Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

Maria Secrier

born in Radauti, Romania

Oral examination: June 5th, 2013

Visualization and analysis strategies for dynamic gene-phenotype relationships and their biological interpretation

Referees: Prof. Dr. Lars Steinmetz
Prof. Dr. Robert Russell

I dedicate this thesis to my parents, for their continuous support and encouragement.

Dedic această teză părinților mei, care au fost mereu alături de mine și m-au susținut.

Acknowledgements

First and foremost, I would like to thank my main supervisor, Dr. Reinhard Schneider, for giving me the opportunity to do research in his lab and for his mentorship throughout my PhD. I am grateful for the skills I have gained under his supervision and for the scientific and networking opportunities that were created.

I would like to thank Dr. Wolfgang Huber for accepting to take the role of my second supervisor and for his guidance throughout the last stage of my PhD. I have received a lot of valuable advice from him and benefited greatly from interactions with members in his group. I am grateful for this opportunity to broaden my knowledge in the field of statistics.

I would like to thank the members of my Thesis Advisory Committee, Prof. Dr. Lars Steinmetz, Prof. Dr. Rob Russell and Dr. Toby Gibson, for attending the annual meetings and for the useful feedback they have provided. I also thank Dr. Kiran Patil and Prof. Dr. Stefan Wiemann for accepting to be part of my defense committee.

A big thanks goes to Venkata Satagopam, for being a great colleague and friend. He was always ready to help with his vast expertise or a kind word. I thank Dr. Sean O'Donoghue for his helpful visualization-related advice, his entertaining and informative stories and his inspiring attitude towards science and life. I thank the rest of the group (especially Salvador Santiago, Afshin Khan, Janos Binder and Carlos Villacorta Martin) for making the work environment so pleasant, open and entertaining. I feel I have learnt a lot from each of them.

I am grateful to all members of the Huber group for their constructive feedback during the group meetings, as well as the witty and enjoyable informal gatherings. In particular, I would like to acknowledge Simon Anders, Bernd Fischer and Bernd Klaus for their statistical advice.

I would like to thank Jean-Karim Heriché for the helpful discussions and for

providing the data I have used for the biggest part of my project. Many thanks go to Norman Davey for being an excellent collaborator and for sharing his vast knowledge about linear motifs with me. I would also like to acknowledge Holger Dinkel and Bora Uyar for their contributions in this collaboration and for the kind and useful discussions. I am grateful to Aidan Budd for inspiring conversations, for his encouraging words and willingness to share his native speaker skills by correcting parts of my paper.

A special thanks goes to my friends in the EMBL community, who have made the stay here an experience I will always cherish. I thank them for their support, humor, scientific knowledge, artistic contributions, and the many memorable moments we have shared.

Last, but not least, I would like to thank my family for their continuous support throughout my years of study and Claudio for his encouragement, help and patience and for the wonderful times we spent together.

Summary

The complexity of biological systems is one of their most fascinating and, at the same time, most cryptic aspects. Despite the progress of technology that has enabled measuring biological parameters at deeper levels of detail in time and space, the ability to decipher meaning from these large amounts of heterogeneous data is limited. In order to address this challenge, both analysis and visualization strategies need to be adapted to handle this complexity.

At system-wide level, we are still limited in our ability to infer genetic and environmental causes of disease, or consistently compare and link phenotypes. Moreover, despite the increasing availability of time-resolved experiments, the temporal context is often lost. In my thesis, I explored a series of analysis and visualization strategies to compare and connect dynamic phenotypic outcomes of cellular perturbations in a genetic and network context.

More specifically, in the first part of my thesis, I focused on the cell cycle as one of the best examples of a complex, highly dynamic process. I applied analysis and data integration methods to investigate phenotypes derived from cell division failure. I examined how such phenotypes may arise as a result of perturbations in the underlying network. To this purpose, I investigated the role of short structural elements at binding interfaces of proteins, called linear motifs, in shaping the cell division network. I assessed their association to different phenotypes, in the context of local perturbations and of disease.

This analysis enabled a more detailed understanding of the regulatory mechanisms beyond the malfunctioning of cell division processes, but the ability to compare phenotypes and track their evolution was limited. Exploring large-scale, time-resolved phenotypic screens is still a bottleneck, especially in the visualization area. To help address this question, in the subsequent parts of the thesis I proposed novel visualization approaches that would leverage pattern discovery in such heterogeneous, dynamic datasets and enable the generation of new hypotheses.

First, I extended an existing visualization tool, Arena3D, to enable the comparison of phenotypes in a genetic and network context. I used this tool to continue the exploration of phenotype-wide differences between outcomes of gene function suppression within mitosis. I also applied it to an investigation of systemic changes in the network of embryonic stem cell fate determinants upon downregulation of the pluripotency factor Nanog.

Second, time-resolved tracking of phenotypes opens up new possibilities in exploring how genetic and phenotypic connections evolve through time, an aspect that is largely missing in the visualization area. I developed a novel visualization approach that uses 2D/3D projections to enable the discovery of genetic determinants linking phenotypes through time. I used the resulting tool, PhenoTimer, to investigate the patterns of transitions between phenotypes in cell populations upon perturbation of cell division and the timing of cancer-relevant transcriptional events. I showed the potential of discovering drug synergistic effects by visual mapping of similarities in their mechanisms of action. Overall, these approaches help clarify aspects of the consequences of cell division failure and provide general visualization frameworks that should be of interest to the wider scientific community, for use in the analysis of multidimensional phenotypic screens.

