be due to the high importance of balancing proliferation processes against growth stimulating processes in physiological tissue. Further we observe *PTHLH*, the parathyroid hormone-like hormone, to be part of this cluster, which regulates the epithelial-mesenchymal interactions during formation of mammary glands and teeth (Wysolmerski, 2012). Additionally the protein PRKAR2B is part of this cluster, indicating its late activation, which we already observe in the phosphoproteome data individually. However, here we see the confirmation that it is part of the consensus data from the two independent data sets generated on different platforms. Also *MMP2* is part of cluster 2 as well its regulatory counterpart, *TIMP1*, a metallopeptidase inhibitor. As the other metalloproteinases identified in the static consensus graphs (*MMP1* and *MMP10*) are not members of cluster 2, but of the immediately positively regulated cluster 1, it can be assumed, that *TIMP1* activation might also have a negative regulatory impact on these late after EGF stimulation. In the delayed downregulated cluster 3 we observe RARRES3, the retinoic acid receptor responder 3, which is known for its growth inhibitory effects (Hsu and Chang, 2015). A late downregulation thus can have the function of preventing contrasting growth signals. SLC3A2, the solute carrier family 3 member 2, encodes a subunit of a cell surface transmembrane protein complex responsible for regulation of L-type amino acid transport, which is essential for cellular growth and proliferation (Yanagida et al., 2001). Cluster 4, the early negatively regulated cluster, comprises *CTGF*, the connective tissue growth factor, whose downregulation might enhance proliferation of cells upon EGF stimulation. A further member is *IGFBP3*, the insulin-like growth factor binding protein 3, which potentiates insulin-like growth factor action and thereby also stimulates growth promoting effects (Cubbage et al., 1990). Supposedly, the cells do need less proliferating activation via IGF, when there is the growth-promoting stimulation of EGF. This underlines again that signaling patterns are tightly regulated in regard to their dynamics.

Time Course Integration of Consensus Graphs with Proteome Data

We were interested in how far our approach reveals the dynamics of elements in the regulatory cascade of a stimulation induced phosphorylation cascade triggering a specific gene expression, which then leads to the generation of proteins needed in the cellular response to that particular stimulation. Therefore, after integrating the phosphoproteome data in the first pathway layer based integration, we integrated in a second step also the proteome data with the results of our pathway-based integrative analysis dynamically. The delay between consensus transcript generation and their corresponding protein generation reflects the time the cell needs for the complete translational and post-translational process. However, it is known that differences in protein abundance are only attributable to mRNA levels by about 20–40% (Brockmann et al., 2007). This underlines the importance of post-translational modification and is the reason why we assumed the correlation between increasing and decreasing transcript expression and corresponding protein generation to be rather marginal.

For the interpretation of these results we need to be aware of the different ranges of the expression ratios in the data sets of different platforms. Thus, a direct comparison of the expression levels between transcripts and proteins is not possible, however, a dynamic interpretation is feasible.

Dynamically, we observe both correlating and non-correlating expression level patterns between transcripts and corresponding proteins. Based on the time resolution of the measurements we assume the time delay reflecting the translational and post-translational processes to be not necessarily observable in the data, as they can lie in a wide time range. Indeed, correlating behavior seems not to be shifted in time in our analysis for certain transcripts (e.g., for CYR61 up to 4 h after EGF stimulation or THBS1 up to 13 h after EGF stimulation), however, when performed on a time-series data set with higher resolution, such time shifts might be observable. Non-correlating expression level patterns indicate post-translational modifications or a possibly very rapid degradation of mRNA or the protein product, which is not captured in the low resolution time measurements. Of the identified pairs CYR61 is a growth factor inducible protein which promotes the adhesion of endothelial cells (Brigstock, 2002), CCND1 is a protein contributing to coordination of mitosis. High levels of SERPINB2 have been observed to exhibit an anti-proliferative effect (Croucher et al., 2008). In the time courses we see an intermediate increase of its protein levels, but an overall anti-correlating pattern between protein and transcript levels. THBS1, thrombospondin 1, is known as angiogenesis regulator (Chandrasekaran et al., 2000). Its protein levels are similar to that of SERPINB2, however, here we observe rather correlating expression levels, indicating less post-transcriptional modification. Also changes in the correlation behavior can be observed, indicative for a secondary regulatory influence. This could be induced by variations in mRNA degradation, protein degradation rates or post-translational modifications.

From the transcript/protein pairs that are observed as part of the regulatory loops CYR61, THBS1, and CCND1 clearly have a high influence on EGF stimulated cells during cellular proliferation, differentiation and survival, while the detection of SERPINB2 is more intriguing. It is known to inhibit urokinase plasminogen activators (PLAUs), but its physiological function has not been characterized comprehensively, although activity

in the adaptive immune response has been reported (Schroder et al., 2011). As we based the time-course integration on the consensus analysis the discussed time-courses are supported by both transcriptome and proteome data set. Thus, we hypothesize the interaction of SERPINB2 and PLAU, its inhibition target, to be of high relevance for proliferative processes. Our hypothesis is supported also by literature in the context of cancer: SERPINB2 has been associated with increased survival in breast cancer patients (Duffy, 2004).

With the integrated time-courses of phosphoproteins, downstream consensus-graph transcripts and their corresponding proteins the data implies an extensive post-translational modification of a number of proteins. This we see in the transcript/protein pairs investigated in detail here, but also in the downstream transcripts depicted in gray in **Figure 8**, with no corresponding proteins in the list of significantly differentially abundant proteins. Therefore, our results correspond to what is known about the low percentage of protein concentration variations that are affected by mRNA abundances directly (Vogel and Marcotte, 2012). However, our approach not only enables a general overall classification of correlating or anti-correlating transcript/protein pairs, but in addition a time-resolved interpretation of consensus-based regulatory processes.

Comparison of Separate Data Set Analysis with Integrated Consensus-Based Analysis

To comprehensively assess the advantage of our data integration approach based on public pathway knowledge we compared its results with the ones gained by a separate analysis of the individual proteomic and transcriptomic data sets. Waters et al. (2012) performed a separate pathway analysis and reported network statistics, such as the number of nodes in the largest cluster, the number of edges in the network and the two primary hub nodes, however, this analysis was limited to data measured 0–4 h after EGF stimulation. Interestingly, the hub genes identified in the microarray based network were the transcription factors *FOS* and *EGR1*, while the hub proteins identified in the proteome data were EGFR and ITGB1. Comparing these results to our results from the pathway-based integrative analysis, we likewise observe FOS and EGR1 to be highly important regarding regulatory mechanisms during the initial cellular response. Yet, we additionally derived further information than what is given by the separate analysis: We evaluated these genes to play a significant role in the immediate early cellular reaction based on static consensus profiles. Furthermore, we saw that these are mainly influenced by *IL1A* and the phosphorylation of MAPK1 directly as well as indirectly. Based on the time profile clustering we saw on top that they belong to the early positively regulated cluster. The protein hubs that are identified via the separate analysis, however, cannot be found in our consensus analysis, as the consensus is confined to the small set of measured phosphoproteins.

In a second separate analysis of the proteomic and transcriptomic data sets Waters et al. (2012) performed separate gene set enrichment on the basis of differentially expressed proteins and transcripts. The three most significant biological processes identified for the transcriptomic data set were "cell cycle," "mitosis," and "protein folding," while for the proteomic data set the most significant process was "protein synthesis." In a comparison the authors found considerable differences in the gene set enrichment results. Although this type of analysis is widely used for gene expression data it is arguable in how far "gene set" and "protein set" enrichment should be compared directly due to the different biological layers the data and possibly also network knowledge originates from. Thus, we see an inherent problem in the simplified layer-unspecific comparison with subsequent interpretation. Additionally, the results allow no conclusions or hypothesis generation on the molecular level.

In summary, we conclude that the integrated analysis of the two data sets moves the focus to the dynamic interplay of regulatory mechanisms and enables a layer specific and detailed regulatory analysis of the cellular response to external stimulation.

Comparison of Data Integration Approaches in Coupled High-Throughput Data Sets

The data integration approaches applied by Waters et al. (2012) were based on RNA/protein pairs cross-referenced between the platforms. However, no layer-specific analysis was performed. In a canonical correlation analysis the 199 RNA/protein pairs comprising all measurement time points were investigated with the result of intense post-transcriptional regulation on the protein layer. The benefit compared to a simple correlation analysis is that it captures also concordance or disconcordance of pairs when a temporal delay is observed. With our time-course integration we could also observe this effect, individually for specific phosphoprotein initiated signaling cascades. With our approach it is additionally possible to analyze transcriptional and translational dynamics of each cascade individually.

In the integrative analysis of Waters et al. (2012) major cell processes of the combined data were then ranked to early (0–4 h), intermediate (8–13 h) and late (18–24 h) time domains after EGF stimulation. A general shift from categories "cytoskeletal organization" and "regulation of cell cycle" (0–4 h) toward anti-apoptotic and cell adhesion pathways (8–13 h) was observed. An increased representation of the "mitosis" category between 18 and 24 h after stimulation corresponded to an increase of mitotic cells monitored by flow cytometry in parallel. A direct comparison of the analyses results is not possible here, though the results we found in the consensus-based dynamic analysis of the data agree roughly with the results of Waters et al. (2012), when comparing the function of individual consensus molecules with the GO biological process category names. Although having category names enables in general a better overview of the data, it does not allow individual identification of regulatory interactions. Therefore, we consider our approach as valuable additional method in order to get a better understanding of the dynamic biological processes.

Furthermore, integrated signaling networks from all data sets were investigated in Waters et al. (2012). Not surprisingly, the microarray data set contributed the highest number of nodes in the merged network. Compared to the signaling networks from

single data sets, the integrated network comprised increasingly linked nodes, reflected in the number of edges and the degree of the largest cluster reported. The two primary hub nodes of the integrated network were *FOS* and *SRC*, while the hub nodes in the network generated from exclusively microarray data were *FOS* and *EGR1*, generated exclusively from proteome data EGFR and ITGB1 and exclusively from phosphoproteome data STAT3 and MAPK1. Interestingly, we also found *FOS* and *EGR1*, as well as STAT3 and MAPK1 as consensus molecules in our consensus-based dynamic analysis with considerable regulatory influence during the cellular response after EGF stimulation. The proteome hub nodes EGFR and ITGB1, as well as the hub node *SRC* from the integrated network were not part of our results due to the low number of phosphoproteins measured in the study. However, we found already considerable amount of regulatory mechanisms when including only the phosphoproteome data set as initial data set in our analysis. The MMP cascades identified in the integrated analysis from Waters et al. (2012) as most robust response to EGF stimulation were identified as consensus molecule based process by our approach as well.

Unfortunately, in the integrated analysis of Waters et al. (2012) only time domains were considered in contrast to our individual time point analysis. This enables a rough summarized view on the signaling process, yet it does not fully exploit the information encoded in the dynamics. Likewise, the GO term analysis performed is based on a subset of RNA/protein pairs and results in a summarized interpretation, but it does not enable an individual regulatory mechanistic interpretation. Thus, we consider our approach as valuable complement in the analysis of coupled high-throughput data sets.

CONCLUSION

The presented data integration approach shows a way to gain a much deeper understanding of biological processes if time-course measurements and data from different high-throughput platforms representing the different functional layers of the cell are combined. Our approach enables a functional linking of regulatory processes over the transcriptional and translational cycle, even if the temporal resolution of the example data set is quite low, data has only been measured on two functional cellular layers and the phosphoproteome data set is very limited. This sets the basis for the integration of further cellular layers, as following regulation upon external perturbation in a detailed way provides a much deeper understanding of biological processing.

Bioinformatic tools like the R package *pwOmics* promote the generation of coupled data sets as they offer the possibility of an integrated analysis and help to sort the vast data sets in a biologically interpretable manner. By applying the different analysis steps implemented in *pwOmics* we showed that biological interpretation is facilitated and the results correspond to current biological knowledge about EGF stimulation generated in low and high-throughput experiments. Furthermore, we identified interesting regulatory relationships that were not observed yet in physiological EGF signaling. As our approach considers data from the different functional cellular layers individually, it enables to identify the regulatory interplay between these layers. We have demonstrated this in the consensus analysis, which is able to identify the molecular response minutes to hours after stimulation as feedback mechanism with a wave-like regulatory pattern generated by IEGs, DEGs, and SRGs and their corresponding proteins. We could also identify previously published pathway crosstalk via activation of MMPs (Yarden and Sliwkowski, 2001). Furthermore, we could ascertain the link in EGF signaling between the two molecules GADD45A and PCNA, in the investigated data sets, which was previously reported (Chen et al., 1995). Interestingly, we also found PTHLH in the consensus molecules as part of the secondary cellular response, which is involved in the formation of mammary glands (Wysolmerski, 2012). Furthermore, we could identify the regulatory interaction of PLAU and SERPINB2 to be also of high relevance in physiological EGF signaling. Compared with the previously performed integrative analysis on the coupled data set we gain a complementary, and much more detailed view on cellular signaling processes, enabling the generation of biological hypothesis about individual regulatory mechanisms involved in the dynamic interplay of signaling pathways and feedback responses. With the examples stated above we could show, that our integrative approach is able to identify regulatory patterns, molecular interactions and dynamically orchestrated cellular response mechanisms.

In order to link the different functional cellular layers it is beneficial and necessary to integrate knowledge from public databases which builds a frame for placing and linking the individual analysis results. This has the advantage of utilizing a vast amount of collected and curated information, which stays unused otherwise and can add an additional information layer for interpretation of the data. On the other hand this prior knowledge also directs the results in a certain extent, thus the quality of the databases used has to be taken into consideration when interpreting the overall results. A further caveat is that the public database knowledge available in most databases is not cell type or tissue specific resulting in a generalized analysis. However, as more cell type or tissue specific knowledge is collected such databases can be build up and integrated in the presented analysis workflow.

In the consensus-based dynamic analysis we make the simplifying assumption of a gradual change of signaling over time. Clearly, this does not hold true for individual cells and still is a rough assumption for a set of cells as there have been found oscillatory mechanisms which work at high frequencies (Avraham and Yarden, 2011), for example, and which are purely not identifiable via such a time resolution. However, we can still gain a lot of knowledge about the regulatory processes that are encoded in the comparably slow dynamic processes. Of course, there can be even more biologically functional layers measured in high-throughput experiments in a parallel manner over time, such as siRNA, epigenetic influences etc. At the moment such data sets are still rare, but we expect them to be generated increasingly. It will be interesting for future projects to include such additional layers into an integrative analysis.

We showed that the hypotheses on regulatory mechanisms generated via our integrative approach could be confirmed with

independent low-throughput data sets. Although such time-course data sets measured in parallel enable a detailed analysis, it is not yet possible to infer from these data sets every regulatory aspect in detail. Nevertheless, our approach is a step toward portraying the whole picture of regulatory influences on the molecular level.

AVAILABILITY

Main analysis steps of the pathway-based integration approach of coupled time-series omics data described in this manuscript are implemented in the R package *pwOmics* (Wachter and Beissbarth, 2015).

AUTHOR CONTRIBUTIONS

AW developed the method, performed data analysis and wrote the manuscript. TB conceived the design, envisioned the project and revised the manuscript.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support by BMBF e:Bio program grant MetastaSys [0316173A] and by BMBF e:Med grant MMML-Demonstrators [031A428B]. We additionally acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2015.00351

Figure S1 | Static consensus profiles of all members of the static consensus graphs. Color coding corresponds to the one used in the static consensus graphs (red, consensus proteins; yellow, steiner node proteins; lightblue, consensus transcription factors; green, consensus genes).

Figure S2 | Static consensus graphs for time points 1 h after EGF stimulation.

Figure S3 | Static consensus graphs for time points 4 h after EGF stimulation.

Figure S4 | Static consensus graphs for time points 8 h after EGF stimulation.

Figure S5 | Static consensus graphs for time points 13 h after EGF stimulation.

Figure S6 | Static consensus graphs for time points 18 h after EGF stimulation.

Figure S7 | Static consensus graphs for time points 24 h after EGF stimulation.

Figure S8 | Time course integration for phosphoproteins MAPK14 and PRKAR2B. Downstream consensus transcripts identified for MAPK14 and PRKAR2B were mapped to differentially abundant proteins. Note that the measurement range of the expression profiles across platforms can vary. Phosphoprotein time course data is shown in solid, black lines, non-matching transcript data in solid, gray lines and matching transcript and proteome data in rainbow color palette with proteins depicted as solid lines and transcripts depicted as dotted lines.

Table S1 | List of molecule cluster membership in the time profile analysis. Data origin is encoded in the abbreviation after each protein/gene name (_g, microarray data; _p, proteome data).

Table S2 | Lists of pathways identified in the downstream analysis based on the phosphoprotein data for time points 0.25, 1, 4, 8, 13, 18, and 24 h after EGF stimulation. Table includes information about the pathway database used for pathway identification (as part of their ID) and the corresponding pathway names.

Table S3 | Lists of pathways identified in the upstream analysis based on the differentially expressed transcripts for time points 1, 4, 8, 13, 18, and 24 h after EGF stimulation.

REFERENCES

Avraham, R., and Yarden, Y. (2011). Feedback regulation of EGFR signalling: decision making by early and delayed loops. *Nat. Rev. Mol. Cell Biol.* 12, 104–117. doi: 10.1038/nrm3048

Balbin, O. A., Prensner, J. R., Sahu, A., Yocum, A., Shankar, S., Malik, R., et al. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat. Commun.* 4:2617. doi: 10.1038/ncomms3617

Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* 13, 552–564. doi: 10.1038/nrg3244

Brigstock, D. R. (2002). Regulation of angiogenesis and endothelial cell function by connective tissue growth factor (CTGF) and cysteine-rich 61 (CYR61). *Angiogenesis* 5, 153–165. doi: 10.1023/A:1023823803510

Brockmann, R., Beyer, A., Heinisch, J. J., and Wilhelm, T. (2007). Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput. Biol.* 3:e57. doi: 10.1371/journal.pcbi.0030057

Cazzalini, O., Sommatis, S., Tillhon, M., Dutto, I., Bachi, A., Rapp, A., et al. (2014). CBP and p300 acetylate PCNA to link its degradation with nucleotide excision repair synthesis. *Nucleic Acids Res.* 42, 8433–8448. doi: 10.1093/nar/gku533

Chandrasekaran, L., He, C.-Z., Al-Barazi, H., Krutzsch, H. C., Iruela-Arispe, M. L., and Roberts, D. D. (2000). Cell contact–dependent activation of α3β1 integrin modulates endothelial cell responses to thrombospondin-1. *Mol. Biol. Cell.* 11, 2885–2900. doi: 10.1091/mbc.11.9.2885

Chen, I. T., Smith, M. L., O'Connor, P. M., and Fornace, A. J. (1995). Direct interaction of Gadd45 with PCNA and evidence for competitive interaction of Gadd45 and p21Waf1/Cip1 with PCNA. *Oncogene* 11, 1931–1937.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weise, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

Croucher, D. R., Saunders, D. N., Lobov, S., and Ranson, M. (2008). Revisiting the biological roles of PAI2 (SERPINB2) in cancer. *Nat. Rev. Cancer* 8, 535–545. doi: 10.1038/nrc2400

Cubbage, M. L., Suwanichkul, A., and Powell, D. R. (1990). Insulin-like growth factor binding protein-3. Organization of the human chromosomal gene and demonstration of promoter activity. *J. Biol. Chem.* 265, 12642–12649.

Dauer, D. J., Ferraro, B., Song, L., Yu, B., Mora, L., Buettner, R., et al. (2005). Stat3 regulates genes common to both wound healing and cancer. *Oncogene* 24, 3397–3408. doi: 10.1038/sj.onc.1208469

Ding, Y., Chen, M., Liu, Z., Ding, D., Ye, Y., Zhang, M., et al. (2012). atBioNet– an integrated network analysis tool for genomics and biomarker discovery. *BMC Genomics* 13:325. doi: 10.1186/1471-2164-13-325

Duffy, M. J. (2004). The urokinase plasminogen activator system: role in malignancy. *Curr. Pharm. Des.* 10, 39–49. doi: 10.2174/1381612043453559

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi: 10.1093/nar/gks1094

Hamon, J., Jennings, P., and Bois, F. Y. (2014). Systems biology modeling of omics data: effect of cyclosporine a on the Nrf2 pathway in human renal cells. *BMC Syst. Biol.* 8:76. doi: 10.1186/1752-0509-8-76

Han, J., Luby-Phelps, K., Das, B., Shu, X., Xia, Y., Mosteller, R. D., et al. (1998). Role of substrates and products of PI 3-kinase in regulating activation of rac-related guanosine triphosphatases by Vav. *Science* 279, 558–560. doi: 10.1126/science.279.5350.558

He, G., Siddik, Z. H., Huang, Z., Wang, R., Koomen, J., Kobayashi, R., et al. (2005). Induction of p21 by p53 following DNA damage inhibits both Cdk4 and Cdk2 activities. *Oncogene* 24, 2929–2943. doi: 10.1038/sj.onc.1208474

Herbst, R. S. (2004). Review of epidermal growth factor receptor biology. *Int. J. Radiat. Oncol. Biol. Phys.* 59, 21–26. doi: 10.1016/j.ijrobp.2003.11.041

Hollander, M. C., Alamo, I., Jackman, J., Wang, M. G., McBride, O. W., and Fornace, A. J. (1993). Analysis of the mammalian gadd45 gene and its response to DNA damage. *J. Biol. Chem.* 268, 24385–24393.

Hsu, T.-H., and Chang, T.-C. (2015). RARRES3 regulates signal transduction through post-translational protein modifications. *Mol. Cell. Oncol.* 2:e999512. doi: 10.1080/23723556.2014.999512

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Meth.* 12, 115–121. doi: 10.1038/nmeth.3252

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Kramer, F., Bayerlová, M., Klemm, F., Bleckmann, A., and Beissbarth, T. (2013). rBiopaxParser - an R package to parse, modify and visualize BioPAX data. *Bioinformatics* 29, 520–522. doi: 10.1093/bioinformatics/bts710

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A. L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299–313. doi: 10.1038/nrc3721

Kumar, L., and Futschik, M. E. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* 2, 5–7. doi: 10.6026/97320630002005

Lurje, G., and Lenz, H. J. (2009). EGFR signaling and drug discovery. *Oncology* 77, 400–410. doi: 10.1159/000279388

MacDonald, N. J., Freije, J. M. P., Stracke, M. L., Manrow, R. E., and Steeg, P. S. (1996). Site-directed Mutagenesis of nm23-H1 mutation of proline 96 or serine 120 abrogates its motility inhibitory activity upon transfection into human breast carcinoma cells. *J. Biol. Chem.* 271, 25107–25116. doi: 10.1074/jbc.271.41.25107

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108–D110. doi: 10.1093/nar/gkj143

Nishimura, D. (2001). BioCarta. *Biotechnol. Softw. Internet Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol.* 1, 2005.0010. doi: 10.1038/msb4100014

Park, O. K., Schaefer, T. S., and Nathans, D. (1996). *In vitro* activation of Stat3 by epidermal growth factor receptor kinase. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13704–13708. doi: 10.1073/pnas.93.24.13704

Purvis, J. E., and Lahav, G. (2013). Encoding and decoding cellular information through signaling dynamics. *Cell* 152, 945–956. doi: 10.1016/j.cell.2013.02.005

Rau, A., Jaffrézic, F., Foulley, J.-L., and Doerge, R. W. (2010). An empirical bayesian method for estimating biological networks from temporal microarray data. *Stat. Appl. Genet. Mol. Biol.* 9:9. doi: 10.2202/1544-6115.1513

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16, 85–97. doi: 10.1038/nrg3868

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., et al. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* 24, 2894–2900. doi: 10.1093/bioinformatics/btn553

Sadeghi, A., and Fröhlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. *BMC Bioinformatics* 14:144. doi: 10.1186/1471-2105-14-144

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, D674–D679. doi: 10.1093/nar/gkn653

Schroder, W. A., Major, L., and Suhrbier, A. (2011). The role of SerpinB2 in immunity. *Crit. Rev. Immunol.* 31, 15–30. doi: 10.1615/CritRevImmunol.v31.i1.20

Sun, H., Wang, H., Zhu, R., Tang, K., Gong, Q., Cui, J., et al. (2014). iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics* 30, 737–739. doi: 10.1093/bioinformatics/btt576

Treves, S., Feriotto, G., Moccagatta, L., Gambari, R., and Zorzato, F. (2000). Molecular cloning, expression, functional characterization, chromosomal localization, and gene structure of junctate, a novel integral calcium binding protein of Sarco(endo)plasmic reticulum membrane. *J. Biol. Chem.* 275, 39555–39568. doi: 10.1074/jbc.M005473200

Tullai, J. W., Schaffer, M. E., Mullenbrock, S., Sholder, G., Kasif, S., and Cooper, G. M. (2007). Immediate-early and delayed primary response genes are distinct in function and genomic architecture. *J. Biol. Chem.* 282, 23981–23995. doi: 10.1074/jbc.M702044200

Vogel, C., and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232. doi: 10.1038/nrg3185

Wachter, A., and Beissbarth, T. (2015). pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics* 31, 3072–3074. doi: 10.1093/bioinformatics/btv323

Wang, D., Xia, D., and Dubois, R. N. (2011). The crosstalk of PTGS2 and EGF signaling pathways in colorectal cancer. *Cancers* 3, 3894–3908. doi: 10.3390/cancers3043894

Waters, K. M., Liu, T., Quesenberry, R. D., Willse, A. R., Bandyopadhyay, S., Kathmann, L. E., et al. (2012). Network analysis of epidermal growth factor signaling using integrated genomic, proteomic and phosphorylation data. *PLoS ONE* 7:e34515. doi: 10.1371/journal.pone.0034515

Wysolmerski, J. J. (2012). Parathyroid hormone-related protein: an update. *J. Clin. Endocrinol. Metab.* 97, 2947–2956. doi: 10.1210/jc.2012-2142

Yanagida, O., Kanai, Y., Chairoungdua, A., Kim, D. K., Segawa, H., Nii, T., et al. (2001). Human L-type amino acid transporter 1 (LAT1): characterization of function and expression in tumor cell lines. *Biochim. Biophys. Acta* 1514, 291–302. doi: 10.1016/s0005-2736(01)00384-4

Yarden, Y., and Sliwkowski, M. X. (2001). Untangling the ErbB signalling network. *Nat. Rev. Mol. Cell Biol.* 2, 127–137. doi: 10.1038/35052073

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* 41, 316–323. doi: 10.1038/ng.337

Zhang, W., and Liu, H. T. (2002). MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res.* 12, 9–18. doi: 10.1038/sj.cr.7290105

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4 Integration of phosphoproteome and transcriptome data to link B cell receptor activation with gene expression dynamics

Reference

This manuscript is in preparation. Supplementary material of the manuscript is listed in Chapter 7 *Appendix*.

Original Contribution

AW, TO, HU, ES, CL and TB conceived and designed the study; AW, SS, JC, JB, SL and CL were involved in data preprocessing; SM and CD performed wet lab experiments; AW and SS performed bioinformatic analysis; AW and TB developed and implemented the software package; AW wrote the paper; TB and TO contributed to manuscript writing. All authors read the manuscript.

Integration of phosphoproteome and transcriptome data to link B cell receptor activation with gene expression dynamics

Astrid Wachter*[1], Thomas Oellerich[2,3,4], Jasmin Corso[5], Silvia Münch[2], Carmen Doebele[2], Salma Sohrabi[5], Julia Beck[6], Stephan Lorenzen[7], Christof Lenz[8,9], Ekkehard Schütz[6], Henning Urlaub[8,9] and Tim Beissbarth[1]

[1]Institute of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen

[2]Department of Medicine II, Hematology/Oncology, Goethe University, Theodor-Stern-Kai 7, 60590 Frankfurt, Germany

[3]German Cancer Research Center and German Cancer Consortium, Heidelberg, Germany

[4]Department of Haematology, Cambridge Institute of Medical Research, University of Cambridge, Hills Road, Cambridge, CB2 0XY, United Kingdom

[5]Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Am Fassberg 11, 37077 Göttingen, Germany

[6]Chronix Biomedical, Goetheallee 8, 37073 Göttingen, Germany and 5941 Optical Court Suite 203E, San Jose, CA 95138, USA

[7]Department of Molecular Parasitology, Bernhard Nocht Institute for Tropical Medicine, Bernhard-Nocht-Straße 74, D-20324, Hamburg, Germany

[8]Bioanalytics, University Medical Center, Institute for Clinical Chemistry, Robert-Koch-Straße 40, 37075 Göttingen, Germany

[9]Bioanalytical Mass Spectrometry Group, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

*Corresponding author

Running title: Link receptor activation with gene expression dynamics

Abstract

Integrating time-course information from different data types has emerged as essential element of systems biology approaches to improve our understanding of dynamic cellular responses. Driven by the growing interest in generation of matched high-throughput datasets, we developed a methodology to systematically identify individual signaling axes that are triggered by activated receptors and to link them to their transcriptional response. For this purpose we used a public time-resolved phosphoproteome dataset and generated time-shifted transcriptome data to analyze activated B cell receptor signaling dynamics. We integrated these datasets by a cellular layer-specific pathway-based approach, using public knowledge from biological databases. By construction of consensus graphs, reflecting layer-specific concordant information between the two data types, we were able to confirm known B cell receptor signaling links, e.g. PLCG2 activation following phosphorylation of the tyrosine kinase SYK. In addition, we found hitherto unknown relationships that we hypothesize to be implicated in BCR signaling. Furthermore, we were able to determine dynamically activated individual signaling axes by cross-platform analysis. Thus, integration of matched high-throughput datasets from different cellular layers is a promising approach to broaden our view on complex cellular signaling processes and thereby refine our mechanistic understanding of the cell.

Keywords

B cell receptor / lymphoma / data integration / signaling

Introduction

Functional integration of different molecular layers of the cell has gained more and more interest recently, in parallel with the increased generation of high-throughput datasets by different technical methods. These efforts have already revealed interesting findings in the context of systems biology and systems medicine (Ritchie et al., 2015; Du and Elemento, 2015; Wang et al., 2015). However, so far less emphasis has been put on the integration of time-series datasets from different technology platforms, although these datasets may enable a detailed cross-examination of effectors and effect propagation throughout different cellular layers (Bar-Joseph et al., 2012). Aberrant signal propagation is often linked to malignant transformation of cells, giving rise to major diseases. One such example are B cell malignancies, which are characterized by aberrantly activated B cell receptor (BCR) signaling. The B cell receptor is a control instance for B cell differentiation, homeostasis and function, as both maturation and survival of the cell are regulated by BCR signaling (Rickert, 2013).

Because of the pivotal role of BCR signaling for B cell physiology and pathophysiology, an improved understanding of BCR signaling axes is needed. Detailed investigation has been performed by transcriptional profiling of e.g. chronic lymphocytic leukemia patient samples (Ferreira et al., 2014). BCR engagement induces multiple signaling events by reversible phosphorylation of effector proteins that have been studied by phosphoproteomics (Satpathy et al., 2015; Corso et al., 2016). In addition signaling crosstalk and various feedback regulations have been described for BCR signaling pathways (Seda and Mraz, 2014; Song et al., 2013; Wang et al., 2012). However, all of these studies focus on one individual molecular layer of the complex signaling response, rather than integrating different layers.

Nonetheless, time-resolved layer-specific studies have demonstrated that signaling mechanisms cannot be fully captured when only single time points within a signaling response are monitored (Irish et al., 2006; Corso et al., 2016). Hence a detailed molecular characterization of signaling mechanisms acting not only in space but also in time requires time-series approaches. A time-resolved study demonstrating different temporal modes of RNA- and protein-layer regulations in response to misfolding stress in mammalian cells was published recently by Cheng et al. (2016). Following time-series data on different molecular layers thus clearly adds information with the potential to decipher molecular mechanisms in more detail (Wachter and Beißbarth, 2016).

To date open questions remain, for example, whether specific associations between dynamic transcriptional and post-translational processes can be determined that are intrinsic to a specific biological system. The scope of this study was (i) to systematically characterize the phosphoproteome and the transcriptome dynamics in BCR stimulated cells, (ii) to link the two processes over time, and (iii) to integrate them to improve our understanding of BCR signaling and its downstream consequences. Using a pathway-based layer-specific integration of phosphoproteome and transcriptome data we were able to identify time-dependent consensus molecule sets. Moreover, we could characterize those individual signaling axes in the cellular response that were supported by cross-platform analysis and gain further insights into signaling patterns by a correlation trajectory analysis.

Results

BCR signal propagation on phosphoproteome and transcriptome levels

In a previous study we systematically characterized BCR signaling dynamics by a stable-isotope labeling by amino acids in cell culture (SILAC) based phosphoproteome analysis (Corso et al., 2016). In this study phosphosite levels were measured in unstimulated Burkitt lymphoma cell lines and after 2, 5, 10 and 20 min of BCR stimulation (**Fig 1A**). In total, we detected 1024 phosphosites with significantly differential abundance in at least one of the analyzed stimulation durations, with approximately three times more upregulated than downregulated phosphosites for each stimulation duration (**Fig 1B**). To elucidate the corresponding downstream transcriptional activation in these cells we have now analyzed gene expression by RNA-sequencing in cells that were left unstimulated or were stimulated through their BCR for various stimulations (10, 20, 60 and 120 min) (**Fig 1A, Fig S1**-**3**). While we observed both up- and downregulated phosphosites on the phosphoproteome level, we almost exclusively found upregulated transcripts upon BCR stimulation (**Fig 1B**-**D, Tab S1**). Notably, early upregulation was observed for immediate early response genes, such as *FOS* or *EGR-1*.

Pathway-based integration of phosphoproteome and transcriptome data identifies key players of BCR signaling

We used the pathway-based data integration approach implemented in our R package 'pwOmics' for data integration of the different data sets, as it takes the different molecular layers that we analyzed by different technologies into consideration (Wachter and Beißbarth, 2015; https://bioconductor.org/packages/devel/bioc/html/pwOmics.html). Furthermore, we extended this approach by including signaling axes identification steps, which were incorporated into the 'pwOmics' package. The pathway-based integration compares the molecules identified in a 'downstream analysis' starting from the differentially phosphorylated phosphosites with the molecules identified in an 'upstream analysis' starting from differentially expressed transcripts in a layer specific manner. The different identification steps are performed using biological database knowledge, in particular pathway databases Biocarta (Nishimura, 2001), KEGG (Kanehisa et al., 2014), Pathway Interaction Database (PID) (Schaefer et al., 2009) and Reactome (Fabregat et al., 2016), as well as transcription factor (TF) target gene relations from the TRANSFAC® database (Matys et al., 2006). In addition, we incorporated information from the PhosphoSitePlus database about activatory and inhibitory downstream signaling of phosphoproteins to prefilter consensus molecules in the 'upstream' and the 'downstream analysis'. Moreover, we integrated a filtering step based on phosphorylation/dephosphorylation of phosphoproteins and up-/downregulation of transcript expression levels, respectively. **Fig 2** summarizes these new functions of the 'pwOmics' package.

In the 'downstream analysis' of the differentially phosphorylated phosphoproteins we first identified signaling pathways that were potentially affected according to public knowledge from the pathway databases mentioned above. As expected, these pathways contained the Biocarta 'BCR signaling pathway', the PID 'BCR signaling pathway' and the KEGG 'B cell receptor signaling pathway' for all analyzed stimulation time points (**Tab S2**). Accordingly, these PID and Biocarta pathways could be identified for all time points in the 'upstream analysis', confirming the pathway-level to be the adequate level for integration of these data sets. The lacking identification of the KEGG 'B cell receptor signaling pathway' in the 'upstream analysis' can be attributed to low numbers of significantly regulated transcripts, resulting from a conservative preprocessing of the dataset.

For all time points with available phosphoproteome and transcriptome data we performed a pathway-based integration to find functional links throughout the signaling axes. Via the 'downstream' and the

'upstream analysis' we arrived at potential molecules implicated in the different molecular layers, such that building the intersect on each layer reduced false positive identifications in a layer-specific manner. Thus sets of 'consensus molecules' represented by consensus proteins, consensus TFs and consensus genes/transcripts were identified for their role in BCR signaling at individual time points. **Fig 3** displays the corresponding time-shifted consensus graph, comprising the intersect of consensus proteins identified in intersection analysis pooling phosphoproteome data from 2, 5 and 10 min of stimulation and transcriptome data from 60 and 120 min of stimulation. This pooled integration allows to capture the overall signaling effect over time throughout the different molecular layers. Individual time-shifted consensus graphs for subsequent time points demonstrate that early signaling is governed to a greater extent by the phosphoproteome layer, whereas later time points are characterized by a more pronounced impact of the transcriptional layers. These time-resolved consensus graphs are depicted in **Fig EV1**, a consensus graph for phosphoproteome and transcriptome data that were obtained after the same BCR stimulation duration is shown in **Fig S4.** These consensus graphs exemplify that by our method the complex individual datasets could be condensed to interpretable heterogeneous graphs comprising integrated knowledge from both high-throughput technology platforms.

Next we investigated how much the integrated consensus graph reflects known regulatory signaling nodes. We found, that indeed knowledge about BCR signaling was collected so far either by layer-specific high-throughput data or by individual experiments. Such individual experiments link upstream protein signaling to downstream transcriptional changes, but are typically restricted to a certain signaling axis or pathway. In total we could map around 50 % of the identified consensus phosphoproteins and approximately 80 % of the identified consensus transcripts to previously published resources confirming the ability of our integration approach to identify relevant regulatory molecules. Phosphoproteins frequently associated with BCR signaling and identified by our integration approach included in particular CD19, SYK, MAPK1, MAPK3, BLNK, PLCG1, PLCG2 , CBL and PIK3CA.

While CD19 is a co-receptor of the BCR, SYK being recruited to the BCR leads to MAPK1 and MAPK3 activation. SYK can phosphorylate both PLCG1 and BLNK. The latter, upon phosphorylation, leads to the generation of docking sites that bind BTK and PLCG2, which are involved in initiation of

downstream NFAT activation via intracellular calcium levels. Another substrate of SYK, enhanced by Y317 phosphorylation, is the E3 ubiquitin ligase CBL, known as an inhibitor of SYK-dependent signaling. In parallel PIK3CA is triggered SYK-dependently, activating the AKT/mTOR pathway. (Young and Staudt, 2013; Geahlen, 2009)

As the described signaling regulations are highly dependent on the phosphorylation patterns, our next step was to dissect the individual contribution of consensus phosphoprotein sites on downstream signaling. Therefore, we analyzed their dynamic profiles. In **Fig 4** these phosphorylation profiles are linked to corresponding downstream consensus transcription factors and consensus transcripts. We observe a clear separation of four clusters in these sites, with SYK(Y525), SYK(Y526), SYK(Y348) and SYK(Y352) constituting the most upregulated cluster. SYK(Y525) and SYK(Y526) are known to be phosphorylated after BCR engagement in an autophosphorylation reaction, SYK(Y348) and SYK(Y352) are modified by autophosphorylation in vitro upon crosslinking of the BCR. All of these modifications induce enzymatic activity. (Geahlen, 2009)

Furthermore, we observe an intermediately upregulated cluster dominated by SYK which is represented by different phosphorylation patterns. CBL, APC, PIK3CA, CRKL, BLNK, NCK1, PLCG1, IQGAP1 and MAPK3 are exclusively found in this cluster, whereas CD22, RPS6KA3, TSC2, ABI1, MAPK7, RAF1, CD19, PTPN6 and LYN are exclusively found in the third, just slightly upregulated cluster. The latter includes a number of negative regulators such as CD22, an inhibitory co-receptor of the BCR. Its humanized anti-CD22 monoclonal antibody was previously tested in clinical trials in order to raise the threshold of BCR activation (Sieger et al., 2013). Also PTPN6 (SHP-1) negatively regulates signaling via the BCR (Ono et al., 1997), together with LYN in an inhibitory loop via CD22 and SHIP1 (Packard and Cambier, 2013). More interesting, though, are the SYK phosphosites upregulated in addition to the highly upregulated cluster of SYK phosphosites, as they might be possible therapeutic targets for pathologically activated BCR signaling. CBL is an inhibitor of SYK-dependent signaling by targeting SYK for ubiquitination (Geahlen, 2009; Sohn et al., 2003). Interestingly, all phosphosites of CBL that were at least identified to be significantly regulated at one time point show high expression values early on with subsequent decrease of the ratios. This pattern could also be observed in the highly up-phosphorylated cluster for SYK(Y525) and SYK(Y526) and in the intermediately up-phosphorylated cluster for the SYK phosphosites Y348, Y631, Y296, Y630,

Y525, S295, Y352, Y323 and Y296, which indicates a possible regulatory interaction. However, other SYK sites and specifically MAPK3 sites are showing an opposite pattern with increasing expression levels over time. While CBL, PIK3CA, CRKL, BLNK, NCK1, PLCG1 and MAPK3 are well known to be implicated in BCR signaling, the function of APC and IQGAP1 is less well described, but clearly identified via our pathway-based layer-specific integration approach.

In addition, we observe a downregulated phosphosite cluster, which consists of EPS15(S790), PAG1(Y181), SNIP1(S49, S52), ASAP1(S839) and SQSTM1(S266). While SNIP1, ASAP1 and SQSTM1 are only observed in the downregulated cluster, EPS15 and PAG1 are also found in the slightly upregulated cluster with different phosphorylation patterns. SNIP1 has been described as inhibitor of NFkB-signaling (Kim et al., 2001) and as part of a signature of pre-germinal center-derived B-Cell Non-Hodgkin Lymphomas (Rolland et al., 2014). In B cell receptor signaling activation of NFkB signaling can take place via BCL10/MALT1 and the recruitment of IKK (Ferch et al., 2007). SQSTM1, in contrast, has been described as ubiquitin-binding scaffold protein positively regulating NFkB signaling (Long et al., 2010). ASAP1 belongs to the Ras superfamily of small GTPases and is involved in cytoskeletal rearrangement (Büchse et al., 2011). Besides its known functions in intracellular trafficking and potential function in transcriptional regulation, EPS15 was described as regulator of B-cell lymphopoieses (van Bergen en Henegouwen, 2009; Pozzi et al., 2012). However, the individual phosphosites implicated here have not been described in regard to human B cell receptor signaling before to the best of our knowledge. Regardless, PAG1 phosphorylation on Y181 has been described in BL cell lines before (Rolland et al., 2014). Downregulation of this cluster of specifically modified proteins upon BCR stimulation can be interpreted as a propagation effect necessary for the stimulation signal to be transmitted throughout the cell.

Downstream of the phosphoprotein layer, however, we see a combinatorial response of transcriptional changes affected by propagation through the transcription factor layer. We observe early immediate response genes like *FOS*, *EGR1*, *EGR2* and *EGR3* highly upregulated early after BCR stimulation, as expected. Furthermore, we observe *PIM1* and *TXNIP* to be downregulated at late time points. *PIM1* is a protooncogene encoding a serine/threonine protein kinase (Zhu et al., 2002). Overexpression of *PIM1* in mice leads to tumor formation, inhibitors of *PIM1* have been shown to induce death of cancer cells

(Magnuson et al., 2010). *TXNIP* was described previously as antitumor gene as it forms a transcriptional repressor complex (Han et al., 2003). Interestingly, *CCL4* exhibits a positive regulation peak at 20 min of BCR stimulation. The expression of this chemokine is needed to attract T cells in the immune response (Takahashi et al., 2015). While the integrated consensus graph allows to draw a functional link between the upstream phosphoprotein activation and downstream transcriptional response, it does not provide information on individual signaling propagation axes so far.

Systematic characterization of signaling axes and feedback mechanisms

We were interested in following BCR dependent signaling from the activated receptors via protein phosphorylation to the transcriptome response. As typically layer-specific characterization of signaling is performed, we investigated if time-series datasets for different molecular layers enable a new and biologically more reasonable perspective of signal transduction. We therefore performed a systematic analysis of downstream signaling starting from the set of phosphosites being at least significantly regulated at one stimulation duration (**Fig 5A**). This resulted in a time-resolved overview on the number of target genes in activated pathways downstream of the corresponding phosphoproteins. Additionally, the number of matching transcripts to these target genes could be identified in the transcriptome data, individually for up- and downregulation. This is demonstrated in **Fig 5B** on the example of the phosphoprotein Epidermal Growth Factor Receptor Pathway Substrate 15 (EPS15) in detail. EPS15 is involved in receptor-mediated endocytosis of epidermal growth factor. We observe that according to pathway databases EPS15 is implicated in 13 signaling pathways, with seven pathways having a large number (>600) of target genes. Most of these are linked to Erbb1 signaling. As expected, we observe more matching transcripts at late transcriptome measurement time points and higher numbers of matching transcripts in pathways with high numbers of downstream target genes. Signaling axes downstream of the tyrosine kinase SYK are provided in **Fig S5.**

Correlation trajectories – identification of time-resolved correlation patterns

Next, we systematically analyzed time resolved correlation patterns for our set of consensus phosphoproteins. By comparing the same order of measurements (phosphoproteome 2 min vs. transcriptome 10 min, phosphoproteome 5 min vs. transcriptome 20 min, phosphoproteome 10 min vs. transcriptome 60 min and phosphoproteome 20 min vs. transcriptome 120 min), we assumed a delay in

signaling from the phosphorylation cascade to RNA synthesis of about 8 min, with a slower RNA synthesis compared to the phosphorylation cascade itself. These assumptions are based on measurements of human RNA synthesis rates that were previously measured to be 1.3-4.3 kb/min (Maiuri et al., 2011). Based on these numbers we checked the first transcriptome data time point (10 min) for the approximate maximal time durations needed to synthesize its significantly differentially expressed transcripts, in case transcript length information from UCSC Genome Browser (Kent et al., 2002) was available. Only a small number of transcripts from three genes (*DNAJB1*, *EGR2* and *NR4A1*) exceeded the threshold of 8 min (**Fig EV2**). While the additional time needed for further RNA processing steps is hard to estimate, we can still assume that 8 min after the first phosphoproteome data measurement we might capture almost all transcripts in the measurements.

Fig 6 exemplarily shows correlations of PAG1, PLCG2 and PTPN6 phosphoprotein expression levels with expression levels of some of their transcripts affected downstream. Only those transcripts were taken into consideration that were found to be differentially regulated in the transcriptome data set. Complete correlation results for these consensus phosphoproteins are provided in **Fig S6**. Such correlation trajectories give detailed insights into regulatory relationships from a response-specific point of view instead of a layer-specific one. Very similar correlation patterns can be observed for different phosphosites of one protein, e.g. for PAG1. However, also diverse correlation patterns of differently phosphorylated proteins such as for PLCG2 are identified. Similar correlation patterns might indicate same upstream regulators, whereas diverse phosphoprotein patterns hint to varying upstream regulatory influences of the investigated sites. While for the two sites depicted for PAG1 upstream regulation seems to be similar starting at 2 min of BCR stimulation, there is clearly a higher change in phosphorylation levels for PAG1(Y417) up to 2 min. Very similar patterns observed for a set of differently phosphorylated phosphoproteins on different downstream transcripts (e.g. for PTPN6 starting from 5 min phosphoproteome/ 20 min transcriptome) indicates very similar transcriptional regulation of these transcripts with upstream influence of similarly regulated phosphosites. In this case PTPN6 influence is negative and expression levels of the immediate early genes *EGR2* and *EGR3* are decreasing after early regulatory involvement.

The presented results are certainly biased towards information in the databases which were used for

identification of the signaling axes. Nevertheless, they provide a layer-integrated and summarized signal-specific view on the BCR signaling response and enable detailed investigation of individual signaling axes.

Discussion

Layer-specific high-throughput measurements have been the basis of experimental studies in past years, yet more and more emphasis is now given to a multi-omics characterization in order to investigate biological hypotheses with a systems focus. Furthermore, different studies show that time-series data sets are particularly important to understand complex cellular responses, as these responses can be composed of different temporal modes interacting in a time-dependent manner (Buescher et al., 2016). With the emergence of multi-omics time-series data sets, that are still rare, but expected to be generated increasingly (Bar-Joseph et al., 2012; Rajasundaram and Selbig, 2016), the need of appropriate analyses workflows arises. Using a parallely measured phosphoproteome and transcriptome data set of BCR stimulated human cells, we demonstrate that a response-specific signal propagation tracking enables a more focused characterization than a layer-specific one. We incorporated time-series data on the different cellular layers to track regulatory relationships in their particular signaling axes throughout different layers, thus arriving at a very detailed and systematic characterization of BCR stimulated downstream cellular adaptations.

Several studies have investigated paired links between phosphoproteome and transcriptome data sets before with slightly different foci. Oyama et al. (2011) linked SILAC-LC/MS time course data to GeneChip time course data through prediction of TF motif activity for understanding the molecular mechanisms of tamoxifen resistance at a system level in breast cancer. Rotival et al. (2015) generated LC-MS/MS and transcriptome data to identify regulators of macrophage multinucleation in the rat. They first characterized multinucleation-specific transcription factors by transcription factor binding site enrichment analysis, before mapping those together with the phosphopeptide data to a protein interaction network and identifying pairs in closer than random vicinity. These examples show various

applications, yet are alike in their aim to characterize different cellular response layers in one system. This shows that a flexible integration workflow as presented here has multiple potential applications, especially if time course data is considered. Our approach to systematically integrate phosphoproteome and transcriptome time series data sets will thus provide a useful option for many experimental applications in which such parallel data sets are measured to characterize cellular response in detail.

To facilitate the integrative analysis of comparable time-series high-throughput data sets we have presented here a pathway-based layer-specific integration approach with a focus on cellular response that covers more than one cellular layer. As part of the integration we deliberately accept a data reduction step, mapping phosphoproteome data to signaling pathways. However, individual activating or inhibitory downstream effects are considered, as part of individual signaling axes at specific stimulation durations. The potentiation of downstream effects due to a high number of signaling axes acting in concert results in a combinatorial transcriptional effect. With a comparison of 'downstream analysis' results to the transcriptome data set, we aim to filter out those phosphorylation changes that are not strong enough to result in a significant transcriptional change of target genes downstream, thereby controlling the number of false positives and setting a threshold for identification. With the incorporation of a filtering step based on regulatory concordance and further public knowledge from the PhosphoSitePlus database, we arrived at heterogeneous consensus graphs that are feasible for biological interpretation on the one hand, but also capture signaling contributions confirmed by both data types on the other hand. These graphs could highlight molecular players of the cellular response as confirmed by published resources over different cellular layers. Although the presented approach is inherently biased through the intense use of databases for linking purposes, resulting consensus graphs and signaling axes do benefit from combining various knowledge domains from different cellular layers. The downside of this public knowledge-based approach is that it will not predict nor emphasize newly discovered associations. Nevertheless, we were able to unfold the signaling axes affected by BCR signaling in a very detailed manner. This allowed us to investigate individual correlation trajectories of phosphosites and their affected transcripts downstream. These correlation patterns show that there are different classes of downstream effects for individual phosphorylation patterns, starting from almost same effects on different downstream transcripts through very similar transcriptional regulation effects of different transcripts towards very diverse phosphorylation patterns which also hint

to diverse functions in downstream signaling.

We conclude that a response-specific instead of a layer-specific investigation of signaling axes can lead to further insights into regulatory mechanisms. We strongly believe that integration of different data types is an indispensable step towards this response-specific perspective. Furthermore, we hope that the presented integrative approach can contribute to improve our understanding of regulatory mechanisms in cellular responses and thus help to identify required therapeutic interventions in deregulated signaling pathways.

Materials and Methods

Cell culture, BCR stimulation and cell lysis

The human DG75 lymphoma cell line was kindly provided by A. Rosenwald, Institute of Pathology, University of Wuerzburg, Wuerzburg, Germany. Cell culture, BCR stimulation and cell lysis was performed as described in Corso et al. (2016).

Phosphoproteome analysis & Mass spectrometry data analysis

Protein digestion for phosphoproteome analysis, phosphopeptide enrichment for pYome analysis, LC-MS/MS analysis and data processing was performed as described in Corso et al. (2016). Downstream data analysis of MaxQuant (Version 1.5.0.25) results was performed with Perseus (Version 1.5.0.15). Global phosphoproteome and pYome datasets of the SILAC-labelled cell line DG75 were analysed in independent sessions. Briefly, reverse and contaminant entries were removed, as were phospho-sites with a localization probability lower than 0.75. Sites were considered as quantified if at least 50% of biological replicates (global phosphoproteome 2/4 and pYome 1/2) had valid values. Ratios were logarithmized (log_2). For the global phosphoproteome and the pYome, sites with a SILAC ratio < -2 SD or >2 SD of at least one time point were considered as significantly regulated. Global phosphoproteome and pYome datasets were merged, filtering out pY measurements in global phosphoproteome that were measured in pYome and pS/pT measurements in pYome that were measured in global phosphoproteome. Raw files and MaxQuant search results have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE

partner repository (Vizcaíno et al., 2015) with the dataset identifier PXD003492.

RNA-Sequencing

Six biological replicates, each of DG75 cells either unstimulated or BCR-stimulated for 10 minutes, 20 minutes, 1 hour and 2 hours, were prepared. Additionally, for each time point unstimulated control cells were harvested and immediately frozen in liquid nitrogen. Pellets were thawed in RNAlater (Qiagen) and total RNA was extracted by using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions for extraction of total RNA from human cells. The RNA Integrity Number (RIN) was determined for all samples on an Agilent 2100 Bioanalyzer by using the Eukaryote Total RNA Nano Chip (Agilent). All RNA samples had an RIN of > 8. Sequencing libraries were prepared from the poly-An RNA fraction of 1 µg total RNA by using the TruSeq RNA Sample Preparation Kit according to the manufacturer's instructions (Illumina). Paired-end sequencing was performed on an Illumina HiSeq2000; 100 bp were generated for each read. A mean of 63 M (standard deviation 12 M) reads was generated for each of the 60 samples. Sequences were aligned to the RefSeq human transcriptome using bwa, and raw 'hits' per transcript were merged genewise (Li and Durbin, 2009; Pruitt et al., 2014). These counts per gene were analysed using DeSEQ (Anders and Huber, 2010) describing gene expression as a generalised linear model including treatment (BCR vs. CONTROL), a factorial time effect and the combined effect as full model; a model lacking the combined effect was used as a reduced model. The resulting p values (indicating a change over time depending on the treatment) were adjusted according to Benjamini-Hochberg (Benjamini and Hochberg, 1995). Raw data was deposited on NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE90120.

Data integration

Data integration of phosphoproteome and transcriptome data was done with R package 'pwOmics' (Wachter and Beißbarth, 2015). Functionality of the package was updated and extended throughout this analysis. Input phosphoproteome data was prefiltered to include only phosphoproteins with sites showing at least one significant regulation of at least one time point. For generation of the consensus molecule set pathway databases Biocarta (Nishimura, 2001), KEGG (Kanehisa et al., 2014), Pathway Interaction Database (Schaefer et al., 2009) and Reactome (Fabregat et al., 2016) as well as

TRANSFAC® database (Matys et al., 2006), version Biobase 2015.4, were used. Consensus graphs were generated based on protein-protein-interaction (PPI) database STRING (Franceschini et al., 2013) and TRANSFAC data, using the shortest path approximation of the Steiner tree algorithm (Sadeghi and Fröhlich, 2013). The utility of the pwOmics package was extended to compare not only differentially regulated molecules, but to include regulation concordance. Prior to analysis, PhosphoSitePlus (Hornbeck et al., 2014) database knowledge about downstream signaling activation or inhibition was included in the analysis: Regulatory sites were downloaded (03/2016) and prefiltered for human phosphorylation sites. Sites annotated as 'activity, induced' and 'activity, inhibited' were included filtering step of the analysis. In case no PhosphoSitePlus database information was available for a certain site, a direct comparison of downstream and upstream analyses was performed. Data integration analyses steps were done in R version 3.2.2.

Visualization

Heatmaps generated with the 'ComplexHeatmap' R package (Gu et al., 2016) show supervised hierarchical clustering using euclidean distance and complete linkage. Missing values of sites were imputed with the 'impute' R package (Hastie et al., 2016) based on 10 nearest neighbours prior to plotting. In consensus phosphoprotein heatmap four clusters were identified with k-means clustering. Network graphs were generated with the 'pwOmics' (Wachter and Beißbarth, 2015) package using 'igraph' (Csardi and Nepusz, 2006) and then visualized with Cytoscape (Cline et al., 2007) using communication R package 'RCy3' (Shannon et al., 2013) and Cytoscape App 'CyREST' (Ono et al., 2015). References used in **Fig 3** include Niiro and Clark, 2002; Pauls et al., 2016; Niiro et al., 2012; Su et al., J Biol Chem, 1999; Yin et al., 2007; Ingham et al., 1996; Castello et al., 2013; Goldfeld et al., 1992; Wen et al., 2003; Franke et al., 2011; Tabrizi et al., 2009; Dörner et al., 2015; Krzysiek et al., J Immunol, 1999.

Acknowledgements

The authors gratefully acknowledge financial support by the German Ministry of Education and Research (BMBF) for the e:Bio program grant MetastaSys (0316173A) and e:Med program grants HER2LOW (031A429C), MMML-Demonstrators (031A428B) and MyPathSem (FKZ031L0024A).

Author contributions

A.W., T.O., H.U., E.S., C.L. and T.B. conceived and designed the study; A.W., S.S., J.C., J.B., S.L. and C.L. were involved in data preprocessing; S.M. and C.D. performed wet lab experiments; A.W. and S.S. performed bioinformatic analyses; A.W. and T.B. developed and implemented the software package; A.W. wrote the paper; T.B. and T.O. contributed to manuscript writing. All authors approved the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Alexa, A., and Rahnenfuhrer, J. (2016). topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0.
2. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.
3. Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet *13*, 552–564.
4. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289–300.
5. van Bergen En Henegouwen, P.M. (2009). Eps15: a multifunctional adaptor protein regulating intracellular trafficking. Cell Commun. Signal *7*, 24.
6. Büchse, T., Horras, N., Lenfert, E., Krystal, G., Körbel, S., Schümann, M., Krause, E., Mikkat, S., and Tiedge, M. (2011). CIN85 interacting proteins in B cells-specific role for SHIP-1. Mol. Cell Proteomics *10*, M110.006239.
7. Buescher, J.M., and Driggers, E.M. (2016). Integration of omics: more than the sum of its parts. Cancer & Metabolism *4*, 4.
8. Castello, A., Gaya, M., Tucholski, J., Oellerich, T., Lu, K.-H., Tafuri, A., Pawson, T., Wienands, J., Engelke, M., and Batista, F.D. (2013). Nck-mediated recruitment of BCAP to the BCR regulates the PI(3)K-Akt pathway in B cells. Nat. Immunol. *14*, 966–975.
9. Cheng, Z., Teo, G., Krueger, S., Rock, T.M., Koh, H.W., Choi, H., and Vogel, C. (2016).

Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. Mol. Syst. Biol. *12*, 855.

10. Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat Protoc *2*, 2366–2382.
11. Corso, J., Pan, K.-T., Walter, R., Doebele, C., Mohr, S., Bohnenberger, H., Ströbel, P., Lenz, C., Slabicki, M., Hüllein, J., et al. (2016). Elucidation of tonic and activated B-cell receptor signaling in Burkitt's lymphoma provides insights into regulation of cell survival. Proc. Natl. Acad. Sci. U.S.A. *113*, 5688–5693.
12. Csardi, G., Nepusz, T. (2006). The igraph software package for complex network research, InterJournal, Complex Systems 1695. http://igraph.org
13. Dörner, T., Shock, A., Goldenberg, D.M., and Lipsky, P.E. (2015). The mechanistic impact of CD22 engagement with epratuzumab on B cell function: Implications for the treatment of systemic lupus erythematosus. Autoimmun Rev *14*, 1079–1086.
14. Du, W., and Elemento, O. (2015). Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene *34*, 3215–3225.
15. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucl. Acids Res. *30*, 207–210.
16. Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome pathway Knowledgebase. Nucleic Acids Res. *44*, D481–D487.
17. Ferch, U., zum Büschenfelde, C.M., Gewies, A., Wegener, E., Rauser, S., Peschel, C., Krappmann, D., and Ruland, J. (2007). MALT1 directs B cell receptor-induced canonical nuclear factor-kappaB signaling selectively to the c-Rel subunit. Nat. Immunol. *8*, 984–991.
18. Ferreira, P.G., Jares, P., Rico, D., Gómez-López, G., Martínez-Trillos, A., Villamor, N., Ecker, S., González-Pérez, A., Knowles, D.G., Monlong, J., et al. (2014). Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. Genome Res *24*, 212–226.
19. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. *41*, D808–D815.
20. Franke, A., Niederfellner, G.J., Klein, C., and Burtscher, H. (2011). Antibodies against CD20 or B-Cell Receptor Induce Similar Transcription Patterns in Human Lymphoma Cell Lines. PLOS

ONE *6*, e16596.

21. Geahlen, R.L. (2009). Syk and pTyr'd: Signaling through the B cell antigen receptor. Biochim. Biophys. Acta *1793*, 1115–1127.
22. Goldfeld, A.E., Flemington, E.K., Boussiotis, V.A., Theodos, C.M., Titus, R.G., Strominger, J.L., and Speck, S.H. (1992). Transcription of the tumor necrosis factor alpha gene is rapidly induced by anti-immunoglobulin and blocked by cyclosporin A and FK506 in human B cells. Proc Natl Acad Sci U S A *89*, 12198–12201.
23. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics btw313.
24. Han, S.H., Jeon, J.H., Ju, H.R., Jung, U., Kim, K.Y., Yoo, H.S., Lee, Y.H., Song, K.S., Hwang, H.M., Na, Y.S., et al. (2003). VDUP1 upregulated by TGF-beta1 and 1,25-dihydorxyvitamin D3 inhibits tumor cell growth by blocking cell-cycle progression. Oncogene *22*, 4035–4046.
25. Hastie, T., Tibshirani, R., Narasimhan, B. and Chu G. (2016). impute: impute: Imputation for microarray data. R package version 1.44.0.
26. Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520.
27. Ingham, R.J., Krebs, D.L., Barbazuk, S.M., Turck, C.W., Hirai, H., Matsuda, M., and Gold, M.R. (1996). B cell antigen receptor signaling induces the formation of complexes containing the Crk adapter proteins. J. Biol. Chem. *271*, 32306–32314.
28. Irish, J.M., Czerwinski, D.K., Nolan, G.P., and Levy, R. (2006). Altered B-cell receptor signaling kinetics distinguish human follicular lymphoma B cells from tumor-infiltrating nonmalignant B cells. Blood *108*, 3135–3142.
29. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. *42*, D199–D205.
30. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res. *12*, 996–1006.
31. Kim, R.H., Flanders, K.C., Reffey, S.B., Anderson, L.A., Duckett, C.S., Perkins, N.D., and Roberts, A.B. (2001). SNIP1 Inhibits NF-κB Signaling by Competing for Its Binding to the C/H1 Domain of CBP/p300 Transcriptional Co-activators. J. Biol. Chem. *276*, 46297–46304.
32. Krzysiek, R., Lefèvre, E.A., Zou, W., Foussat, A., Bernard, J., Portier, A., Galanaud, P., and Richard, Y. (1999). Antigen Receptor Engagement Selectively Induces Macrophage

548 Inflammatory Protein-1α (MIP-1α) and MIP-1β Chemokine Production in Human B Cells. J
549 Immunol *162*, 4455–4463.
550 33. Kumar, L., and E. Futschik, M. (2007). Mfuzz: A software package for soft clustering of
551 microarray data. Bioinformation *2*, 5–7.
552 34. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
553 transform. Bioinformatics *25*, 1754–1760.
554 35. Liu, Y., and Aebersold, R. (2016). The interdependence of transcript and protein abundance:
555 new data–new complexities. Mol Syst Biol *12*.
556 36. Long, J., Garner, T.P., Pandya, M.J., Craven, C.J., Chen, P., Shaw, B., Williamson, M.P.,
557 Layfield, R., and Searle, M.S. (2010). Dimerisation of the UBA Domain of p62 Inhibits
558 Ubiquitin Binding and Regulates NF-κB Signalling. Journal of Molecular Biology *396*, 178–
559 194.
560 37. Magnuson, N.S., Wang, Z., Ding, G., and Reeves, R. (2010). Why target PIM1 for cancer
561 diagnosis and treatment? Future Oncol *6*, 1461–1478.
562 38. Maiuri, P., Knezevich, A., De Marco, A., Mazza, D., Kula, A., McNally, J.G., and Marcello, A.
563 (2011). Fast transcription rates of RNA polymerase II in human cells. EMBO Rep. *12*, 1280–
564 1285.
565 39. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I.,
566 Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module
567 TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108–
568 D110.
569 40. Niiro, H., and Clark, E.A. (2002). Regulation of B-cell fate by antigen-receptor signals. Nat Rev
570 Immunol *2*, 945–956.
571 41. Niiro, H., Jabbarzadeh-Tabrizi, S., Kikushige, Y., Shima, T., Noda, K., Ota, S.-I., Tsuzuki, H.,
572 Inoue, Y., Arinobu, Y., Iwasaki, H., et al. (2012). CIN85 is required for Cbl-mediated regulation
573 of antigen receptor signaling in human B cells. Blood *119*, 2263–2273.
574 42. Nishimura, D. (2001) BioCarta. Biotech Software & Internet Report. 2, 117–120.
575 43. Ono, K., Muetze, T., Kolishovski, G., Shannon, P., and Demchak, B. (2015). CyREST:
576 Turbocharging Cytoscape Access for External Tools via a RESTful API. F1000Research.
577 44. Ono, M., Okada, H., Bolland, S., Yanagi, S., Kurosaki, T., and Ravetch, J.V. (1997). Deletion of
578 SHIP or SHP-1 reveals two distinct pathways for inhibitory signaling. Cell *90*, 293–301.
579 45. Oyama, M., Nagashima, T., Suzuki, T., Kozuka-Hata, H., Yumoto, N., Shiraishi, Y., Ikeda, K.,
580 Kuroki, Y., Gotoh, N., Ishida, T., et al. (2011). Integrated Quantitative Analysis of the

Phosphoproteome and Transcriptome in Tamoxifen-resistant Breast Cancer. J. Biol. Chem. *286*, 818–829.

46. Packard, T.A., and Cambier, J.C. (2013). B lymphocyte antigen receptor signaling: initiation, amplification, and regulation. F1000Prime Rep *5*.
47. Pauls, S.D., Ray, A., Hou, S., Vaughan, A.T., Cragg, M.S., and Marshall, A.J. (2016). FcγRIIB-Independent Mechanisms Controlling Membrane Localization of the Inhibitory Phosphatase SHIP in Human B Cells. J. Immunol. *197*, 1587–1596.
48. Pozzi, B., Amodio, S., Lucano, C., Sciullo, A., Ronzoni, S., Castelletti, D., Adler, T., Treise, I., Betsholtz, I.H., Rathkolb, B., et al. (2012). The Endocytic Adaptor Eps15 Controls Marginal Zone B Cell Numbers. PLOS ONE *7*, e50818.
49. Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res *42*, D756–D763.
50. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.
51. Rajasundaram, D., and Selbig, J. (2016). More effort — more results: recent advances in integrative “omics” data analysis. Current Opinion in Plant Biology *30*, 57–61.
52. Rickert, R.C. (2013). New insights into pre-BCR and BCR signalling with relevance to B cell malignancies. Nat Rev Immunol *13*, 578–591.
53. Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet *16*, 85–97.
54. Rolland, D., Basrur, V., Conlon, K., Wolfe, T., Fermin, D., Nesvizhskii, A.I., Lim, M.S., and Elenitoba-Johnson, K.S.J. (2014). Global Phosphoproteomic Profiling Reveals Distinct Signatures in B-Cell Non-Hodgkin Lymphomas. Am J Pathol *184*, 1331–1342.
55. Rotival, M., Ko, J.-H., Srivastava, P.K., Kerloc’h, A., Montoya, A., Mauro, C., Faull, P., Cutillas, P.R., Petretto, E., and Behmoaras, J. (2015). Integrating Phosphoproteome and Transcriptome Reveals New Determinants of Macrophage Multinucleation. Mol Cell Proteomics *14*, 484–498.
56. Sadeghi, A., and Fröhlich, H. (2013). Steiner tree methods for optimal sub-network identification: an empirical study. BMC Bioinformatics *14*, 144.
57. Satpathy, S., Wagner, S.A., Beli, P., Gupta, R., Kristiansen, T.A., Malinova, D., Francavilla, C., Tolar, P., Bishop, G.A., Hostager, B.S., et al. (2015). Systems-wide analysis of BCR

signalosomes and downstream phosphorylation and ubiquitylation. Mol Syst Biol *11*.

58. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic Acids Res. *37*, D674–D679.
59. Seda, V., and Mraz, M. (2015). B-cell receptor signalling and its crosstalk with other pathways in normal and malignant cells. Eur J Haematol *94*, 193–205.
60. Shannon, P.T., Grimes, M., Kutlu, B., Bot, J.J., and Galas, D.J. (2013). RCytoscape: tools for exploratory network analysis. BMC Bioinformatics *14*, 217.
61. Sieger, N., Fleischer, S.J., Mei, H.E., Reiter, K., Shock, A., Burmester, G.R., Daridon, C., and Dörner, T. (2013). CD22 ligation inhibits downstream B cell receptor signaling and Ca(2+) flux upon activation. Arthritis Rheum. *65*, 770–779.
62. Sohn, H.W., Gu, H., and Pierce, S.K. (2003). Cbl-b negatively regulates B cell antigen receptor signaling in mature B cells through ubiquitination of the tyrosine kinase Syk. J. Exp. Med. *197*, 1511–1524.
63. Song, W., Liu, C., Seeley-Fallen, M.K., Miller, H., Ketchum, C., and Upadhyaya, A. (2013). Actin-mediated feedback loops in B-cell receptor signaling. Immunol. Rev. *256*, 177–189.
64. Su, L., Rickert, R.C., and David, M. (1999). Rapid STAT Phosphorylation via the B Cell Receptor. J Biol Chem *274*, 31770–31774.
65. Tabrizi, S.J., Niiro, H., Masui, M., Yoshimoto, G., Iino, T., Kikushige, Y., Wakasaki, T., Baba, E., Shimoda, S., Miyamoto, T., et al. (2009). T cell leukemia/lymphoma 1 and galectin-1 regulate survival/cell death pathways in human naive and IgM+ memory B cells through altering balances in Bcl-2 family proteins. J. Immunol. *182*, 1490–1499.
66. Takahashi, K., Sivina, M., Hoellenriegel, J., Oki, Y., Hagemeister, F.B., Fayad, L., Romaguera, J.E., Fowler, N., Fanale, M.A., Kwak, L.W., et al. (2015). CCL3 and CCL4 are biomarkers for B cell receptor pathway activation and prognostic serum markers in diffuse large B cell lymphoma. Br J Haematol *171*, 726–735.
67. Vizcaíno, J.A., Csordas, A., del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., et al. (2015). 2016 update of the PRIDE database and its related tools. Nucl. Acids Res. gkv1145.
68. Wachter, A., and Beißbarth, T. (2015). pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. Bioinformatics *31*, 3072–3074.
69. Wachter, A., and Beißbarth, T. (2016). Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data. Front Genet *6*.

70. Wang, R.-S., Maron, B.A., and Loscalzo, J. (2015). Systems medicine: evolution of systems biology from bench to bedside. WIREs Syst Biol Med *7*, 141–161.
71. Wang, X., Li, J.-P., Kuo, H.-K., Chiu, L.-L., Dement, G.A., Lan, J.-L., Chen, D.-Y., Yang, C.-Y., Hu, H., and Tan, T.-H. (2012). Down-regulation of B Cell Receptor Signaling by Hematopoietic Progenitor Kinase 1 (HPK1)-mediated Phosphorylation and Ubiquitination of Activated B Cell Linker Protein (BLNK). J. Biol. Chem. *287*, 11037–11048.
72. Wen, R., Chen, Y., Xue, L., Schuman, J., Yang, S., Morris, S.W., and Wang, D. (2003). Phospholipase Cγ2 provides survival signals via Bcl2 and A1 in different subpopulations of B cells. The Journal of Biological Chemistry *278*, 43654–43662.
73. Yin, Q., Wang, X., McBride, J., Fewell, C., and Flemington, E. (2008). B-cell Receptor Activation Induces BIC/miR-155 Expression through a Conserved AP-1 Element. J. Biol. Chem. *283*, 2654–2662.
74. Young, R.M., and Staudt, L.M. (2013). Targeting pathological B cell receptor signalling in lymphoid malignancies. Nat Rev Drug Discov *12*, 229–243.
75. Zhu, N., Ramirez, L.M., Lee, R.L., Magnuson, N.S., Bishop, G.A., and Gold, M.R. (2002). CD40 signaling in B cells regulates the expression of the Pim-1 kinase via the NF-kappa B pathway. J. Immunol. *168*, 744–754.

Figure legends

Fig 1: Paired phosphoproteome and transcriptome data sets show characteristic expression changes after BCR stimulation. (A) BCR stimulation durations of phosphoproteome and transcriptome measurements. Time scale is log-transformed. **(B)** Number of significantly regulated sites/transcripts at corresponding BCR stimulation durations. Bars above zero-level indicate upregulation numbers, bars below zero-level downregulation numbers. Phosphoproteome data is shown individually for serine phosphorylated sites (pS), threonine phosphorylated sites (pT) and tyrosine phosphorylated sites (pY). Regulated transcripts are abbreviated as T. **(C)** Heatmap displaying log2 ratios of phosphosites being at least differentially phosphorylated at one stimulation duration compared

to no stimulation. **(D)** Heatmap displaying fold changes of transcripts being at least significantly regulated at one time point compared to no stimulation.

Fig 2: R software tool 'pwOmics' enables pathway-based and layer-specific integration of phosphoproteome and transcriptome data. (A) In the 'downstream analysis' preprocessed phosphoproteome data is used to identify signaling pathways of differentially phosphorylated sites. These pathways are scanned for transcription factors. In a next step downstream target genes are identified. In the 'upstream analysis' differentially expressed transcripts are used to identify upstream transcription factors. Signaling pathways containing these transcription factors are then evaluated in regard to potential proteomic regulators. 'Downstream' and 'upstream analyses' are performed for each stimulation duration. Intersecting molecules can be defined as a consensus molecule set C on each of the three molecular layers – protein layer depicted in red, transcription factor layer depicted in blue and transcript/gene layer depicted in green. **(B)** The extension of our R software tool 'pwOmics' provides a more sophisticated approach to define consensus molecule sets: Both direction of regulation and phosphorylation information from the PhosphoSitePlus database about enzymatic downstream activity are used to define the consensus sets for the protein, the transcription factor and the transcript/gene layer. If database information for an individual phosphosite is available it is used to prefilter the consensus sets taking into consideration concordance of regulation, otherwise no prefiltering step is performed. P refers to protein layer, TF refers to transcription factor layer, T refers to transcript/gene layer, ↓ refers to downregulation, ↑ refers to upregulation.

Fig 3: Integrated omics data based consensus graph can be confirmed to a large extent by literature. Time-shifted consensus graph based on the intersect of consensus proteins identified in intersection analysis pooling phosphoproteome data from 2, 5 and 10 min of stimulation and transcriptome data from 60 and 120 min of stimulation. Graph displays consensus proteins in red oval shapes, transcription factors in blue hexagons and genes/transcripts in green rectangles. Protein-protein dependencies are shown in solid lines, whereas transcription factor target gene relations are represented in dashed lines. Molecules are framed in colors according to literature references of studies of B cell receptor signaling. Multiple frames were used for multiple references. Reference annotation is not exhaustive, but clearly shows that studies so far were mainly either individual experiments that

investigated individual proteins or signaling pathways, or high-throughput based and focusing on one molecular layer.

Fig 4: Highly regulated phosphorylation patterns identified in integrative analysis can be attributed mostly to SYK. Heatmaps show temporal changes of phosphosite log2 ratios and fold changes of those consensus proteins and consensus transcripts, respectively, which are part of the consensus graph in Figure 4. Phosphosites of proteins are annotated with S (serine), T (threonine) or Y (tyrosine) phosphorylation, together with the annotation of multiple phosphorylation events (M). Missing values are displayed in 'grey' inside the heatmap. The color bar encodes different phosphosites of one protein with the same color, proteins with just one phosphosite differentially phosphorylated at least on one stimulation duration are shown in 'grey'. Connecting lines between TFs and transcripts show regulatory relationships as depicted in Figure 3.

Fig 5: Signaling axes triggered by BCR stimulation can be identified for individual phosphoproteins. (A) Downstream signaling of individual phosphoproteins affects target gene expression via signal propagation through pathways. To reduce false positive identifications, target genes are cross-checked against transcriptome data by comparison with differentially regulated transcripts. Each phosphoprotein can affect multiple pathways at each BCR stimulation duration, further affecting different sets of target genes. Comparison with transcriptome data can be performed in the same temporal order of measurements (as indicated here) or with larger time shifts. Feedback signaling will take place changing protein levels corresponding to transcripts that were affected initially by BCR stimulation, as indicated by blue, dashed arrows. This has a further impact on different signaling pathways at later time points. **(B)** Upper panel: Downstream signaling of phosphoprotein EPS15. Number of potentially affected target genes identified for each pathway found in signaling axes downstream of EPS15. Lower panel: Affected target genes matching differentially regulated transcripts. Left plot shows number of matching upregulated transcripts, right plot shows number of matching downregulated transcripts when comparing affected target genes of the indicated pathways with transcriptome data. Colors indicate BCR stimulation durations with 'red' corresponding to 10 min, 'green' corresponding to 20 min, 'cyan' corresponding to 60 min and 'purple' corresponding to 120 min of BCR stimulation.

Fig 6: Correlation trajectories of phosphosites affecting downstream transcripts, exemplarily for phosphoproteins PAG1, PLCG2 and PTPN6. Each protein specific panel shows selected transcripts. In each plot log2 ratios of indicated phosphosites are plotted against fold changes of affected transcripts. BCR stimulation times of phosphoproteome and transcriptome data are indicated in the legend, with filled circles indicating 2 min of BCR stimulation in phosphoproteome and 10 min of BCR stimulation in transcriptome data, filled diamonds indicating 5 min and 20 min of BCR stimulation in phosphoproteome and transcriptome data, respectively, filled triangles indicating 10 min and 60 min of BCR stimulation in phosphoproteome and transcriptome data, respectively, and filled squares indicating 20 min of BCR stimulation in phosphoproteome and 120 min of BCR stimulation in transcriptome data. Short distances between symbols indicate small phosphorylation and transcript level changes, whereas large horizontal and vertical distances between symbols indicate large phosphorylation and transcript level changes, respectively. Incomplete trajectories are shown in case of missing values (e.g. PTPN6_Y525_M2).

Expanded View Figure Legends

Fig EV1: Individual time-shifted consensus graphs for subsequent time points of phosphoproteome and transcriptome measurements. Graph displays consensus proteins in red oval shapes, transcription factors in blue hexagons and genes/transcripts in green rectangles. Protein-protein dependencies are shown in solid lines, whereas transcription factor target gene relations are represented in dashed lines. Phosphoproteome/transcriptome stimulation durations are indicated for each graph.

Fig EV2: Significantly regulated transcripts (10 min after stimulation) and maximal time durations needed for RNA synthesis. UCSC Genome Browser was used to assess the transcript lengths. Maximal synthesis duration of Maiuri et al. (2011) was used to calculate maximal synthesis durations.

Appendix material

Fig S1: Principal component analysis of RNA-Seq data set.

Fig S2: Sample heatmap of RNA-Seq data set.

Fig S3: Venn diagram showing overlap of significantly regulated transcripts for different stimulation times.

Fig S4: Consensus graph based on same measurement time points of phosphoproteome and transcriptome data.

Fig S5: Signaling axes downstream of SYK.

Fig S6: Exemplary correlation trajectories of PAG1, PLCG2 and PTPN6 signaling axes.

Tab S1: Significantly regulated transcripts. Transcriptome data analysis revealed 35 transcripts differentially regulated after 10 min of BCR stimulation, 59 transcripts differentially regulated after 20 min of BCR stimulation, 65 transcripts differentially regulated after 60 min of BCR stimulation and 87 transcripts differentially regulated after 120 min of BCR stimulation.

Tab S2: BCR signaling related 'upstream' and 'downstream pathways' identified in integrative analysis.

Figure 1

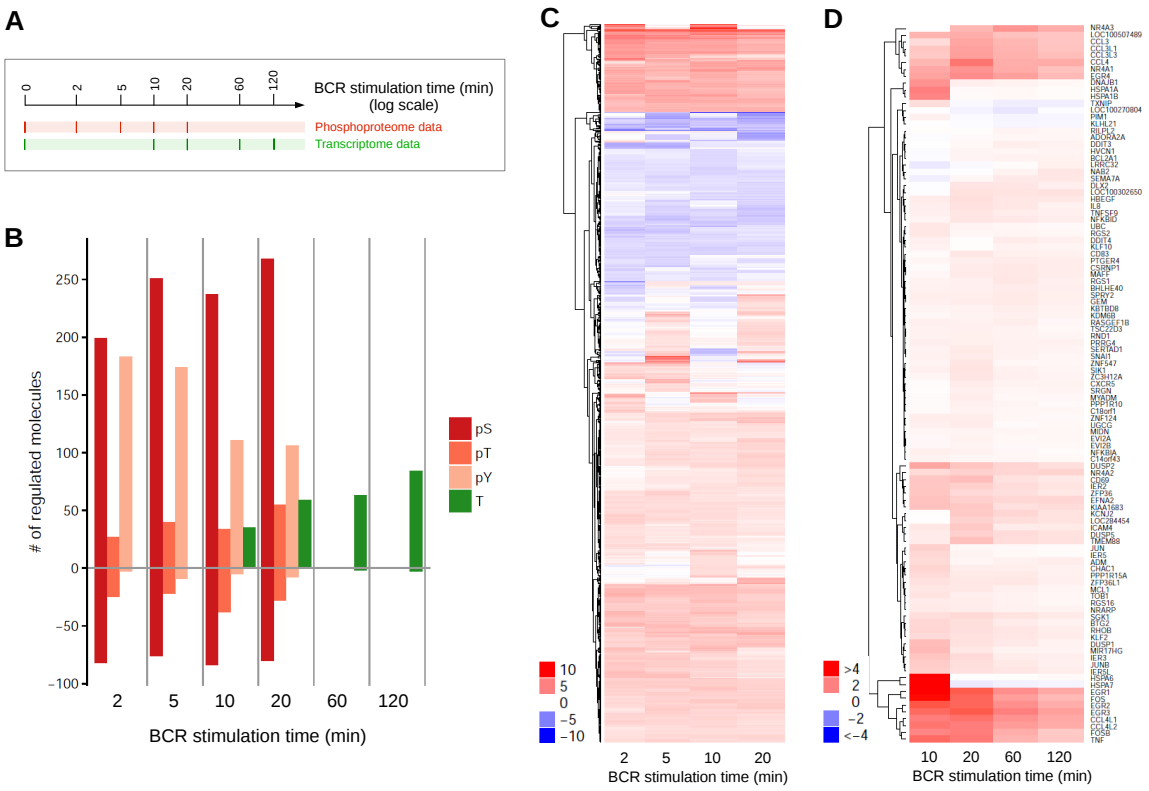


Figure 2

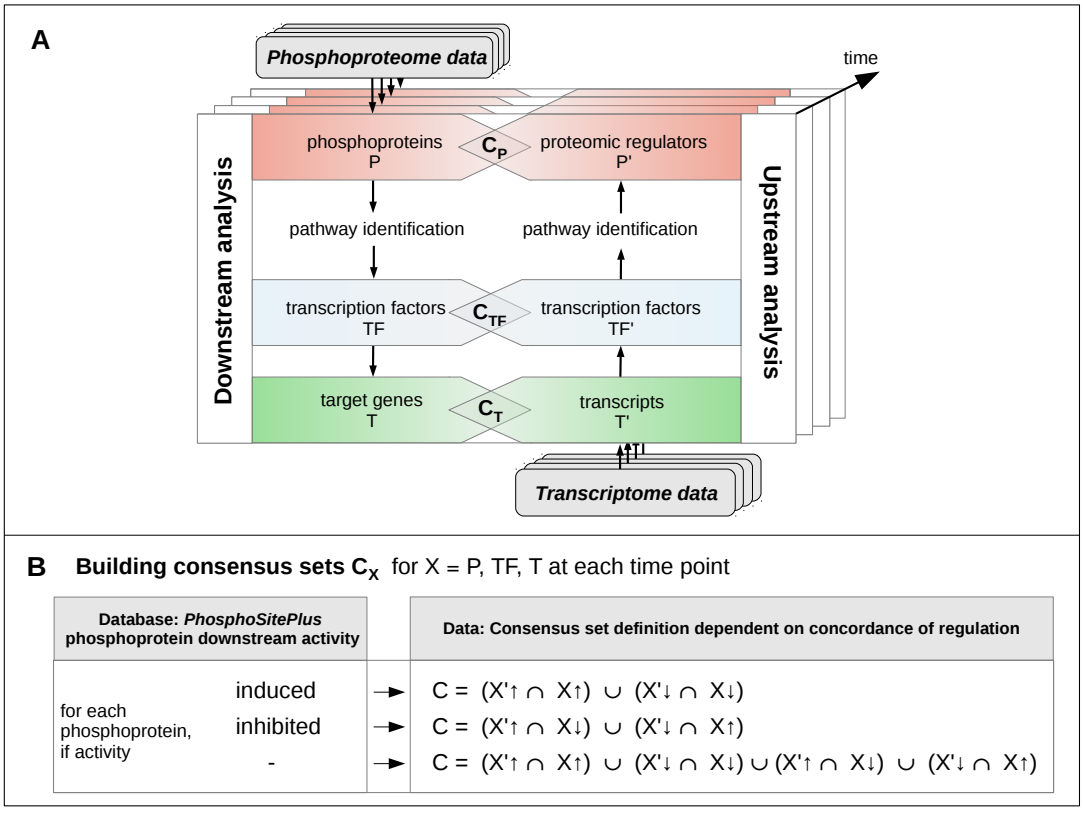


Figure 3

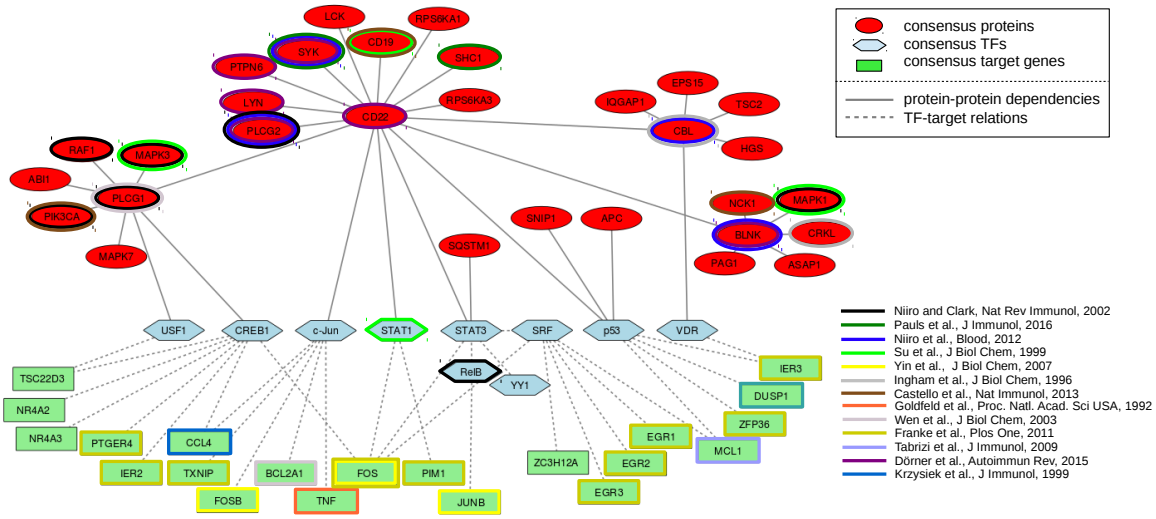


Figure 4

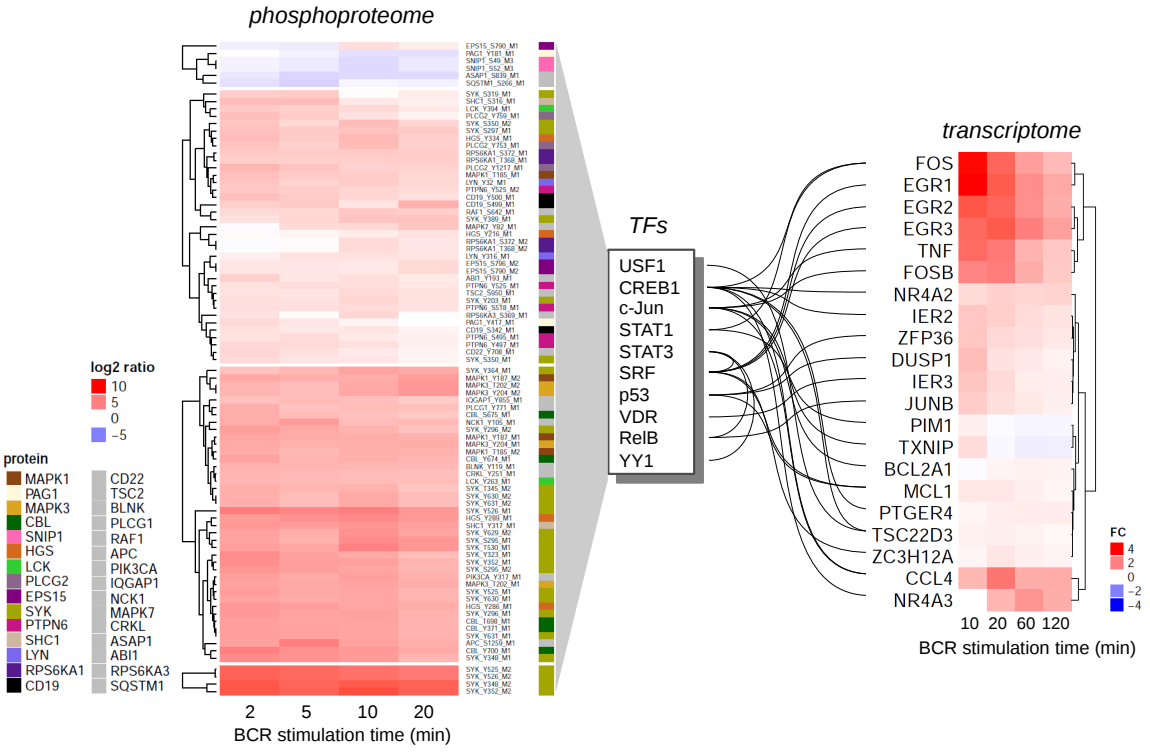


Figure 5

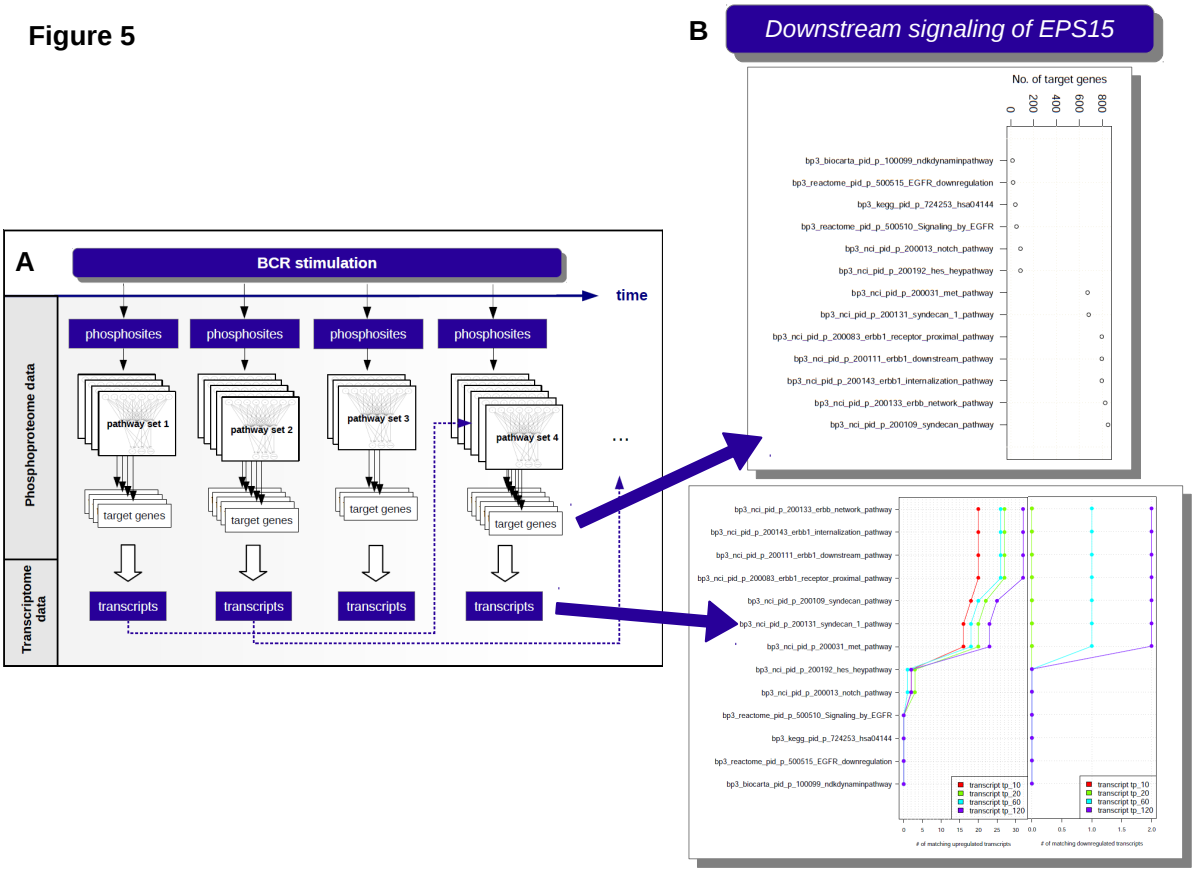


Figure 6

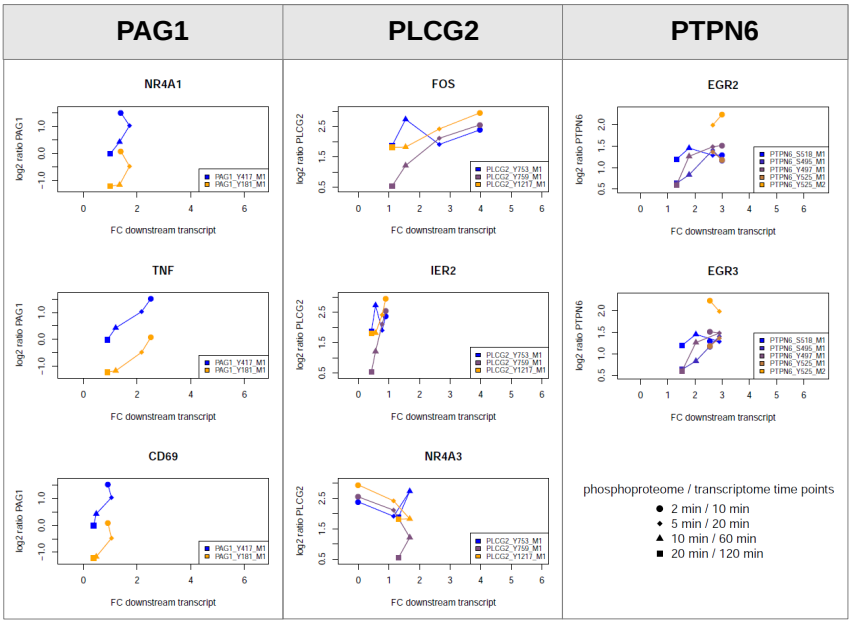
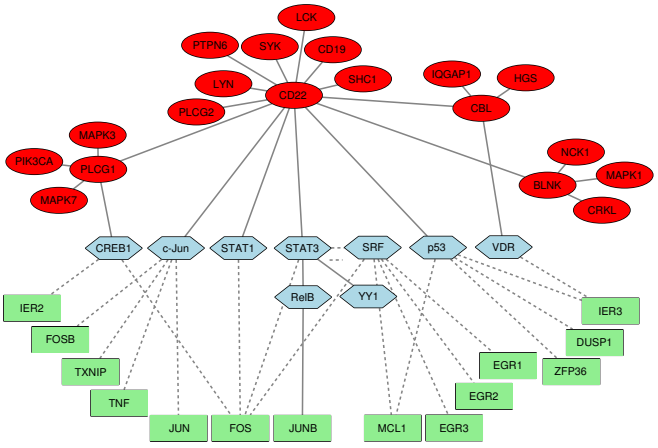
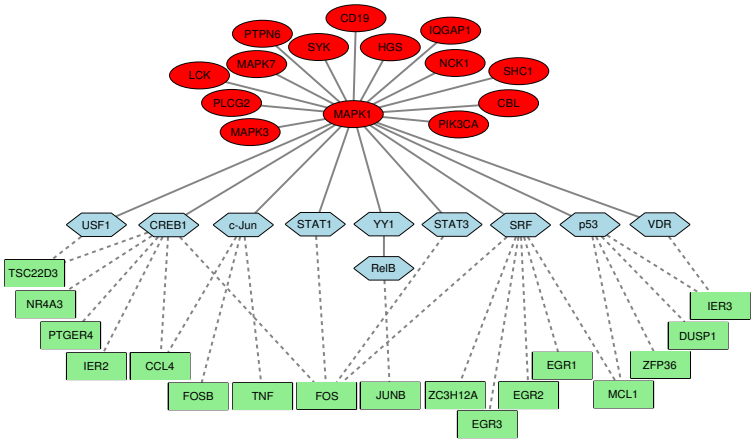
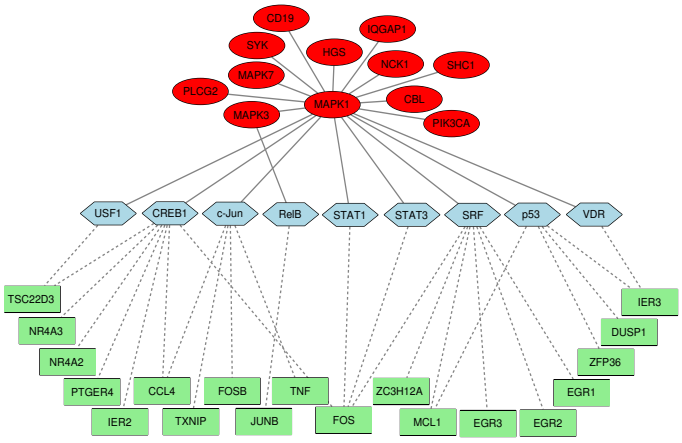
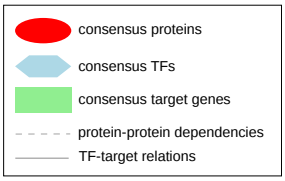


Figure EV1







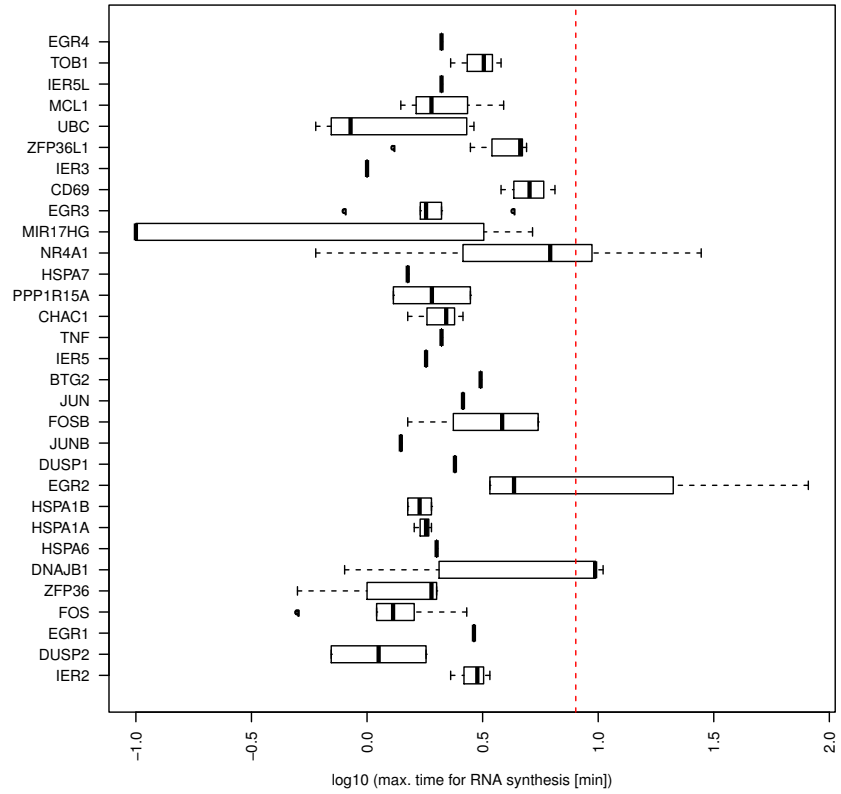


5 min / 10 min

10 min / 20 min

20 min / 60 min

Figure EV2


EGR4
TOB1
IER5L
MCL1
UBC
ZFP36L1
IER3
CD69
EGR3
MIR17HG
NR4A1
HSPA7
PPP1R15A
CHAC1
TNF
IER5
BTG2
JUN
FOSB
JUNB
DUSP1
EGR2
HSPA1B
HSPA1A
HSPA6
DNAJB1
ZFP36
FOS
EGR1
DUSP2
IER2
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
log10 (max. time for RNA synthesis [min])

5 Discussion

In this work, I adressed a very specific integration problem, composed of linking different but specific data types in a biologically meaningful way. These data sets containing (phospho-)proteomic and transcriptomic data measured in parallel are still very rarely generated, however, with the increasing use of different high-throughput platforms to characterize a single biological experiment, such data sets gain in importance and require appropriate analysis workflows (Gomez-Cabrero et al., 2014).

In the previous chapters, I introduced a pathway-based level-specific data integration method for (phospho-)proteomic and transcriptomic data. This method takes into consideration the molecular levels on which the data is generated and finds common regulatory influences between the different molecular levels. Signaling analysis based on the integrated data enables a comprehensive analysis of co-regulation patterns, consensus networks and inferred causal links between consensus molecules. This is complemented by the response-specific identification of signaling axes. In Chapter 2 *pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge* the implementation of this method as R package 'pwOmics' is presented. To demonstrate its functions on parallely measured phosphoproteome, transcriptome and proteome data, two data sets were investigated. They were generated with different high-throughput methods and analyzed with a focus on different biological questions. Thus a rough comparison of the data integration approach applied on different input data is possible. Chapter 3 *Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data* demonstrates the proposed approach on the example of physiological EGFR signaling. It reveals feedback signaling events and pathway crosstalk that has been described previously, but also identifies new regulatory interactions that might give rise to generation of new biological hypothesis. Chapter 4 *Integration of phosphoproteome and transcriptome data to link B cell receptor activation with gene expression dynamics* gives a systematic overview on time-resolved signaling events throughout different molecular levels and thereby identifies response-specific signaling axes instead of layer-specific signaling axes.

The results of the presented integration approach reveal that the linkage of the different data domains via biological databases and the subsequent integration allows identifying basic molecular components. Regulatory interactions, which are of biological importance for the cellular response, can be captured whereas it is not possible to cover different molecular levels just based on individual data sets. However, the use of databases gives a strict and

biased filtering towards those components that have been characterized already, which is symptomatic for a database-guided approach.

Here, results have been confirmed as far as possible based on already available knowledge on the different signaling pathways in the specific experimental settings. Furthermore, the investigated signaling pathways are already well characterized on different molecular levels individually (Oda et al., 2005; Reddy et al., 2016; Ferreira et al., 2014; Satpathy et al., 2015). This permits to evaluate results independently, however, further experimental validation will be necessary. Often, integration approaches are validated by checking certain individual processes that are testable in individual experiments. This validation approach has the advantage that individual regulatory mechanisms can be cross-checked, yet, it omits a considerable number of regulatory mechanisms in case they are not directly testable in the laboratory. Further validation might also include further regulatory linkage between data sets over time as signaling is expected to be reflected on each molecular level.

The R package 'pwOmics' developed during the course of this work is available on *Bioconductor*, an open source software framework for the development of tools for the analysis and comprehension of genomic high-throughput data.

5.1 Deciphering level-exceeding molecular mechanisms

The biological rationale of coupling different data types as presented in Section 1.3.1 *Underlying biological rationale* stays a prevailing need in the light of diseases that could not be characterized even with very sensitive high-throughput methods applied on one or even several molecular levels individually (Haider and Pal, 2013). In order to understand cellular responses it is therefore crucial to not only compare the different data sets, but to integrate them in a biologically meaningful way.

This includes a very thorough consideration of both the links between the different levels as well as the finely regulated temporal sequences of cellular signaling. Surely, this consideration needs to go together with careful drafting of the scope of the applied approach. So far, a lot of very individual proteomic and transcriptomic data integration approaches have been published with different hypotheses and scopes (Section 1.3.3, *Integration approaches*), which shows that there is a high need and a high potential for such approaches. When exclusively regarding the scope of understanding molecular mechanisms and relationships between and within different types of molecular structures, many approaches have been developed, yet most of them are not implemented in a way that allows public use. Still, in the context of increasing focus on systems-wide characterization, many researchers generate data sets on more than one molecular level and would benefit considerably by having access to corresponding analysis tools and pipelines (Gomez-Cabrero et al., 2014).

In this work, I chose the integration to be performed on the level of signaling pathways, implying a link of the different data types via pathway knowledge. This includes a data reduction step in terms of downstream effects of phosphoproteins, yet it channelizes the

regulatory influence between the different molecular levels and thus links the effectors with the affected molecules. The downstream and upstream analyses of the phosphoproteome data set and the transcriptome data set were chosen to be primarily performed individually. This has the advantage that intermediate results can be retrieved, such as the set of signaling pathways, in which there are differentially abundant phosphoproteins or the set of signaling pathways which is involved upstream of the differentially expressed transcripts. Furthermore, it enables level-specific data integration based on the intersect of molecules on each level, thereby reducing false positive identifications introduced in the database knowledge extraction step. According to Table 1.1 (Section 1.3.3, *Integration approaches*) the basic idea of this integration approach falls into the category of topological network approaches with an identification of common regulators.

The presented approach relies on a sound preprocessing of the individual data sets, such that in case different significance thresholds are chosen for the individual data sets also a more or less conservative result can be generated. This gives high flexibility to investigate more and less conservative settings in the integration tool 'pwOmics'.

The choice of the integration level and the methodical success of linking the effectors to the affected molecules are further discussed in Section 5.1.1 *Pathway-based integration: Linking effects and effectors*. The data sets investigated with the presented approach illustrate different analysis foci, dependent on either the number of parallely measured data points or the high-throughput methods used. This is discussed in Section 5.1.2 *Data set characteristics and potential*.

In addition, further optimization would be possible by the use of disease specific instead of general pathway models. One example for cancer-specific pathway models has been published by (Kuperstein et al., 2015), however, generation of disease-specific models needs a very careful data curation. Although at the moment consideration of such models in the presented integration approach is still very limited, future optimization is possible, as there is an increasing number of data sources providing disease-specific information (Wu et al., 2010; Mizuno et al., 2012).

5.1.1 Pathway-based integration: Linking effects and effectors

Signaling pathways have developed throughout evolution and have been adjusted in different species according to the particular environmental challenges. They channel information flow, cross-link between signaling axes and provide the interface to metabolic changes, ion fluxes and all changes a cell has to undergo in case of external stimulation. Likewise, they keep up basic cellular functions (Jordan et al., 2000). As pathways are the systems in which both initial cellular reactions take place and late-response changes are initiated forming the basis for an enduring cellular adjustment, they build the biological layer on which these effects can be measured easily with high-throughput methods today. These measurements also reflect positive or negative combinatorial feedback influence that enables an adjustment of the cellular signaling system over time. Yet, when analyzing these data, also input and

output of information have to be interlinked to understand the individual pathway effects in more detail.

Multi-layered data sets enable a step-by-step tracking of the information flow. With the different layers also mechanistic dependencies are predefined, which are usually flexibly dependent on the cellular environment. Given the complex signaling interdependencies, this tracking can be performed in a more refined manner the higher the temporal resolution of measurements is.

The consequence of not considering signaling pathways individually in data integration but using just a union or intersect of molecule sets, is the erroneous assumption of perfect correlation between protein and RNA expression (see Section 1.3.1, *Underlying biological rationale*) and a neglection of cellular dynamics. In addition, significance thresholds can either lead to very small subsets of identified molecules, which cannot be interpreted easily, or to very large numbers of molecules constituting a challenge for biological interpretation, as well. Nevertheless, there is still a big challenge associated with the use of pathway knowledge as integration basis, which is the bias inherent in pathway databases. This is further discussed in Section 5.3.2, *Database biases and restrictions*.

Furthermore, the 'classical' signaling pathways commonly do not describe the three-dimensional signaling space. Hence, if not associated with certain structures such as e.g. membrane proteins, the localization of molecules is disregarded. This means a simplification in regard to concentration gradients and molecular transport times takes place. Such detailed information can nowadays be used to set up detailed bottom-up mathematical models, however, it requires data sets to be highly resolved, both in time and in space. An example for the importance of this interdependency is colocalization of signaling molecules, as stochastic effects have been described to have an impact on signaling speed (Josić et al., 2011). Both detailed spatial information and stochastic influences are not taken into account when integrating different high-throughput data sets with pathway models. Instead measurements are considered as results of signaling in cellular space. Consequently, the idea presented in this work is inference of cellular mechanisms activated with stimulation rather than providing a model which can reflect cellular processes completely.

5.1.2 Data set characteristics and potential

While the queried biological databases have a high influence on the presented integration approach, the data set characteristics provide different potential for it, dependent on the number of parallely measured data points and the high-throughput methods used. While the EGFR signaling data set comprises few, but carefully selected phosphoproteins, the BCR signaling data includes a large phosphoproteome data set that is not preselected by researchers, but confined by the measurement method only. In addition, the measurement methods deviate strongly: The EGFR signaling data set comprises PowerBlot^{TM} measurements on the phosphoprotein and microarray measurements on the transcriptome level, while phosphosites in the BCR signaling data set have been measured with mass spectrometry and transcriptome data with RNA-Seq. As a result, the first setting allows a much more focused view on the

integrated data, but at the same time it encounters the problem of having a biased selection of phosphoproteins. The second setting of BCR signaling avoids this problem and gives a more unbiased view on the integrated results.

A higher number of parallely measured time points in the EGFR signaling data set enables retrieval of time-course information by inference of a probabilistic network, while the shifted time points measured on the different molecular layers in the BCR signaling data set allow for a detailed tracking of signal propagation over time through the different layers. Furthermore, the more sensitive methods used in the BCR signaling data set provide a more detailed insight into molecular mechanisms. The EGFR signaling data set reflects physiological signaling conditions, while in the BCR stimulation setting, the cellular response of a BL cell line is measured. For these cells, it is not clear yet, in how far activated BCR signaling is affected by the pathological state of the cell.

In both cases a comparison of the analysis results with a comparable pathological respectively physiological setting would improve understanding of molecular changes caused by the disease. Larger data sets consisting of more regulatory layers, e.g. miRNA expression or epigenetics, would have further potential to broaden the knowledge of cellular responses from a systems point of view.

5.2 Data integration findings: From known regulatory patterns towards newly identified cellular response characteristics

The data integration findings presented in Chapter 3 *Decoding Cellular Dynamics in Epidermal Growth Factor Signaling Using a New Pathway-Based Integration Approach for Proteomics and Transcriptomics Data* and Chapter 4 *Integration of phosphoproteome and transcriptome data to link B cell receptor activation with gene expression dynamics* exemplify the prospects of systems biology in terms of understanding complex biological systems. Both EGF signaling and BCR signaling are very well characterized due to their high impact on different human malignancies. This made it possible to address two different foci in the presented studies.

1. Network inference just based on data already pre-filtered according to the consensus molecules identified in the integratory analysis, enabling hypothesis formation on signaling crosstalk and time-dependent signal propagation.

2. Systematic identification of signaling axes in a response-specific way instead of level-specific way, allowing to track effect propagation throughout different molecular levels and to then correlate upstream signaling with downstream transcriptional effects.

5.2.1 Decoding cellular dynamics in epidermal growth factor signaling

Data integration for the EGF signaling data set enabled to derive specific dynamic patterns. Findings of the integrated analysis could confirm known regulatory response patterns of the cell upon an external stimulation. These are characterized by a primary and by a secondary cellular response (Tullai et al., 2007), reflecting early adaptation processes and long-lasting cellular changes. Also in the analysis of the co-regulation patterns via time profile clustering main signaling patterns could be characterized, including early immediate signaling (both up- and downregulation) and a late cellular response.

Static consensus profiles, reflecting the molecules' membership in the static consensus graphs at each time point of measurement, hint towards auto-feedback signaling. These observation could be done e.g. for PLAU, urokinase-type plasminogen activator, and CTGF, connective tissue growth factor. Auto-feedback of CTGF could be confirmed in the dynamic consensus analysis, where CTGF could be mapped to early, but sustained gene expression changes. Interestingly, CTGF is part of the early negatively regulated cluster, so that its downregulation might contribute to enhanced proliferation of cells upon EGF stimulation.

As time-series data allow for the identification of interactions, pathway crosstalk was one of the main foci of data integration in this data set. Crosstalk not characterized by posttranslational modification, such as the activation of matrix metalloproteinases (MMPs) by G-protein-coupled receptors (Yarden and Sliwkowski, 2001) could be identified very clearly at late time points after stimulation (4-24 hrs after EGF stimulation). This included MMP1, MMP2, MMP10 and ErbB ligands AREG and EGF, the latter one triggering time-shifted self-induction. However, the results are constrained by the small amount of phosphoprotein data available in this dataset.

Nevertheless, inference of a probabilistic network enabled identification of regulatory effects between the different molecular layers. A subsequent mapping to the time domains of the consensus molecules revealed the main trajectory of the system, thus making it possible to follow signaling changes. Known regulatory patterns were identified such as the indispensable task of MAPK1 and STAT3 to activate downstream transcriptional changes. As STAT3 is also a transcription factor, which builds homo- and heterodimers upon phosphorylation, translocates to the nucleus and activates transcription (Park et al., 1996), it also triggers a cellular feedback response. This feedback can go through JAK2, a target gene of STAT3, activating positive auto-feedback to STAT3 (Dauer et al., 2005).

Interestingly, SERPINB2 was identified as a consensus molecule which is differentially abundant in the protein dataset. SERPINB2 is known to inhibit urokinase plasminogen activators (PLAUs), which I hypothesized to be feedback-regulated based on the static consensus profiles. As both identification of SERPINB2 and PLAU is based on the integrated measurements of EGF stimulation effects on different platforms, relevance in the proliferation processes can be assumed. This hypothesis is e.g. supported by a study showing SERPINB2 to by associated with increased survival in breast cancer patients (Duffy, 2004).

The success of integrating gene expression dynamics and protein abundance dynamics is heavily influenced by translational and post-translational processes as well as molecular degradation processes. Due to the wide dynamic range of such processes they are not always observable in the data. Results indeed do not very clearly show time-shifted correlations in the dynamics, as expected. This might be attributed to the low time resolution of measurements, to post-translational modifications of the proteins, to a rapid degradation of mRNA or of protein products.

The presented study on EGF signaling thus revealed new hypotheses in regard to regulatory patterns and demonstrated the gain of an integratory data analysis over purely level-specific analysis.

5.2.2 *Systematic data integration of DG75 B cell receptor stimulation - phosphoproteome and transcriptome data in concert*

Integration of B cell receptor signaling data from different molecular levels enabled identification of signaling axes in a response-specific instead of a level-specific way. Pathway-based consensus analysis of phosphoproteome and transcriptome data could confirm known key players in B cell receptor signaling, but also identify so far unknown signaling links. In this study, time-dependent signaling patterns as well as transcriptional changes were identified and linked by identification of signaling axes.

With a systematic tracking of signaling events from the side of the receptor via phosphorylation cascades in responding pathways to their downstream effects on transcription I could clearly characterize different dynamic phases in the cellular response in the integrated consensus molecule sets. These could be separated in a first very active phase on the phosphoproteome level, that leads to increasing transcriptional changes during further time points. To generate a more general view on the cellular response phosphoproteome data from early time points after stimulation and transcriptome data from late time points after stimulation were pooled in the integration process. The resulting pooled consensus view on the cellular response covers a high number of molecules known to be implicated in BCR signaling, as shown in a comparison with literature ressources. While the presented cross-platform integration approach enabled level-specific comparison of the molecular levels and thus covers consensus molecules derived from different measurement platforms, previous knowledge was generated either in high-throughput on just one molecular level or in experiments just covering very specific signaling axes.

The consensus set of molecules implicated in BCR signaling could then be checked further in regard to phosphorylation dynamics. Besides phosphosites known to be implicated in BCR signaling and highly upregulated in our data, such as SYK(Y525) and SYK(Y526) which are autophosphorylated after BCR engagement, we could observe a high number of additionally upregulated SYK phosphosites. These are very interesting in regard to therapeutic intervention in BCR signaling.

Apart from well-known BCR signaling members, we also identified APC and IQGAP1 as consensus proteins. These belong to the phosphoproteins upregulated after BCR stimulation, however, the sites identified in this study have not been listed in the PhosphoSitePlus database (Hornbeck et al., 2015) yet. APC is a tumor suppressor known as negative WNT signaling pathway regulator and its activity is correlated with its phosphorylation status. IQGAP1 is involved in the regulation of cell morphology and motility. This finding is an example which constitutes an interesting result for designing further experimental studies. Apart from upregulated phosphosites, we also identified a number of downregulated sites, including PAG1(Y181), previously described in BL cell lines (Rolland et al., 2014).

A further benefit of tracking individual signaling propagation is the possibility to observe dynamic changes in the phosphoproteome data as a function of its transcripts being affected downstream. Such correlation trajectories can give insights into upstream phosphorylation of phosphoproteins in the investigated cellular response and into the downstream regulation in regard to transcription. However, cautious interpretation of these correlations is necessary as the transcriptional pattern is essentially a combinatorial effect of upstream regulation.

The response-specific integrated information compiled by the presented data integration method introduces a database bias on analysis results, but it nevertheless forms a step towards a systems characterization. This systems view can be refined further by including a higher number of regulatory molecular levels. In summary, these results help to identify regulation patterns in the complex chain of effectors and effects constituting BCR signaling.

5.3 Limitations of the presented cross-platform integration approach

The presented cross-platform integration approach is facing a number of methodological as well as computational challenges. As linking of different molecular levels is performed purely based on database knowledge, the results clearly depend on the selection of databases. While database curation and currentness of data can vary strongly across different databases, a research bias towards those molecules being investigated more intensely is observed. Thus, the results can only be as reliable as the input database knowledge used and will therefore not reflect interactions between molecules discovered only recently. Furthermore, mapping between the different databases and molecular IDs proves to be an additional challenge, for which a generalized standard would greatly improve the efforts of systems biology and systems medicine approaches.

As input data for the presented approach is the pre-processed data from individual platforms, the integrative analysis is very flexible on the one hand, while it is dependent on the selected significance thresholds of the input data on the other hand.

Biological simplification of protein activity is done in the course of data reduction. Therefore it is crucial to ensure it is complemented by bottom-up knowledge collected in individual experiments. Here, this knowledge was introduced by including PhosphoSitePlus

database (Hornbeck et al., 2015) knowledge when information regarding protein activation was available. This allowed to prefilter consensus molecule sets accordingly.

Methodologically and due to financial considerations it is not feasible to generate data sets with a very high dynamic resolution of measurements. Currently, this prevents applying classical time-series methods and makes it necessary to use an estimation procedure to generate dense data points for the dynamic consensus analysis. In this way dynamic changes are used to retrieve more information, however, only low frequency changes can be captured.

Individual limitations regarding genomic data integration, database biases, time resolution of measurements and data standardization are described in more detail in the following subsections.

5.3.1 Limits of genomic data integration

While prospects of genomic data integration are very promising, there are also multiple limitations and caveats. More evidence supporting a certain molecular link should increase confidence in a prediction as multiple features from different sources increase knowledge coverage. However, the increase of predictive power is usually limited and depends on whether the features are well selected and independent (Lu et al., 2005). Furthermore, for integratory methods it proves difficult to assess the statistical power of the approach universally, so that an interpretation or evaluation of results has to be performed with caution. Data reduction steps might neglect relevant functional associations in favor of non-relevant associations. In addition, the extraction of primary variables often used during data reduction might be challenging in terms of arriving at interpretable models (Ritchie et al., 2015).

The presented approach performs a filtering step to identify features that are of relevance based on two different data sets, thus uses increased knowledge coverage. As variables are not derived by factor analysis, but directly selected in the integratory filter process, interpretation of primary variables is not a problematic issue in this case. However, data from multiple molecular layers intrinsically has dependent correlation structures. Nevertheless, incorporation of time-series data and dynamic bayesian network inference was able to identify of causal influences as indicated by literature comparison.

At the moment, a further restriction of genomic data integration is the ongoing debate whether it is preferable to generate knowledge in width covering rather more molecular layers or in depth generating data with more sensitive techniques, but on fewer molecular layers. Analysis methods need to be customized according to these decisions. With improving methods and decreasing costs for data generation, this issue will loose its importance in future. At the moment, it is still reasonable to challenge whether additional data from a different molecular layer would add to the overall understanding of underlying molecular mechanisms.

5.3.2 Database biases and restrictions

Besides the obvious research bias that is found in databases due to more intensely investigated hub genes, database quality or a lack thereof needs to be taken into account when setting up models and assessing analysis results. Especially when using this biological knowledge for linking effects and effectors, it needs to be taken into consideration that most databases include mixed findings of different experiments, conducted in different tissues, cell lines etc. Thus, an estimate of the extent of actual knowledge that can be transferred undoubtedly on the experiment of interest is challenging, given the complex structure of cellular signaling and molecular links making up a cellular response. Additionally, there might be biases during the construction of the database itself. Such biases have been studied e.g. for the case of the miRBase (Griffiths-Jones et al., 2006). Analysis of historic versions of this database as subsets of today's or the final database revealed a strong dependence of the network topology on the point of time at which the data was retrieved (Saturnino et al., 2014). Such a bias is often neglected when interpreting results based on database knowledge.

Furthermore, the current offer of signaling databases and further biological databases is structurally divided into modular parts, e.g. individually reflecting signaling pathways from receptor layer to transcription factors, individually reflecting transcription factor binding, individually reflecting protein-protein interactions. This can be attributed to the complexity of modeling universal signaling in the cell, but also leads to i) problems in combining knowledge from different databases and ii) the tendency to judge one of these modules as an independent unit. As reflected in the idea of the rather holistic view of systems biology, this clearly is of questionable value from a biological point of view. Therefore, there is a parallel trend in building up integratory databases, based on e.g. data centralization, data warehousing, dataset integration, or direct links between the data (Lapatas et al., 2015). These raise awareness on the challenge of standards adoptions and common file formats, and additionally foster solutions that are taking into account the different structures of different data types. These efforts might lead to a rather response-specific instead of a unit- or module-specific investigation of data, which would reflect the true molecular interplay between different molecular layers more appropriately.

As the presented integration approach is restricted by only allowing to retrieve e.g. pathway information provided in a BioPAX format (a standardized pathway exchange format), the number of pathway databases which can be scanned for biological information is limited. With further integration of other databases either data integration methods or database output formats need to be very flexible in order to successfully combine information.

5.3.3 Time resolution effects on network inference

When considering time-series data, an important question to address is the sampling rate. If sampling during the experiment is insufficient, the system is underdetermined and cannot lead to a uniquely identifiable model. The number of measurements should be determined dependent on the measurement error, the variation in the data, the number of biomarkers

investigated in an experiment and the sparseness of the connectivity of the network. The sampling rate should in theory be i) adjusted to the scale in which the biomarker variation is expected to occur, ii) adjusted to the speed in which changes are occurring in certain time spans, if known in advance, and iii) higher than $2^K[K + log(N)]$, with K denoting regulatory inputs per gene and N denoting the number of biomarkers. (McKinney, 2009)

As the number of regulatory inputs per gene is typically unknown, the presented approach uses interpolation between measurement time points. Nevertheless a careful interpretation of the results gained by the presented data integration method is required because high noise levels in the system might cause deviation from smooth profiles. As stated above, inferred molecular links are thus rather a basis for further experimental validation than the high confidence outcome of an in silico experiment.

Furthermore, usage of high sampling rates is usually restricted due to limited financial ressources. This requires a thorough deliberation of the required experimental output and the actual aim of the study beforehand. Both from a methodological as well as biological point of view defining the focus of the study proves to be crucial, as for certain research questions dynamic cellular processes play a role e.g. cell cycle processes. In that case cellular synchronization is necessary to avoid interfering molecular dependencies.

5.3.4 Data standardization

Data standardization is a major challenge when working with multiple data sets from different platforms. This is reflected in efforts of integrative databases, in enforcement of database standards (Field et al., 2009), in agreements on a minimal set of information when publishing experimental data (Burgoon, 2006), and in efforts for unique nomenclature (Gray et al., 2016). Apart from issues addressing rather individual data set annotations, an important problem in data integration that is frequently underestimated is thus the challenge of integrating data sets from different platforms, which are archived in different databases. Here, questions on data file formats are an issue, as well as annotation on possible preprocessing steps performed on the data. On top, if further prior knowledge is involved when integrating data, different data types and pathway information from pathway databases needs combined processing, which today still requires individual solutions.

Dependent on the methodological approach of integration, the actual pooling of the data can take place in very different steps of the data analysis (compare Subsection 1.3.3 *Integration approaches*). This gives a certain range of complexity to the formal integration problem, but the selection of the integration step should always be done in regard to the biological question.

In this work HUGO gene symbols were used for ID matching with databases. Different nomenclature was translated and then mapped against databases using these IDs. Furthermore, the BioPAX format using standard OWL (RDF/XML) syntax (a pathway language exchange format) was used to extract pathway knowledge such as gene sets and their corresponding topology. This enabled using large and widely known pathway databases but also

entailed the exclusion of databases not providing their pathways in BioPAX format. Though it was not the focus of this work, translation between such different formats is possible, yet it might require working with partly different information from different databases, showing again the importance of standardization.

Furthermore, certain simplifications in the integration process connected to data annotation and hence connected to standardized data storage were accepted. These include i) non-unified origin of knowledge in biological databases (different cell lines, different tissues) as discussed in Section 5.3.2 *Database biases and restrictions* and ii) no individual consideration of combinations of phosphorylations (multiplicity) compared to single phosphorylations, affecting downstream signaling of phosphoproteins as well. The first is commonly the case when pathway-based methods are used and can be resolved with more specific pathway models in future. The second is part of the data reduction process during integration and needs to be addressed in future work in a more detailed way.

6 Conclusions and Outlook

The focus of this thesis was to develop an integration approach for proteome and transcriptome high-throughput data, as growing numbers of coupled omics data sets on different molecular levels are publicly available. However, time-series data sets covering multiple molecular levels are still rare. To scientifically evaluate any improvement in terms of modeling a molecular system functionally, a systematic analysis on the added value of integration of additional data types is needed. Evaluation criteria could include identification of biomarkers or identification of promising therapeutic targets. Different molecular layers include e.g. miRNA expression, epigenetic regulation, mutation data etc. and are of considerable interest for approaching a more holistic view of multi-layer data analysis. Such additional layers can be included into the presented integrative analysis very easily in case prior biological knowledge is available for linking the layers in a database-guided way.

Though with the presented integration approach some limitations are faced as discussed in Section 5.3 *Limitations of the presented cross-platform integration approach*, it can be easily extended to other species for which a sufficient amount of public biological knowledge is available. Further specification could be accomplished by using specific disease databases or databases that are tissue-specific. In fields with very dense data availability, such as cancer research, this specification could be based on specific cancer pathway databases (e.g. Atlas of Cancer Signalling Network, Kuperstein et al. (2015)). Also the so-called disease map approach is currently further explored for a number of diseases with high research focus.

Moreover, the stimulation data presented here deals with cell population measurements, such that the observed expression changes and the integration results cannot be directly transferred to individual cell signaling. Yet, the number of single cell approaches for expression measurements is constantly increasing. These approaches can give deeper insights into individual cellular mechanisms and cell-cell-communication in case of e.g. tumor samples which include both tumor and stroma cells so that a clear distinction of cellular expression levels can only be applied when single cell measurements are used. Our presented integration approach could easily be applied to such measurements, in case parallel extraction and measurement of different molecular levels is feasible.

It is important to note that the presented results are reflecting only one specific state of either physiological or pathological signaling. For an in-depth understanding, an additional comparison of these integration results to the opposite state would be highly beneficial.

In this way, identifying either temporal or molecular deregulation would be possible and potential therapeutic targets could be predicted.

The presented integration approach combines diverse information sources in order to yield better linkage between the measured data sets by making use of already known molecular interactions. In this context molecular dependencies were identified for future experimental validation that are already known in other contexts. Furthermore, response-specific signaling can be tracked through different molecular layers with the presented approach. However, as discussed in Section 5.3 *Limitations of the presented cross-platform integration approach* no knowledge on interactions of newly measured molecules is possible so far. Thus benefits of high sensitivity screening methods are not exploited, as no newly identified interactions are highlighted. Nevertheless, this might form an interesting objective for further optimization.

In summary, the presented integration approach can clearly provide guidance in the formation of further experimental hypotheses to elucidate the complex cellular signaling response upon perturbation of the system. Furthermore, it can prove consideration of individual molecular levels to be valuable for cross-platform integration in terms of structuring and focusing results for biological interpretation. The R package which was developed in the course of this thesis ('pwOmics') is publicly available, can be applied for data from single time points or time series data sets and facilitates exploiting different open source databases.

7 Appendix

This appendix provides the supplementary material of Chapter 4 *Integration of phosphoproteome and transcriptome data to link B cell receptor activation with gene expression dynamics*:

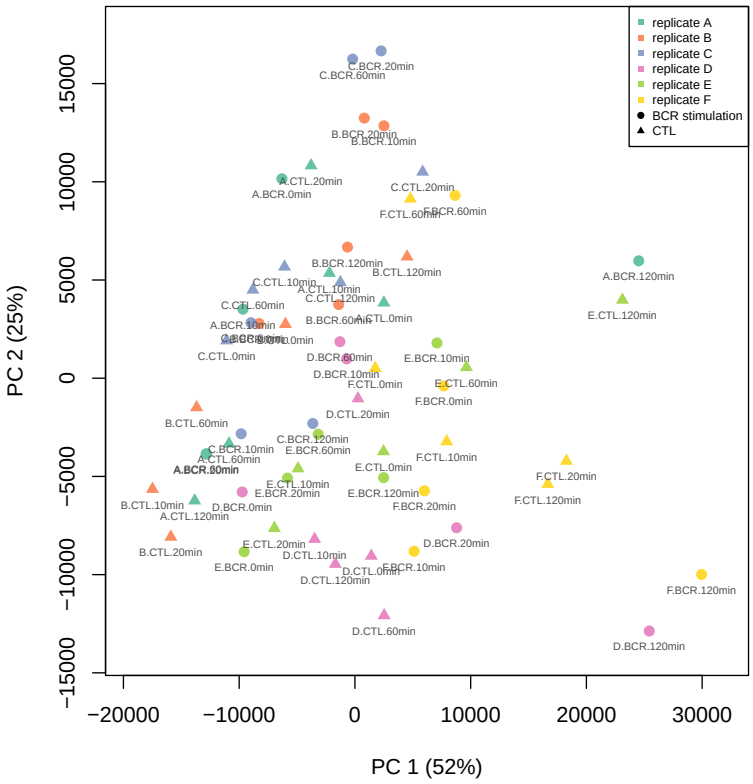


Figure S1. *Principal component analysis of RNA-Seq data set. Normalized log2 counts were prefiltered to those 500 transcripts showing the highest variance over all samples prior to principal component analysis. Replicates are visualized in different colors, while BCR stimulation and control measurements are plotted in filled circles and triangles, respectively. No strong outlier can be detected, replicate measurements predominantly cluster together.*

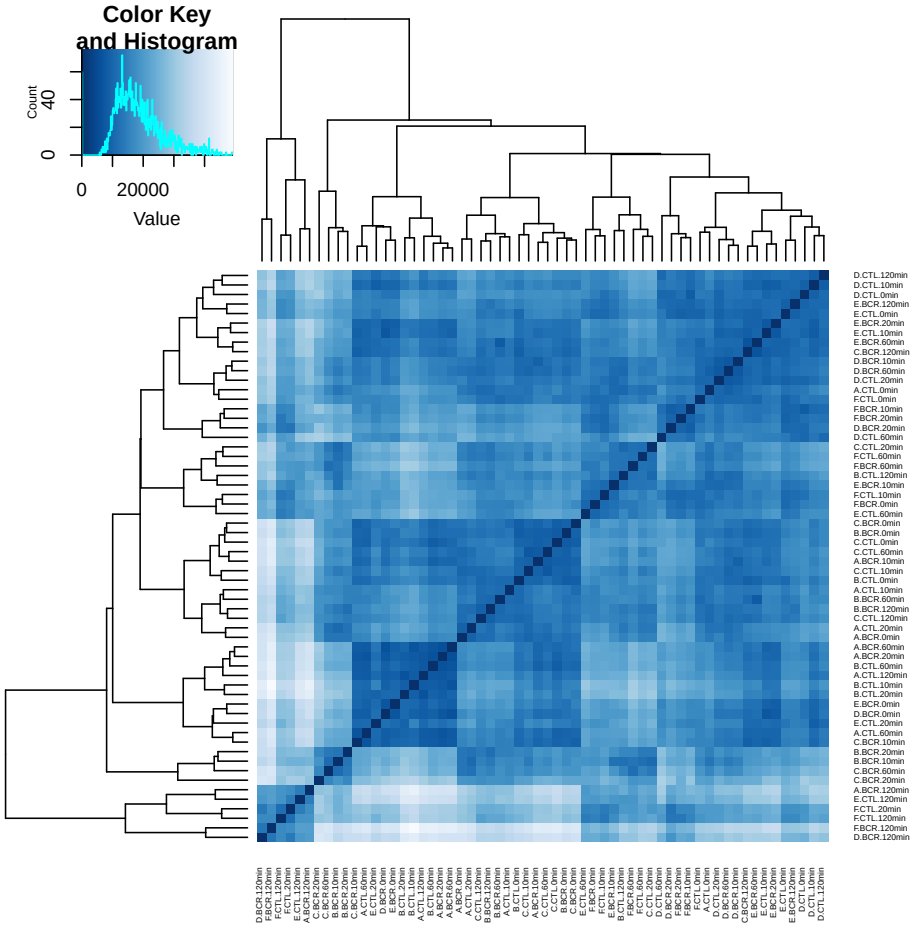


Figure S2. *Sample heatmap of RNA-Seq data set. Transcriptome data is displayed plotting sample-to-sample distances in a heatmap showing Euclidean distance between the samples and individual replicates. Normalized log2 counts were used to ensure stabilized variance.*

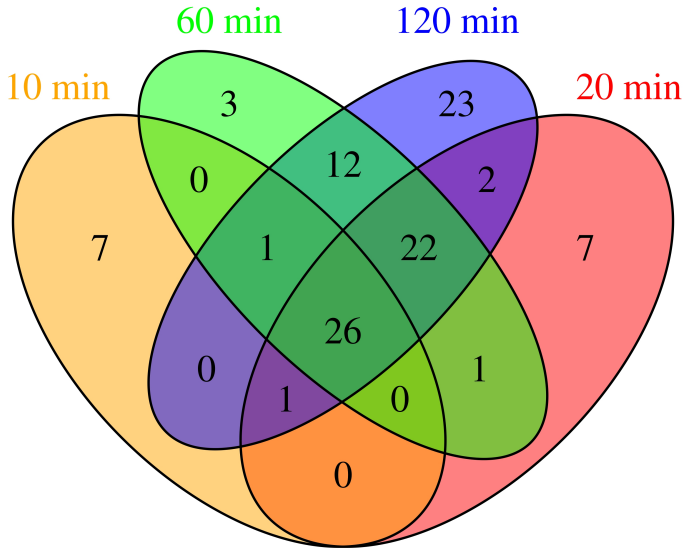


Figure S3. *Venn diagram showing overlap of significantly regulated transcripts for different stimulation times. High overlap of significantly regulated transcripts at late time points (60 and 120 min of BCR stimulation) can be observed, whereas short BCR stimulation durations show less overlap.*

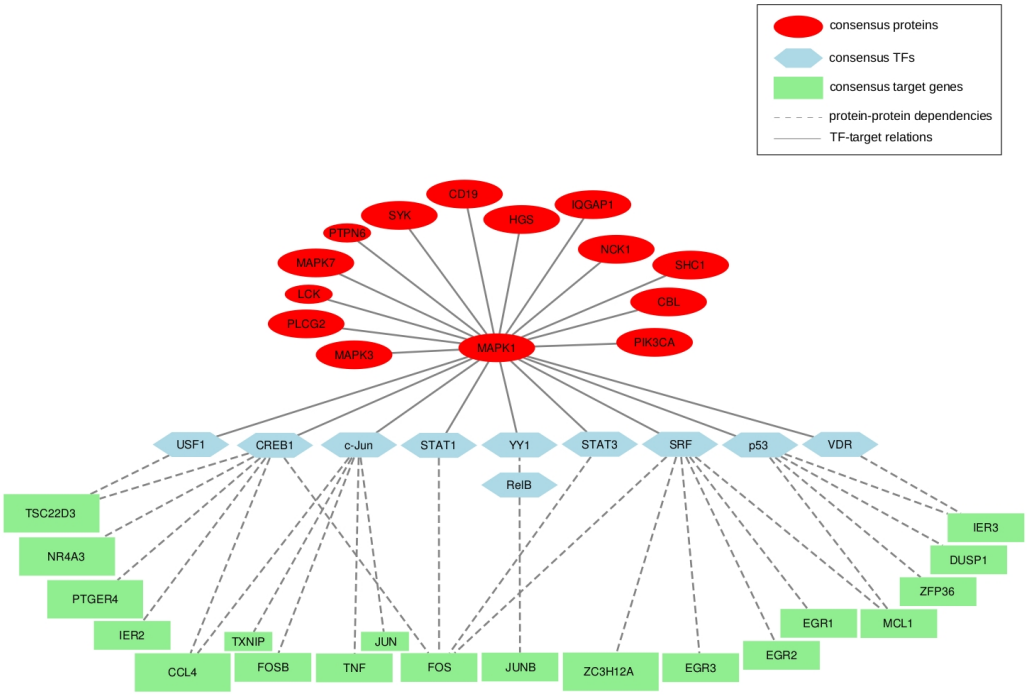


Figure S4. *Consensus graph based on same measurement time points of phosphoproteome and transcriptome data. Small node sizes indicate nodes identified in consensus graph based on data from 10 min of BCR stimulation, big node sizes correspond to nodes identified in consensus graphs based on data from 20 min of BCR stimulation. Intermediate node sizes indicate nodes identified in consensus graphs based on data from both 10 min and 20 min BCR stimulation times.*

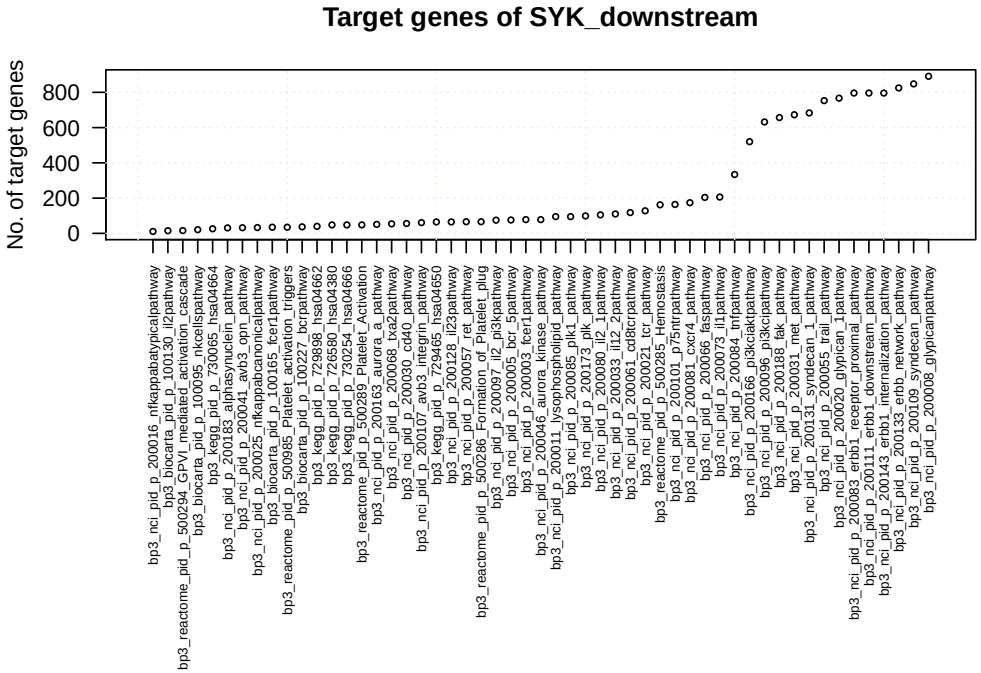


Figure S5, A. *Signaling axes downstream of SYK. Identified signaling pathways with the corresponding numbers of their target genes are displayed. For each pathway the biopax version, internal pathway IDs and pathway names are given.*

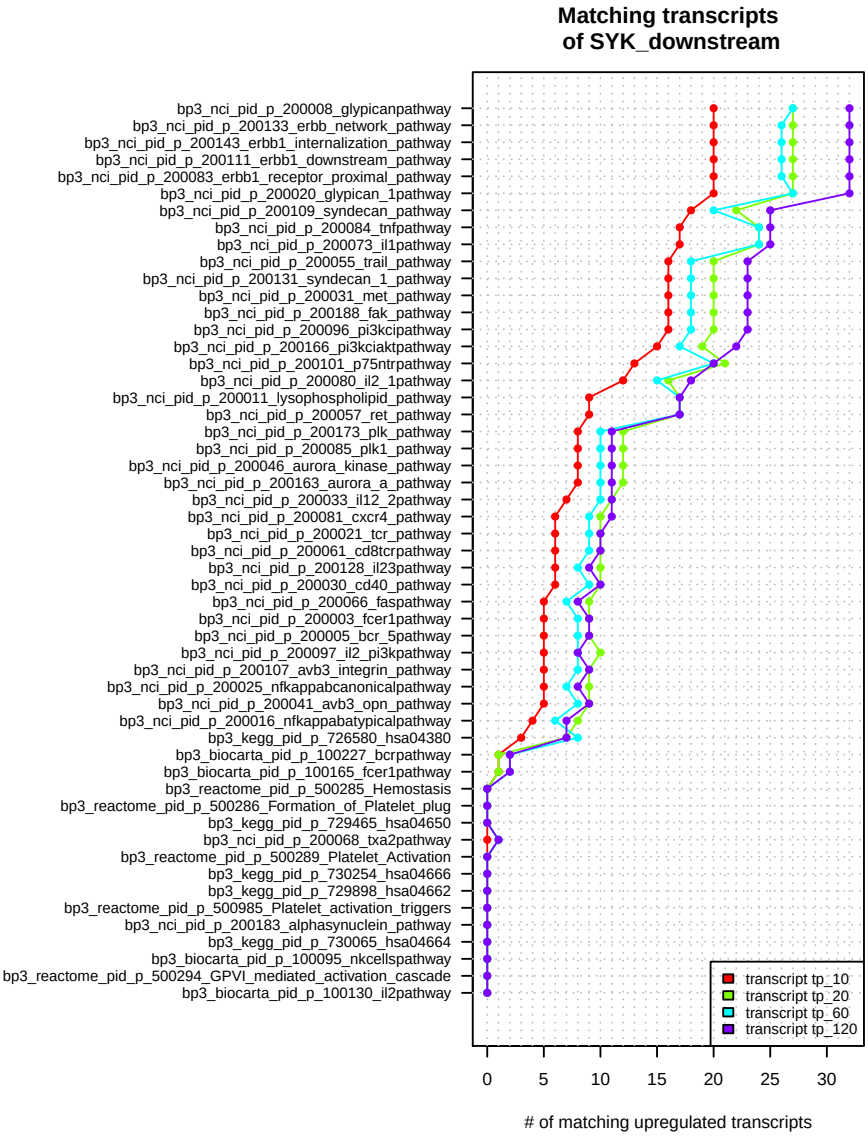


Figure S5, B. *Signaling axes downstream of SYK. Number of target genes matching to upregulated transcripts per signaling pathway. For each pathway the biopax version, internal pathway IDs and pathway names are given. 'Red' color indicates 10 min, 'green' indicates 20 min, 'cyan' indicates 60 min and 'purple' indicates 120 min of BCR stimulation in transcriptome data set.*

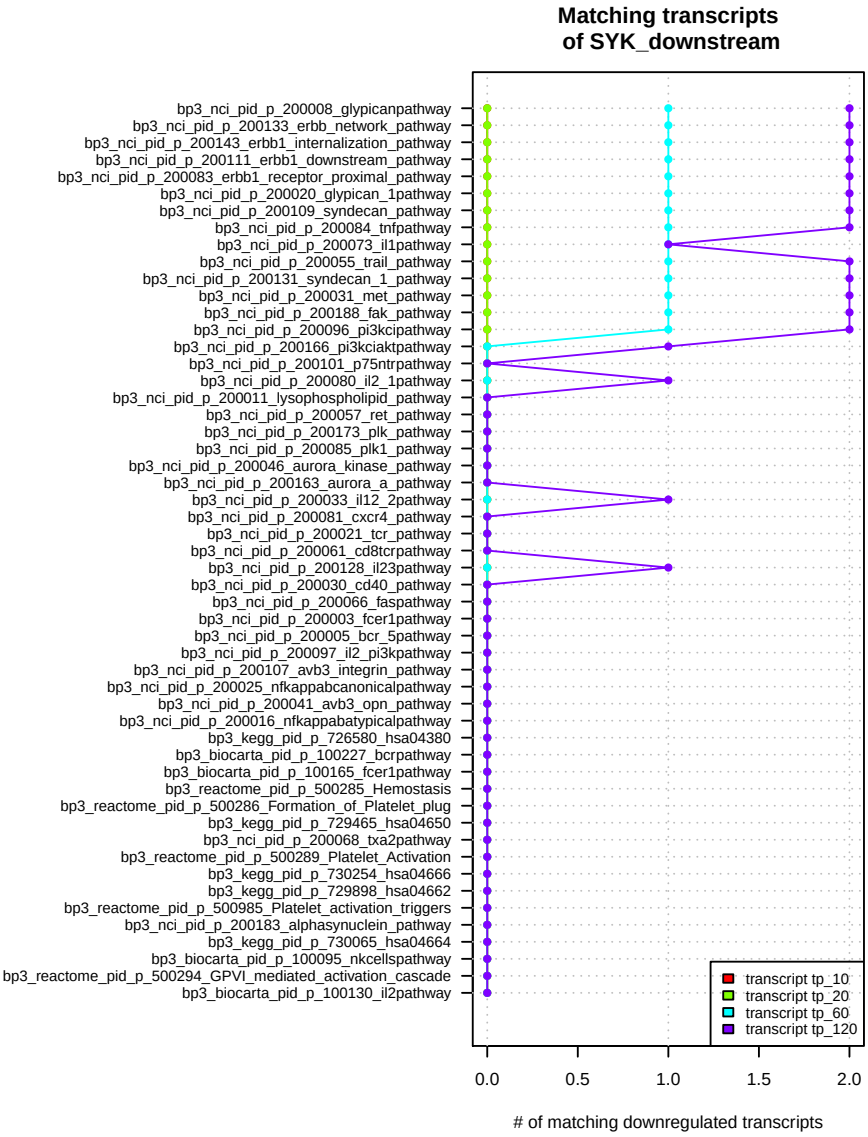


Figure S5, C. *Signaling axes downstream of SYK. Number of target genes matching to downregulated transcripts per signaling pathway. For each pathway the biopax version, internal pathway IDs and pathway names are given. 'Red' color indicates 10 min, 'green' indicates 20 min, 'cyan' indicates 60 min and 'purple' indicates 120 min of BCR stimulation in transcriptome data set.*

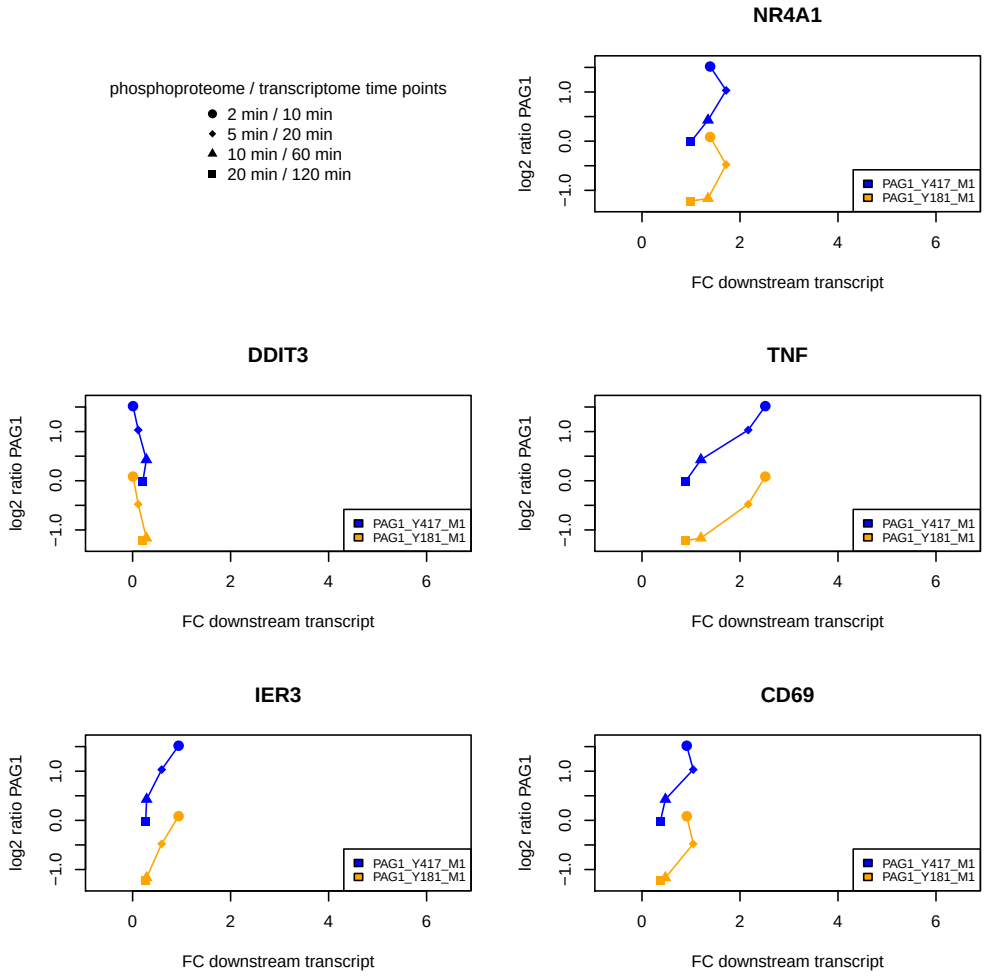


Figure S6, A. *Exemplary correlation trajectories of PAG1. Shown are correlations of PAG1 phosphosites with downstream transcripts NR4A1, DDIT3, TNF, IER3 and CD69. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

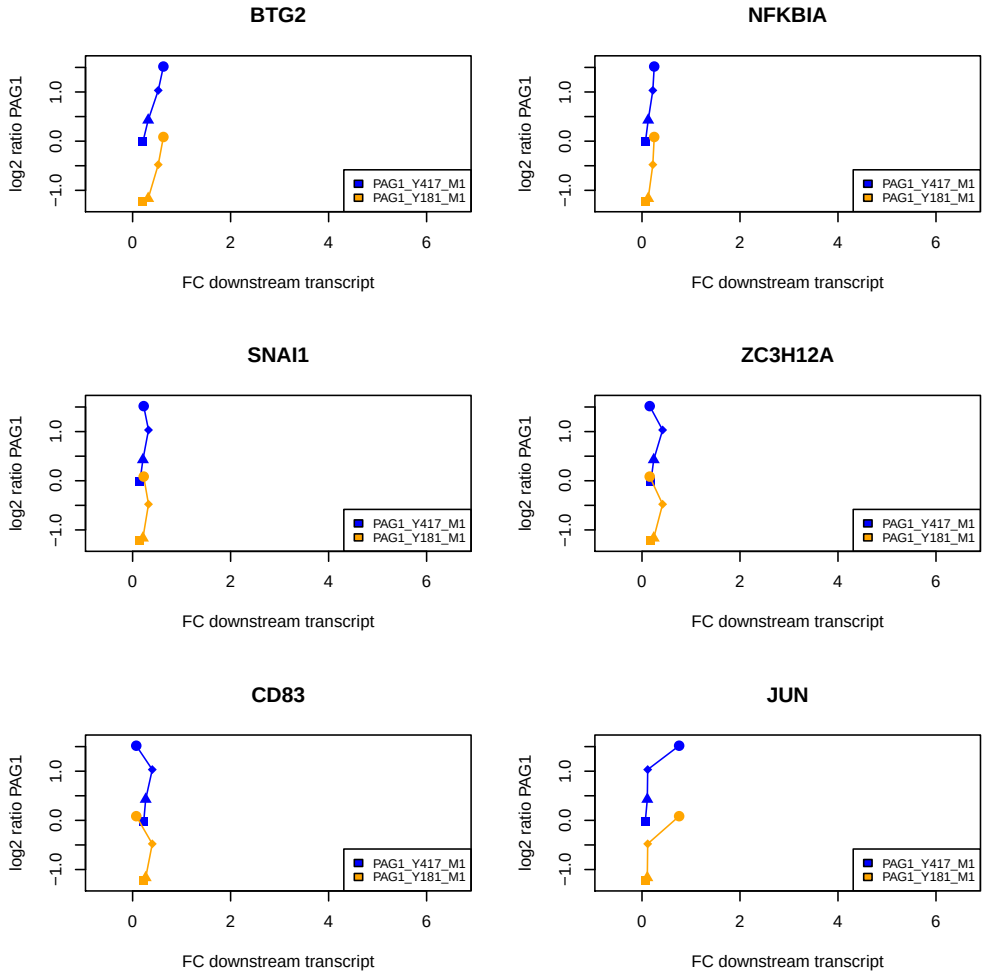


Figure S6, B. *Exemplary correlation trajectories of PAG1. Shown are correlations of PAG1 phosphosites with downstream transcripts BTG2, NFKBIA, SNAI1, ZC3H12A, CD83 and JUN. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

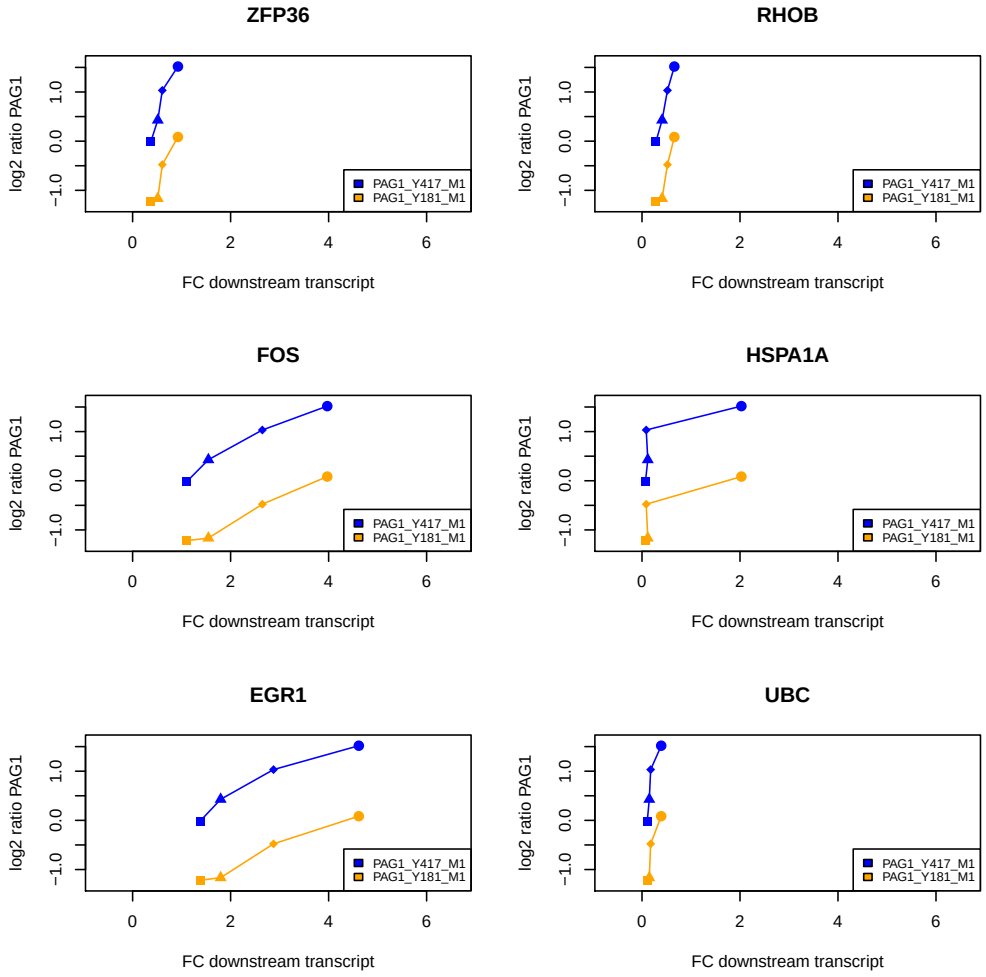


Figure S6, C. *Exemplary correlation trajectories of PAG1. Shown are correlations of PAG1 phosphosites with downstream transcripts ZFP36, RHOB, FOS, HSPA1A, EGR1 and UBC. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

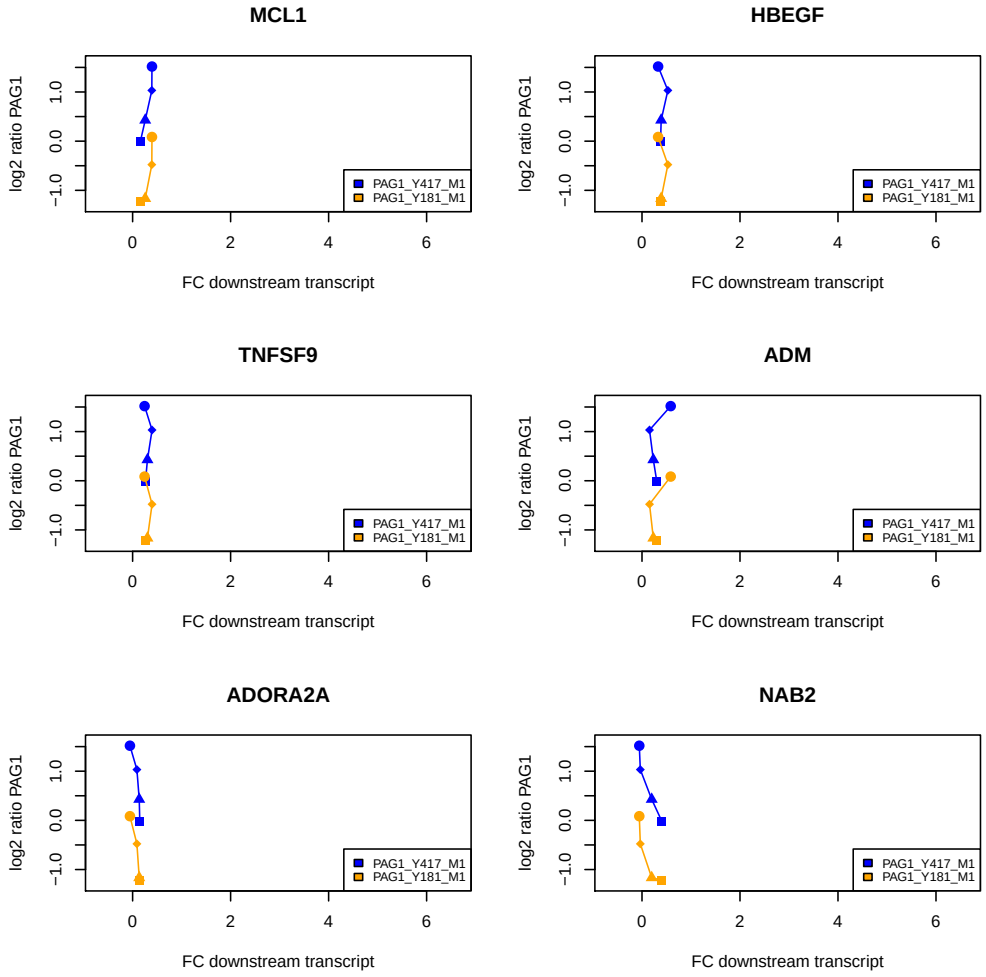


Figure S6, D. *Exemplary correlation trajectories of PAG1. Shown are correlations of PAG1 phosphosites with downstream transcripts MCL1, HBEGF, TNFSF9, ADM, ADORA2A and NAB2. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

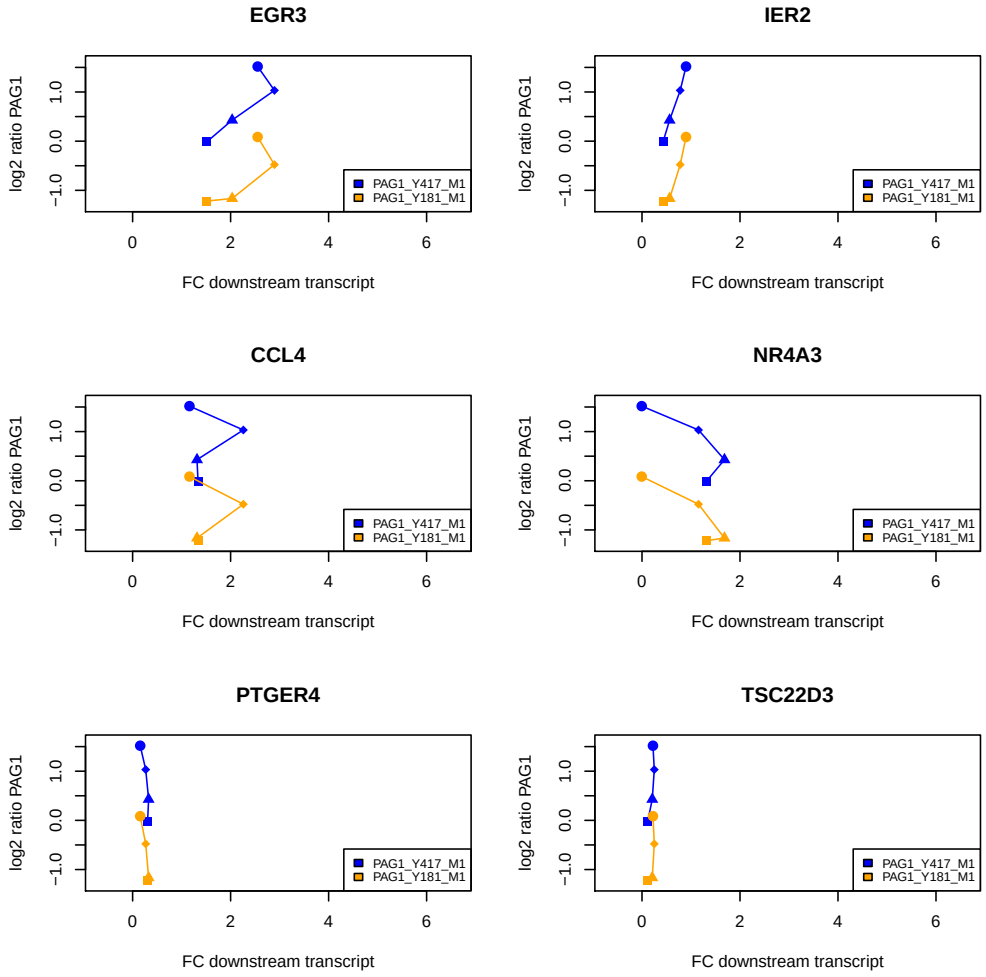


Figure S6, E. *Exemplary correlation trajectories of PAG1. Shown are correlations of PAG1 phosphosites with downstream transcripts EGR3, IER2, CCL4, NR4A3, PTGER4 and TSC22D3. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

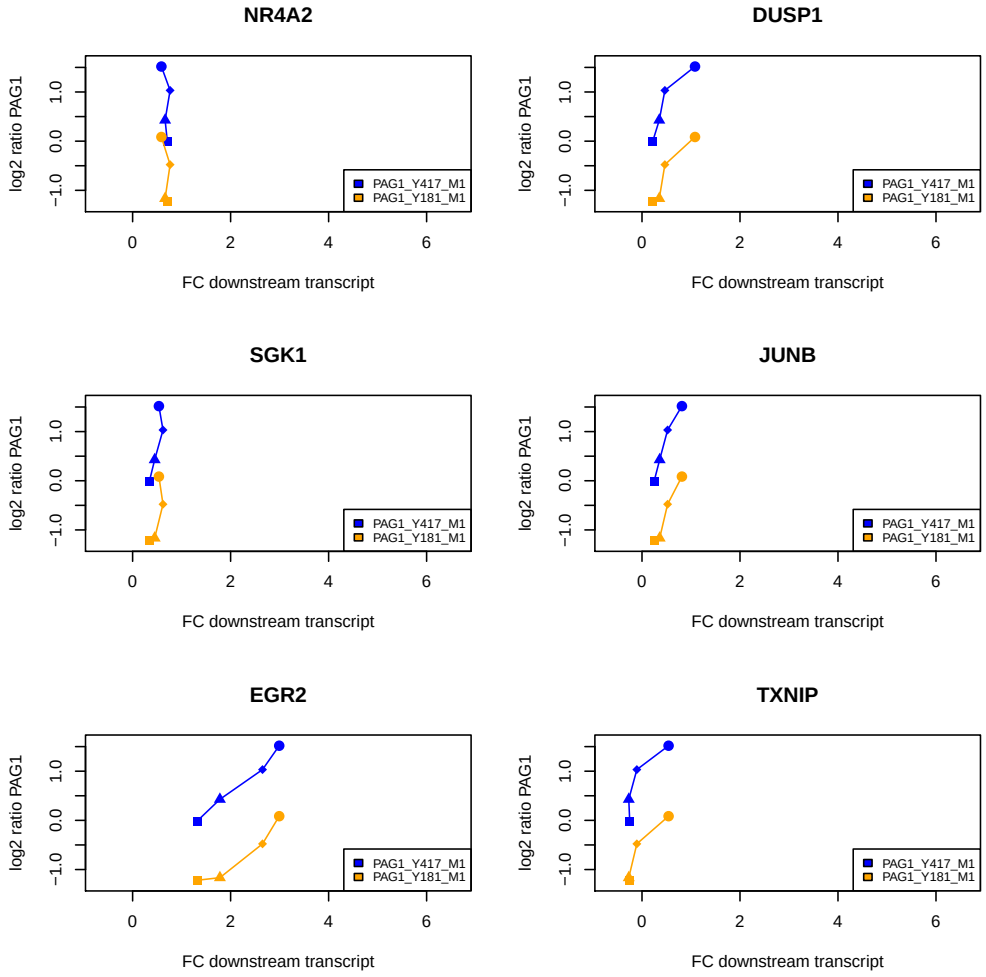


Figure S6, F. *Exemplary correlation trajectories of PAG1. Shown are correlations of PAG1 phosphosites with downstream transcripts NR4A2, DUSP1, SGK1, JUNB, EGR2 and TXNIP. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

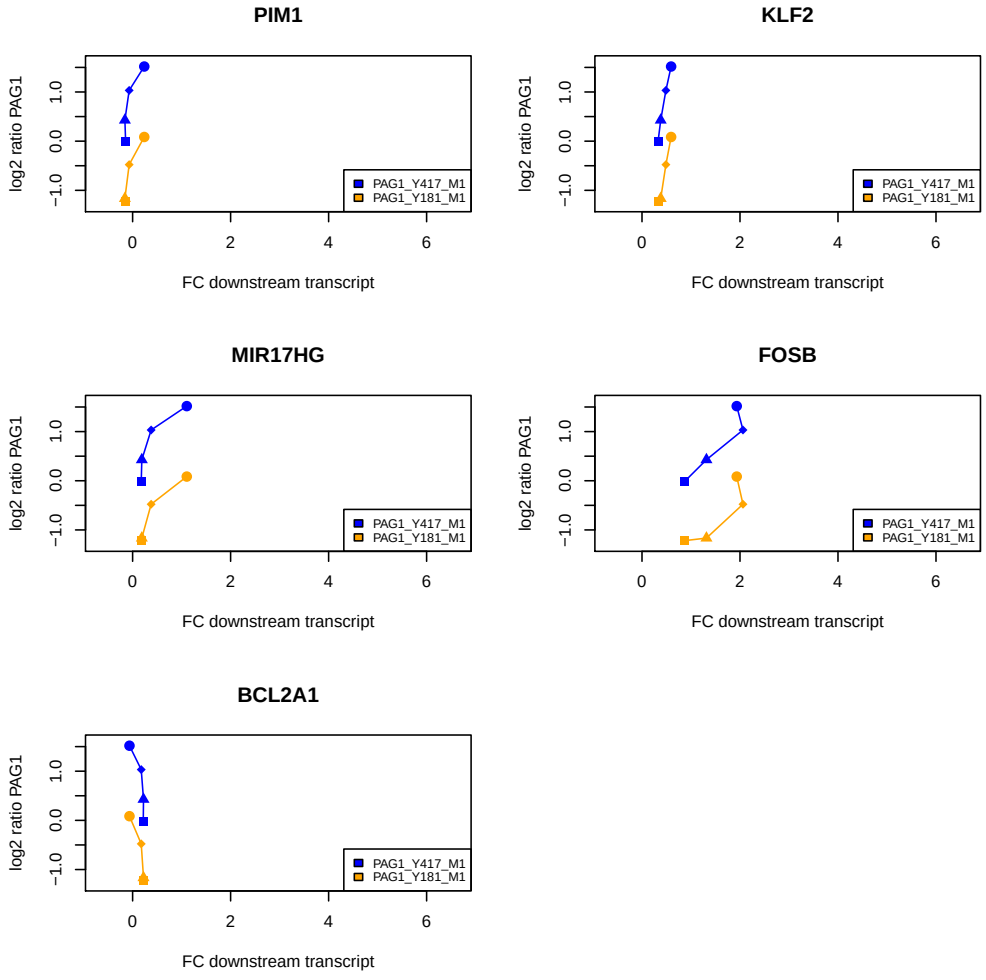


Figure S6, G. *Exemplary correlation trajectories of PAG1. Shown are correlations of PAG1 phosphosites with downstream transcripts PIM1, KLF2, MIR17HG, FOSB and BCL2A1. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

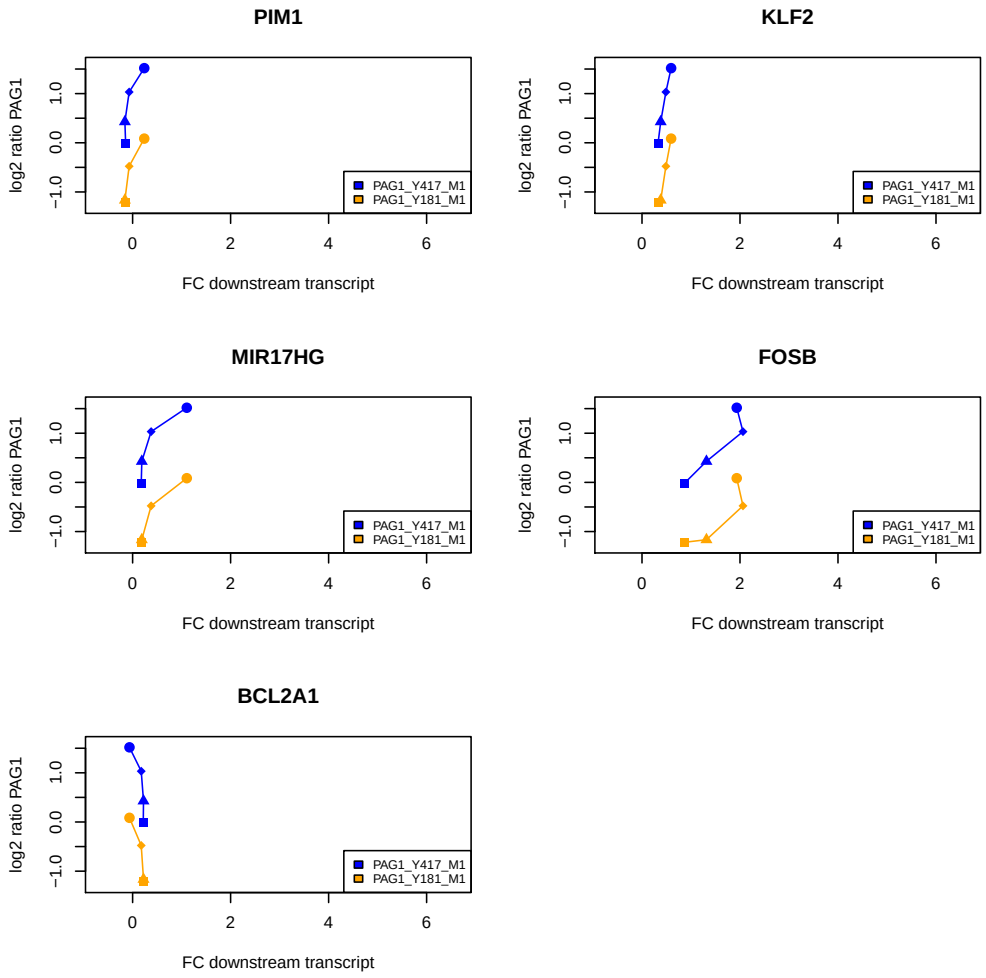


Figure S6, H. *Exemplary correlation trajectories of PLCG2. Shown are correlations of PLCG2 phosphosites with downstream transcripts FOS, EGR3, IER2, CCL4 and NR4A3. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

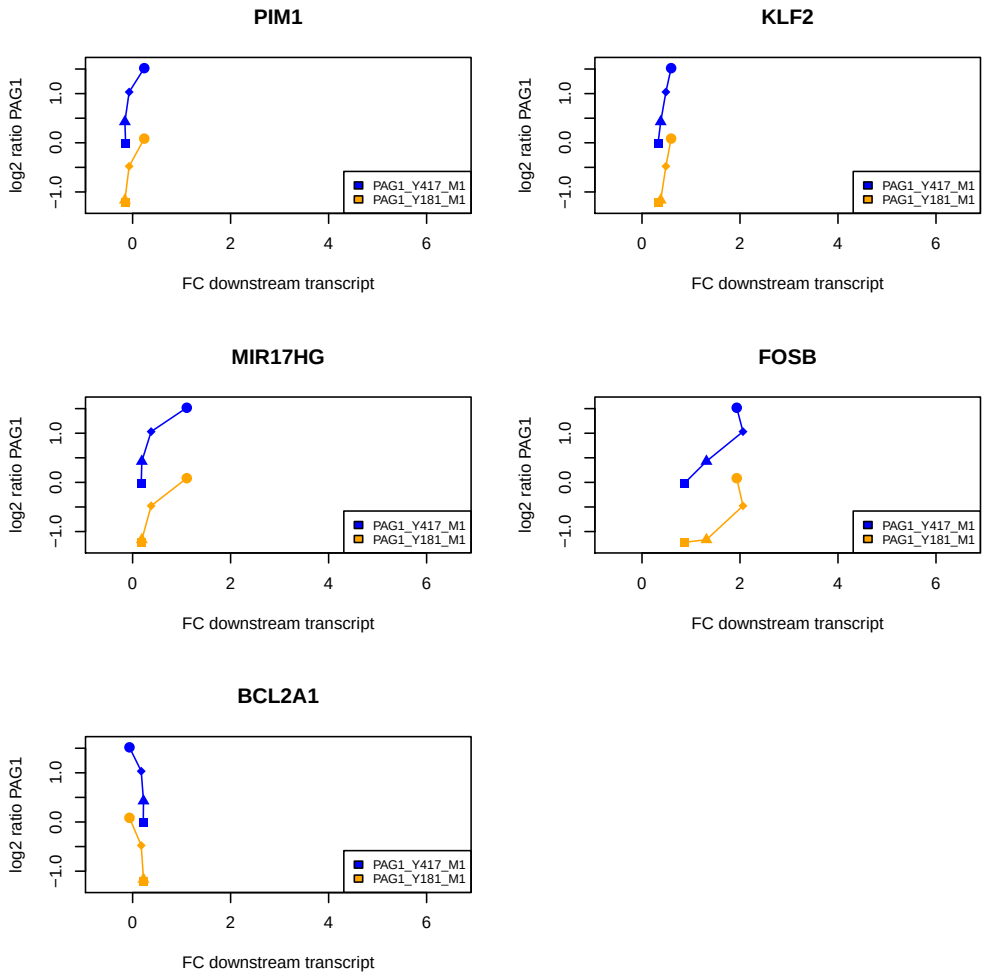


Figure S6, I. *Exemplary correlation trajectories of PLCG2. Shown are correlations of PLCG2 phosphosites with downstream transcripts PTGER4, TSC22D3 and NR4A2. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

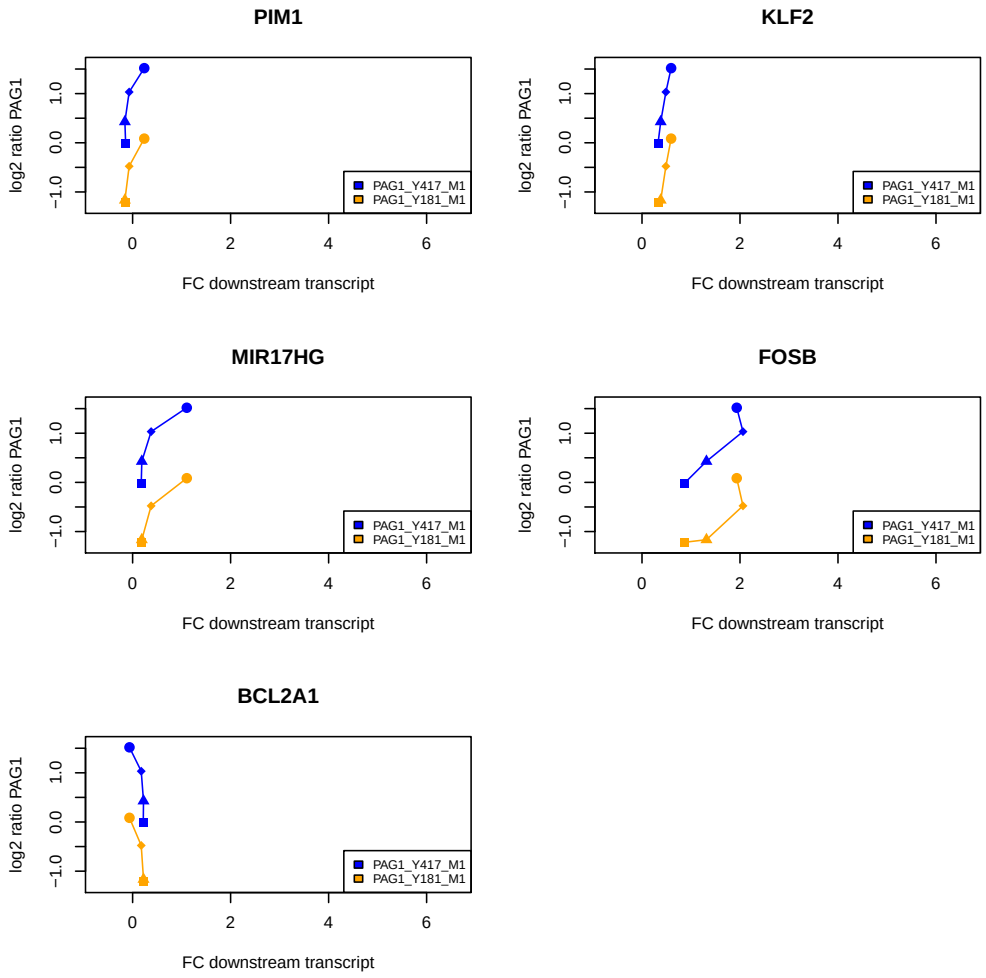


Figure S6, J. *Exemplary correlation trajectories of PTPN6. Shown are correlations of PTPN6 phosphosites with downstream transcripts FOS, EGR2, EGR1, EGR3 and MCL1. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

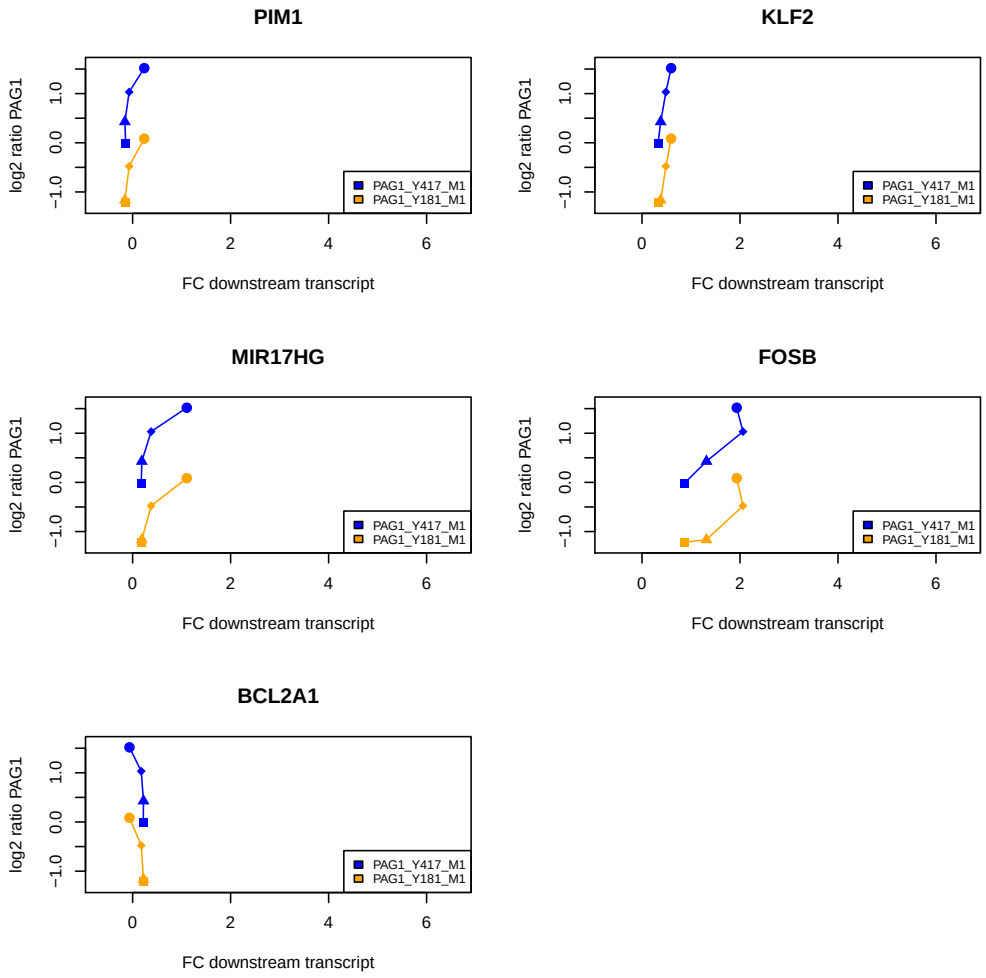


Figure S6, K. *Exemplary correlation trajectories of PTPN6. Shown are correlations of PTPN6 phosphosites with downstream transcripts ZC3H12A, NR4A1, DDIT3 and PIM1. Individual sites are annotated, including indication of single (_M1) or multiple phosphorylation events (_Mx). Plotting symbols denote different BCR stimulation durations used for correlation analysis: a filled circle indicates 2 min of BCR stimulation of phosphoproteome data and 10 min of BCR stimulation duration in transcriptome data. A filled square, a filled triangle and a filled diamond indicate 5 and 10 min of BCR stimulation, 10 and 60 min of BCR stimulation and 20 and 120 min of BCR stimulation of phosphoproteome and transcriptome, respectively.*

transcript	FC 10 min	FC 20 min	FC 60 min	FC 120 min	pval 10 min	pval 20 min	pval 60 min	pval 120 min
IER2	0.898877	0.77991	0.56392	0.429895	0	0	0	0
DUSP2	1.44534	0.939853	0.688184	0.51372	0	0	0	0
EGR1	4.61662	2.87565	1.79698	1.38	0	0	0	0
FOS	3.97273	2.64775	1.54659	1.10914	0	0	0	0
ZFP36	0.925194	0.605675	0.513937	0.37069	0	0	0	0
DNAJB1	1.71619	0.269482	0.191192	0.115444	0	0.000175681	0.0190433	0.0429147
HSPA6	5.11522	0.0307162	0.0519797	0.0890261	0	1	1	1
HSPA1A	2.02843	0.0861656	0.118637	0.0711855	0	1	1	1
HSPA1B	1.92289	0.073793	0.11578	0.0716291	0	1	1	1
EGR2	2.99214	2.65119	1.78136	1.31823	0	0	0	0
DUSP1	1.08208	0.467116	0.3528	0.216036	0	5.76087e-10	8.42387e-08	2.59081e-06
JUNB	0.814865	0.522757	0.360846	0.246827	1.777e-13	0	3.135e-11	4.86337e-11
FOSB	1.93343	2.06105	1.31529	0.870958	8.09986e-10	0	0	0
JUN	0.758139	0.116225	0.10874	0.067803	1.11433e-08	1	1	1
BTG2	0.628427	0.52281	0.317804	0.210104	1.98188e-07	0	1.608e-08	6.22389e-09
IER5	0.625487	0.193644	0.119639	0.0706648	1.23749e-06	0.0926485	1	1
TNF	2.51687	2.16818	1.2023	0.888341	1.27205e-06	0	0	0
CHAC1	0.674111	0.297348	0.27795	0.224405	1.27205e-06	0.000203632	1.54549e-05	7.07964e-09
PPP1R15A	0.593314	0.343875	0.343117	0.231689	2.52977e-06	4.62114e-08	5.3425e-10	6.0353e-11
HSPA7	4.30952	-0.346937	-0.224317	-0.166723	4.27544e-06	1	1	1
NR4A1	1.39024	1.71702	1.34756	0.996219	2.52209e-05	0	0	0
MIR17HG	1.1053	0.378228	0.187848	0.17612	3.12949e-05	0.00772546	0.320875	0.0456173
TXNIP	0.542619	-0.105036	-0.270113	-0.250803	4.81715e-05	1	1.06866e-05	1.1524e-12
EGR3	2.55122	2.89461	2.03173	1.51508	0.000167522	0	0	0
CD69	0.912548	1.04559	0.477789	0.375597	0.000288893	0	0	0
IER3	0.938795	0.59316	0.285875	0.26671	0.000415033	6.42588e-08	0.0258843	1.24141e-05
CCL4L1	2.05041	2.38899	1.7281	1.38554	0.000670149	0	0	0
ZFP36L1	0.475482	0.368497	0.384052	0.253746	0.0010698	1.60213e-09	8.54117e-13	2.67522e-13
CCL4L2	2.26746	2.0456	1.59999	1.34184	0.0010698	2.35029e-11	0	0
CCL3L1	0.924158	1.49624	1.20157	0.964506	0.0179711	0	0	0
UBC	0.392394	0.17699	0.148842	0.107556	0.0266059	0.12939	0.315449	0.0836633
MCL1	0.394966	0.391784	0.257066	0.156282	0.0279714	2.49597e-11	2.54994e-05	0.000135787
IER5L	0.769104	0.497417	0.39815	0.293665	0.0315334	0.000710678	1.5555e-06	7.56668e-06
TOB1	0.435011	0.268778	0.205271	0.153283	0.0315477	0.00092251	0.0122771	0.000909436
EGR4	1.4726	1.83762	1.51255	1.1192	0.0385073	0	0	0
CCL3	0.567912	1.42124	1.15321	0.933006	1	0	0	0
CCL3L3	0.87778	1.53734	1.25659	1.04811	0.136349	0	0	0
DUSP5	0.309166	0.751598	0.563627	0.449869	1	1.12454e-13	0	0
SGK1	0.536627	0.619831	0.450235	0.341725	0.354302	6.40987e-13	2.69759e-10	2.14018e-12
SIK1	0.238515	0.434018	0.231935	0.168373	1	5.02571e-11	0.00163092	0.000159121
ZC3H12A	0.157208	0.420883	0.239586	0.181997	1	2.27337e-09	0.00148142	0.000376916
CD83	0.0743205	0.401075	0.266585	0.228204	1	1.57978e-06	0.00754083	7.90662e-10
NR4A3	-0.00544226	1.15378	1.68301	1.31272	1	1.88483e-05	0	0
SERTAD1	0.203868	0.338169	0.231264	0.17695	1	0.000100254	0.00595974	0.000322009
RGS16	0.330978	0.292666	0.210072	0.133905	0.807781	0.000100323	0.00742463	0.00924937
CSRNP1	0.0778398	0.31181	0.379539	0.299893	1	0.00012462	1.47149e-10	0
IL8	0.275687	0.480222	0.388541	0.224814	1	0.000463558	4.39811e-05	0.00615824
HBEGF	0.329401	0.528039	0.390946	0.379289	1	0.000896118	1.63983e-08	0
CXCR5	0.064964	0.323173	0.187827	0.19024	1	0.00126797	1	0.0086598
PPP1R10	0.0790118	0.251378	0.171453	0.0874702	1	0.00190574	0.109988	0.654776
MYADM	0.0743141	0.300749	0.115954	0.124214	1	0.00205196	1	0.360549
SNAI1	0.226339	0.323897	0.210034	0.149621	1	0.00214554	0.0685462	0.0283572
TSC22D3	0.223783	0.25113	0.209838	0.117673	1	0.00221987	0.00651308	0.0524958
TNFSF9	0.244224	0.396563	0.302602	0.261929	1	0.00230716	3.75129e-05	4.34489e-06
RGS1	0.214969	0.260546	0.390184	0.250579	1	0.00263407	5.59243e-12	2.29305e-11
ZNF547	0.181317	0.412882	0.138806	0.143287	1	0.00354656	1	0.414741
ICAM4	0.388421	0.867151	0.326737	0.339863	1	0.00488034	0.826387	0.232101
NFKBIA	0.24957	0.221316	0.126744	0.0702029	1	0.00708412	1	1
CCL4	1.1601	2.25967	1.31439	1.335	1	0.0086615	0.0432062	0.000288599
NRARP	0.338102	0.301058	0.230631	0.199815	1	0.0123162	0.0192497	0.000560227
KCNJ2	0.0519212	0.876718	0.657892	0.456855	1	0.0132692	0.00162858	0.00315792
ZNF124	0.284766	0.288163	0.14227	0.111868	1	0.0170912	1	0.459707
SPRY2	0.281906	0.277646	0.341832	0.247883	1	0.0181058	4.72061e-07	2.87123e-08
RND1	0.219337	0.247913	0.212622	0.160778	1	0.0244133	0.0166918	0.00153999
TMEM88	0.325295	1.00165	0.555419	0.472756	1	0.0289874	0.0131198	0.0190766
PTGER4	0.154874	0.267611	0.326834	0.305588	1	0.0306995	2.04945e-06	4.28035e-13
UGCG	0.254038	0.317009	0.0760837	0.118064	1	0.0348147	1	0.347963
DDIT4	0.259479	0.0504784	0.325925	0.257938	1	1	4.76113e-09	9.30512e-14
NFKBID	0.176451	0.39693	0.36376	0.298222	1	0.0687118	2.34912e-07	0.000210738
KBTBD8	0.171403	0.208504	0.295293	0.210311	1	0.0575247	1.17252e-06	6.05144e-08
DDIT3	0.00843464	0.114505	0.280935	0.208262	1	1	5.14104e-06	4.79955e-08
BHLHE40	0.23153	0.251245	0.312059	0.268738	1	0.0614772	5.93692e-06	1.4781e-10
LOC284454	0.0346837	0.738084	0.543331	0.452365	1	0.557436	1.47303e-05	0.000455611
MAFF	0.125756	0.336047	0.378709	0.291694	1	0.240532	5.20475e-05	3.35079e-06
GEM	0.254578	0.303712	0.344643	0.240525	1	0.225907	0.000114127	0.000116145
KDM6B	0.121974	0.190032	0.238244	0.187004	1	0.0872722	0.000260815	7.67399e-05
NR4A2	0.587349	0.764722	0.663135	0.709053	1	0.258929	0.00511642	3.08752e-07
KLF10	0.204841	0.051959	0.216179	0.170003	1	1	0.00548948	9.09888e-05
DLX2	-0.0445832	0.428411	0.412269	0.273898	1	0.518874	0.0122771	0.0282669
RASGEF1B	0.258688	0.22648	0.295403	0.122199	1	1	0.0126346	1
RHOB	0.660251	0.519637	0.410082	0.284112	1	0.126685	0.018907	0.0509604
LOC100270804	0.0246037	-0.202219	-0.39149	-0.0626732	1	1	0.0350282	1
NAB2	-0.0549711	-0.0340899	0.198013	0.402793	1	1	0.0768889	0
SRGN	0.0562056	0.25569	0.159058	0.17678	1	0.0547523	0.280674	2.58658e-05
HVCN1	-0.0666527	0.171103	0.152464	0.172623	1	0.471036	0.384898	4.95841e-05
EFNA2	0.985101	0.754158	0.627064	0.661192	1	0.939154	0.0709778	0.000103733
LOC100302650	-0.039493	0.490527	0.430506	0.469478	1	1	0.196009	0.000329886
SEMA7A	-0.255297	0.141834	0.349018	0.380536	1	1	0.190073	0.000547502
ADM	0.586198	0.153728	0.231286	0.302101	1	1	0.338612	0.00127502
PRRG4	0.268596	0.264592	0.225374	0.218519	1	0.707693	0.21935	0.00130399
BCL2A1	-0.0641556	0.175936	0.22193	0.217221	1	1	0.227679	0.00145361
KLF2	0.594434	0.487803	0.382199	0.327254	1	0.0926485	0.0518502	0.00204494
PIM1	0.2376	-0.0704204	-0.157896	-0.142479	1	1	0.298709	0.00350623
LRRC32	-0.347494	-0.0890668	0.0637566	0.262294	1	1	1	0.00528206
RGS2	0.384135	0.13077	0.16667	0.132901	0.0707673	1	0.124289	0.00615824
MIDN	0.159223	0.180665	0.155017	0.132151	1	0.16551	0.273975	0.00650499
EVI2A	0.122465	0.215062	0.115075	0.132207	1	0.497578	1	0.0086598
RILPL2	0.0526796	0.0512769	0.184432	0.139406	1	1	0.0741345	0.0088366
KIAA1683	1.00122	0.879236	0.500903	0.576837	1	0.741506	1	0.0097694
ADORA2A	-0.0547623	0.0874963	0.13503	0.145984	1	1	0.833818	0.0229761
C14orf43	0.209615	0.136761	0.152389	0.119367	1	1	0.283218	0.0272863
C18orf1	0.0842281	0.214164	0.148843	0.1305	1	0.0926485	0.833818	0.0272863
KLHL21	0.0318724	-0.0547449	-0.15429	-0.123129	1	1	0.317873	0.0278579
LOC100507489	1.21098	1.30996	0.998263	0.932946	1	1	0.375295	0.0407371
EVI2B	0.122289	0.17445	0.109703	0.11876	1	0.274631	1	0.0429147

Table S1. *Significantly regulated transcripts. Fold changes (FC) and Benjamini-Hochberg adjusted p-values (pval) are given.*

	Biocarta	KEGG	PID	Reactome
Downstream analysis				
2 min BCR stimulation	x	x	x	
5 min BCR stimulation	x	x	x	
10 min BCR stimulation	x	x	x	
20 min BCR stimulation	x	x	x	
Upstream analysis				
10 min BCR stimulation	x		x	
20 min BCR stimulation	x		x	
60 min BCR stimulation	x		x	
120 min BCR stimulation	x		x	

Table S2. *BCR signaling related 'upstream' and 'downstream pathways' identified in integrative analysis. Pathways identified in the layer-specific pathway-based integration approach based on mapping with pathway databases Biocarta (Nishimura, 2001), KEGG (Kanehisa et al., 2014), Pathway Interaction Database (Schaefer et al., 2009) and Reactome (Fabregat et al., 2016).*

References

Altenbach, S. B., Vensel, W. H., and DuPont, F. M. (2010). Integration of transcriptomic and proteomic data from a single wheat cultivar provides new tools for understanding the roles of individual alpha gliadin proteins in flour quality and celiac disease. *Journal of Cereal Science*, 52(2):143–151.

Bahar Halpern, K., Caspi, I., Lemze, D., Levy, M., Landen, S., Elinav, E., Ulitsky, I., and Itzkovitz, S. (2015). Nuclear Retention of mRNA in Mammalian Tissues. *Cell Reports*, 13(12):2653–2662.

Balbin, O. A., Prensner, J. R., Sahu, A., Yocum, A., Shankar, S., Malik, R., Fermin, D., Dhanasekaran, S. M., Chandler, B., Thomas, D., Beer, D. G., Cao, X., Nesvizhskii, A. I., and Chinnaiyan, A. M. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nature Communications*, 4:2617.

Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8):552–564.

Boisvert, F.-M., Ahmad, Y., Gierliński, M., Charrière, F., Lamont, D., Scott, M., Barton, G., and Lamond, A. I. (2012). A quantitative spatial proteomics analysis of proteome turnover in human cells. *Molecular & cellular proteomics: MCP*, 11(3):M111.011429.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4):365–371.

Burgoon, L. D. (2006). The need for standards, not guidelines, in biological data reporting and sharing. *Nature Biotechnology*, 24(11):1369–1373.

Citri, A. and Yarden, Y. (2006). EGF–ERBB signalling: towards the systems level. *Nature Reviews Molecular Cell Biology*, 7(7):505–516.

Com, E., Boitier, E., Marchandeau, J.-P., Brandenburg, A., Schroeder, S., Hoffmann, D., Mally, A., and Gautier, J.-C. (2012). Integrated transcriptomic and proteomic evaluation of gentamicin nephrotoxicity in rats. *Toxicology and Applied Pharmacology*, 258(1):124–133.

Corso, J., Pan, K.-T., Walter, R., Doebele, C., Mohr, S., Bohnenberger, H., Ströbel, P., Lenz, C., Slabicki, M., Hüllein, J., Comoglio, F., Rieger, M. A., Zenz, T., Wienands, J., Engelke, M., Serve, H., Urlaub, H., and Oellerich, T. (2016). Elucidation of tonic and activated B-cell receptor signaling in Burkitt's lymphoma provides insights into regulation of cell survival. *Proceedings of the National Academy of Sciences of the United States of America*, 113(20):5688–5693.

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., and D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(Database issue):D472–477.

Dagum, P., Galper, A., and Horvitz, E. (1992). Dynamic Network Models for Forecasting. In *Proceedings of the Eight Conference on Uncertainty in Artificial Intelligence*, pages 41–48. AUAI Press.

Dauer, D. J., Ferraro, B., Song, L., Yu, B., Mora, L., Buettner, R., Enkemann, S., Jove, R., and Haura, E. B. (2005). Stat3 regulates genes common to both wound healing and cancer. *Oncogene*, 24(21):3397–3408.

Delmotte, N., Ahrens, C. H., Knief, C., Qeli, E., Koch, M., Fischer, H.-M., Vorholt, J. A., Hennecke, H., and Pessi, G. (2010). An integrated proteomics and transcriptomics reference data set provides new insights into the Bradyrhizobium japonicum bacteroid metabolism in soybean root nodules. *PROTEOMICS*, 10(7):1391–1400.

Duffy, M. J. (2004). The urokinase plasminogen activator system: role in malignancy. *Current Pharmaceutical Design*, 10(1):39–49.

Eberhardt, W., Doller, A., Akool, E.-S., and Pfeilschifter, J. (2007). Modulation of mRNA stability as a novel therapeutic approach. *Pharmacology & Therapeutics*, 114(1):56–73.

Ferreira, P. G., Jares, P., Rico, D., Gómez-López, G., Martínez-Trillos, A., Villamor, N., Ecker, S., González-Pérez, A., Knowles, D. G., Monlong, J., Johnson, R., Quesada, V., Djebali, S., Papasaikas, P., López-Guerra, M., Colomer, D., Royo, C., Cazorla, M., Pinyol, M., Clot, G., Aymerich, M., Rozman, M., Kulis, M., Tamborero, D., Gouin, A., Blanc, J., Gut, M., Gut, I., Puente, X. S., Pisano, D. G., Martin-Subero, J. I., López-Bigas, N., López-Guillermo, A., Valencia, A., López-Otín, C., Campo, E., and Guigó, R. (2014). Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Research*, 24(2):212–226.

Field, D., Sansone, S.-A., Collis, A., Booth, T., Dukes, P., Gregurick, S. K., Kennedy, K., Kolar, P., Kolker, E., Maxon, M., Millard, S., Mugabushaka, A.-M., Perrin, N., Remacle, J. E., Remington, K., Rocca-Serra, P., Taylor, C. F., Thorley, M., Tiwari, B., and Wilbanks, J. (2009). 'Omics Data Sharing. *Science*, 326(5950):234–236.

Glaab, E. (2015). Using prior knowledge from cellular pathways and molecular networks for diagnostic specimen classification. *Briefings in Bioinformatics*, page bbv044.

Gligorijević, V. and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112).

Goldhirsch, A., Wood, W. C., Gelber, R. D., Coates, A. S., Thürlimann, B., Senn, H.-J., and 10th St. Gallen conference (2007). Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, 18(7):1133–1144.

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(2):1–10.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438.

Gray, K. A., Seal, R. L., Tweedie, S., Wright, M. W., and Bruford, E. A. (2016). A review of the new HGNC gene family resource. *Human Genomics*, 10.

Greenbaum, D., Colangelo, C., Williams, K., and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology*, 4(9):117.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database issue):D140–D144.

Haider, S. and Pal, R. (2013). Integrated Analysis of Transcriptomic and Proteomic Data. *Current Genomics*, 14(2):91–110.

Hamon, J., Jennings, P., and Bois, F. Y. (2014). Systems biology modeling of omics data: effect of cyclosporine a on the Nrf2 pathway in human renal cells. *BMC Systems Biology*, 8:76.

Harwood, N. E. and Batista, F. D. (2008). New Insights into the Early Molecular Events Underlying B Cell Activation. *Immunity*, 28(5):609–619.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, 43(D1):D512–D520.

Hunter, T. (1995). Protein kinases and phosphatases: The Yin and Yang of protein phosphorylation and signaling. *Cell*, 80(2):225–236.

Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., Atauri, P. d., Aitchison, J. D., Hood, L., Siegel, A. F., and Bolouri, H. (2005). A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17296–17301.

Imielinski, M., Cha, S., Rejtar, T., Richardson, E. A., Karger, B. L., and Sgroi, D. C. (2012). Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Molecular & cellular proteomics: MCP*, 11(6):M111.014910.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(Database issue):D412–416.

Jordan, J. D., Landau, E. M., and Iyengar, R. (2000). Signaling networks: the origins of cellular multitasking. *Cell*, 103(2):193–200.

Josić, K., López, J. M., Ott, W., Shiau, L., and Bennett, M. R. (2011). Stochastic Delay Accelerates Signaling in Gene Networks. *PLOS Comput Biol*, 7(11):e1002264.

Kamburov, A., Cavill, R., Ebbels, T. M. D., Herwig, R., and Keun, H. C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, 27(20):2917–2918.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–462.

Kristensen, A. R., Gsponer, J., and Foster, L. J. (2013). Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Molecular Systems Biology*, 9:689.

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313.

Kuo, T.-C., Tian, T.-F., and Tseng, Y. J. (2013). 3omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology*, 7:64.

Kuperstein, I., Bonnet, E., Nguyen, H.-A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., Dutreix, M., Barillot, E., and Zinovyev, A. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, 4:e160.

Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R., and Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics (Oxford, England)*, 26(19):2438–2444.

Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., and Schneider, M. V. (2015). Data integration in biological research: an overview. *Journal of Biological Research*, 22(1).

Larance, M., Ahmad, Y., Kirkwood, K. J., Ly, T., and Lamond, A. I. (2013). Global subcellular characterization of protein degradation using quantitative proteomics. *Molecular & cellular proteomics: MCP*, 12(3):638–650.

Larance, M. and Lamond, A. I. (2015). Multidimensional proteomics for cell biology. *Nature Reviews Molecular Cell Biology*, 16(5):269–280.

Leek, J. T. and Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genet*, 3(9):e161.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. (2005). Assessing the limits of genomic data integration for predicting protein networks. *Genome Research*, 15(7):945–953.

Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Älgenäs, C., Lundeberg, J., Mann, M., and Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology*, 6:450.

Maiuri, P., Knezevich, A., De Marco, A., Mazza, D., Kula, A., McNally, J. G., and Marcello, A. (2011). Fast transcription rates of RNA polymerase II in human cells. *EMBO Reports*, 12(12):1280–1285.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database issue):D108–110.

McKinney, B. A. (2009). Informatics approaches for identifying biologic relationships in time-series data. *Wiley Interdisciplinary Reviews. Nanomedicine and Nanobiotechnology*, 1(1):60–68.

McRedmond, J. P., Park, S. D., Reilly, D. F., Coppinger, J. A., Maguire, P. B., Shields, D. C., and Fitzgerald, D. J. (2004). Integration of Proteomics and Genomics in Platelets A PROFILE OF PLATELET PROTEINS AND PLATELET-SPECIFIC GENES. *Molecular & Cellular Proteomics*, 3(2):133–144.

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D'Eustachio, P., and Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, 4(4):1180–1211.

Mizuno, S., Iijima, R., Ogishima, S., Kikuchi, M., Matsuoka, Y., Ghosh, S., Miyamoto, T., Miyashita, A., Kuwano, R., and Tanaka, H. (2012). AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC systems biology*, 6:52.

Monroe, J. G. (2006). ITAM-mediated tonic signalling through pre-BCR and BCR complexes. *Nature Reviews Immunology*, 6(4):283–294.

Murphy, K. and Mian, S. (1999). Modellling Gene Expression Data using Dynamic Bayesian Networks. Technical report, Computer Science Division, University of California, BErkeley, CA.

Nariai, N., Kim, S., Imoto, S., and Miyano, S. (2004). Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 336–347.

Nie, L., Wu, G., Brockman, F. J., and Zhang, W. (2006a). Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins. *Bioinformatics*, 22(13):1641–1647.

Nie, L., Wu, G., and Zhang, W. (2006b). Correlation of mRNA expression and protein abundance affected by multiple sequence features related to translational efficiency in Desulfovibrio vulgaris: a quantitative analysis. *Genetics*, 174(4):2229–2243.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, 2(3):117–120.

Nyati, M. K., Morgan, M. A., Feng, F. Y., and Lawrence, T. S. (2006). Integration of EGFR inhibitors with radiochemotherapy. *Nature Reviews. Cancer*, 6(11):876–885.

Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular Systems Biology*, 1:2005.0010.

Park, O. K., Schaefer, T. S., and Nathans, D. (1996). In vitro activation of Stat3 by epidermal growth factor receptor kinase. *Proceedings of the National Academy of Sciences*, 93(24):13704–13708.

Perco, P., Mühlberger, I., Mayer, G., Oberbauer, R., Lukas, A., and Mayer, B. (2010). Linking transcriptomic and proteomic data on the level of protein interaction networks. *ELECTROPHORESIS*, 31(11):1780–1789.

Piruzian, E., Bruskin, S., Ishkin, A., Abdeev, R., Moshkovskii, S., Melnik, S., Nikolsky, Y., and Nikolskaya, T. (2010). Integrated network analysis of transcriptomic and proteomic data in psoriasis. *BMC Systems Biology*, 4:41.

Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M. I., Jiang, S., McCallum, A., Kirov, S., and Wasserman, W. W. (2009). The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Research*, 37(Database issue):D54–D60.

Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., Stojmirovic, A., Dobrin, R., Braxenthaler, M., Kuentzer, J., Demchak, B., and Ideker, T. (2015). NDEx, the Network Data Exchange. *Cell Systems*, 1(4):302–305.

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I., and Regev, A. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology*, 29(5):436–442.

Rau, A., Jaffrézic, F., Foulley, J.-L., and Doerge, R. W. (2010). An Empirical Bayesian Method for Estimating Biological Networks from Temporal Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Reddy, R. J., Gajadhar, A. S., Swenson, E. J., Rothenberg, D. A., Curran, T. G., and White, F. M. (2016). Early signaling dynamics of the epidermal growth factor receptor. *Proceedings of the National Academy of Sciences*, 113(11):3114–3119.

Reuter, J., Spacek, D. V., and Snyder, M. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4):586–597.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97.

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., and Wiley, H. S. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, 24(24):2894–2900.

Rolland, D., Basrur, V., Conlon, K., Wolfe, T., Fermin, D., Nesvizhskii, A. I., Lim, M. S., and Elenitoba-Johnson, K. S. (2014). Global Phosphoproteomic Profiling Reveals Distinct Signatures in B-Cell Non-Hodgkin Lymphomas. *The American Journal of Pathology*, 184(5):1331–1342.

Satpathy, S., Wagner, S. A., Beli, P., Gupta, R., Kristiansen, T. A., Malinova, D., Francavilla, C., Tolar, P., Bishop, G. A., Hostager, B. S., and Choudhary, C. (2015). Systems-wide analysis of BCR signalosomes and downstream phosphorylation and ubiquitylation. *Molecular Systems Biology*, 11(6).

Saturnino, G. B., Godinho, C. P. d. S., Fagundes-Lima, D., Silva, A. C. e., and Weber, G. (2014). Detection of construction biases in biological databases: the case of miRBase. *arXiv:1407.6570 [q-bio]*. arXiv: 1407.6570.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Research*, 37(suppl 1):D674–D679.

Scharenberg, A. M., Humphries, L. A., and Rawlings, D. J. (2007). Calcium signalling and cell-fate choice in B cells. *Nature Reviews. Immunology*, 7(10):778–789.

Schoenberg, D. R. and Maquat, L. E. (2012). Regulation of cytoplasmic mRNA decay. *Nature Reviews. Genetics*, 13(4):246–259.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

Singh, J. and Padgett, R. A. (2009). Rates of in situ transcription and splicing in large human genes. *Nature Structural & Molecular Biology*, 16(11):1128–1133.

Srivastava, V., Obudulu, O., Bygdell, J., Löfstedt, T., Rydén, P., Nilsson, R., Ahnlund, M., Johansson, A., Jonsson, P., Freyhult, E., Qvarnström, J., Karlsson, J., Melzer, M., Moritz, T., Trygg, J., Hvidsten, T. R., and Wingsle, G. (2013). OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants. *BMC Genomics*, 14:893.

Sun, H., Wang, H., Zhu, R., Tang, K., Gong, Q., Cui, J., Cao, Z., and Liu, Q. (2014). iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics (Oxford, England)*, 30(5):737–739.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J., and von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database issue):D447–452.

Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P.-A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., Seymour, S. L., Souda, P., Tsugita, A., Vandekerckhove, J., Vondriska, T. M., Whitelegge, J. P., Wilkins, M. R., Xenarios, I., Yates, J. R., and Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, 25(8):887–893.

Torres-García, W., Brown, S. D., Johnson, R. H., Zhang, W., Runger, G. C., and Meldrum, D. R. (2011). Integrative analysis of transcriptomic and proteomic data of Shewanella oneidensis: missing value imputation using temporal datasets. *Molecular bioSystems*, 7(4):1093–1104.

Torres-García, W., Zhang, W., Runger, G. C., Johnson, R. H., and Meldrum, D. R. (2009). Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics*, 25(15):1905–1914.

Tullai, J. W., Schaffer, M. E., Mullenbrock, S., Sholder, G., Kasif, S., and Cooper, G. M. (2007). Immediate-Early and Delayed Primary Response Genes Are Distinct in Function and Genomic Architecture. *Journal of Biological Chemistry*, 282(33):23981–23995.

Tuncbag, N., McCallum, S., Huang, S.-S. C., and Fraenkel, E. (2012). SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. *Nucleic Acids Research*, 40(Web Server issue):W505–509.

Wada, Y., Ohta, Y., Xu, M., Tsutsumi, S., Minami, T., Inoue, K., Komura, D., Kitakami, J., Oshida, N., Papantonis, A., Izumi, A., Kobayashi, M., Meguro, H., Kanki, Y., Mimura, I., Yamamoto, K., Mataki, C., Hamakubo, T., Shirahige, K., Aburatani, H., Kimura, H., Kodama, T., Cook, P. R., and Ihara, S. (2009). A wave of nascent transcription on activated human genes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(43):18357–18361.

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

Ward, M. D. and Leahy, D. J. (2015). Kinase activator-receiver preference in ErbB heterodimers is determined by intracellular regions and is not coupled to extracellular asymmetry. *The Journal of Biological Chemistry*, 290(3):1570–1579.

Werhli, A. V. and Husmeier, D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, 9(4):326–332.

Wu, G., Zhu, L., Dent, J. E., and Nardini, C. (2010). A comprehensive molecular interaction map for rheumatoid arthritis. *PloS One*, 5(4):e10137.

Xuan, N. V., Chetty, M., Coppel, R., and Wangikar, P. P. (2012). Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network. *BMC bioinformatics*, 13:131.

Yarden, Y. and Sliwkowski, M. X. (2001). Untangling the ErbB signalling network. *Nature Reviews Molecular Cell Biology*, 2(2):127–137.

Young, R. M. and Staudt, L. M. (2013). Targeting pathological B cell receptor signalling in lymphoid malignancies. *Nature Reviews. Drug Discovery*, 12(3):229–243.

Yunger, S., Rosenfeld, L., Garini, Y., and Shav-Tal, Y. (2010). Single-allele analysis of transcription kinetics in living mammalian cells. *Nature Methods*, 7(8):631–633.

Yuseff, M.-I., Pierobon, P., Reversat, A., and Lennon-Duménil, A.-M. (2013). How B cells capture, process and present antigens: a crucial role for cell polarity. *Nature Reviews Immunology*, 13(7):475–486.

Zhang, Y., Deng, Z., Jiang, H., and Jia, P. (2007). Inferring Gene Regulatory Networks from Multiple Data Sources Via a Dynamic Bayesian Network with Structural EM. In Cohen-Boulakia, S. and Tannen, V., editors, *Data Integration in the Life Sciences*, number

4544 in Lecture Notes in Computer Science, pages 204–214. Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-73255-6_17.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE*, 9(1):e78644.

Herewith I declare, that I prepared this PhD thesis on my own and with no other sources and aids than quoted.

Göttingen, January 26, 2017 Astrid Wachter