Zusammenfassung

Die Komplexität biologischer Systeme ist faszinierend und unverstanden zugleich. Trotz steter Verbesserung der Technologien, die es erlauben biologische Parameter in immer höherer Auflösung sowohl zeitlich als auch räumlich zu untersuchen, ist unsere Fähigkeit die großen Mengen dieser heterogenen Daten zu verstehen noch immer sehr begrenzt. Daher müssen sowohl die Analysewerkzeuge als auch die Visualisierungsinstrumente verbessert werden um diese Komplexität verarbeiten zu können. Auf systemischer Ebene sind wir immer noch begrenzt was unsere Fähigkeit angeht, genetische oder umweltbedingte Krankheitsursachen zu identifizieren und deren Phänotypen einheitlich zu vergleichen. Trotz der Verfügbarkeit von immer detaillierter zeit-aufgelösten Experimenten geht der zeitliche Kontext bei der Analyse oft verloren. In meiner Doktorarbeit untersuchte ich eine Reihe von Analyse- und Visualisierungsstrategien um die dynamischen, phänotypischen Folgen zellulärer Perturbationen im genetischen sowie Netzwerkkontext abzubilden und zu vergleichen.

Im ersten Teil meiner Arbeit konzentrierte ich mich auf den Zellzyklus, da dieser eines der am besten untersuchten Beispiele für komplexe hochdynamische Prozesse darstellt. Ich wandte Analyse- und Datenintegrationsmethoden an, um Phänotypen mit Zellteilungsfehlern zu untersuchen und analysierte durch welche Perturbationen des zugrundeliegenden Netzwerkes derartige Phänotypen auftreten können. Hierzu untersuchte ich kurze, Protein-Protein-Interaktionen vermittelnde Strukturelemente, genannt "short linear motifs" und deren Einfluss auf das Zellteilungsnetzwerk. Ich ermittelte ihre Verbindung zu unterschiedlichen Phänotypen im Kontext lokaler Pertubationen und von Krankheiten.

Die Analysen ermöglichen ein tieferes Verständnis der regulatorischen Mechanismen, die Fehlfunktionen der Zellteilungsprozesse zu Grunde liegen, jedoch gibt es nur beschränkte Möglichkeiten Phänotypen zu vergleichen sowie deren Evolution nachzuverfolgen. Große Datenmengen zeitlich aufgelöster phänotypischer Untersuchungen zu analysieren oder gar zu visualisieren ist noch immer ein ungelöstes Problem. Daher habe ich im zweiten Teil meiner Arbeit neue Visualisierungsansätze entwickelt, die solch heterogene, dynamische Datensätzen für die Mustererkennung zugänglich machen.

Hierzu erweiterte ich zunächst ein existierendes Visualisierungstool "Arena3D", um Phänotypen in einem genetischen- und Netzwerk-kontext vergleichen zu können. Ich benutzte dieses Tool um die phänotypischen Unterschiede zwischen verschiedenen Experimenten, in denen die Funktion spezifischer Gene während der Mitose unterdrückt wurde, zu untersuchen. Darüber hinaus habe ich es zur systematischen Analyse des Einflusses einer verminderten Expression des Pluripotenzfaktors Nanog auf das Differenzierungs Netzwerk embryonaler Stammzellen verwendet.

Darüber hinaus erlaubt diese Visualisierung der Zeit-aufgelösten Verfolgung von Phänotypen zu untersuchen wie genetische und phänotypische Verbindungen über die Zeit evolvieren. Ich habe ein neuartiges Visualisierungstool entwickelt, welches erlaubt mittels 2D/3D Projektionen genetische Determinanten zu finden, welche Phänotypen zeitlich verbinden. Ich wandte dieses Tool "PhenoTimer" an, um die in Zellpopulationen auftretenden übergange zwischen Phänotypen zu untersuchen, welche durch Perturbation der Zellteilung sowie krebs-relevanter Transkriptionsereignisse hervorgerufen werden. Durch das visuelle Abbilden von ähnlichkeiten der Wirkmechanismen von Medikamenten gelang es mir neuartige synergistischer Effekte von Medikamenten nachzuweisen.

Die hier vorgestellten Ansätze helfen die Folgen von Zellteilungsfehlern besser zu verstehen und stellen ein allgemeines Visualisierungsframework zur Verfügung, welches es der wissenschaftlichen Gemeinde erlaubt multidimensionale phänotypische Untersuchungen zu analysieren und zu visualisieren.

Contents

List of Abbreviations

2D	two-dimensional
3D	three-dimensional
APC/C	anaphase-promoting complex or cyclosome
cAMP	cyclic adenosine monophosphate
CDK	cyclin-dependent kinase
CLV	cleavage site
CNV	copy number variation
ESC	embryonic stem cells
GO	Gene Ontology
GUI	graphical user interface
GWAS	genome-wide association study
JOGL	Java Binding for the OpenGL API
LIG	ligand binding site
MOD	post-translational modification site
NGS	next-generation sequencing
ODE	ordinary differential equation

PCA	principal component analysis
PDB	Protein Data Bank
PMCC	Pearson product-moment correlation coefficient
PPI	protein-protein interaction
SBML	Systems Biology Markup Language
SLiM	short linear motif
SNP	single-nucleotide polymorphism
SNR	signal-to-noise ratio
SOM	self-organizing map
TRG	targeting site
UPGMA	unweighted pair group method with arithmetic mean
VRML	Virtual Reality Modeling Language

Chapter 1

Introduction

One of the most fascinating aspects of biology is the complexity of structural organization. This is seen in the details of molecular geometries, in the myriad of interactions that shape regulatory processes, as well as in the higher levels of organization: the cell, the organism, the population and the species (Novikoff, 1945). Complexity develops on a fractal scale, both spatially and temporally: it emerges at all levels of spatial organization, from molecular structures to population distributions throughout the globe, as well as in the dynamics of processes, from atomic (e.g. molecular motions) to evolutionary level (e.g. species changes). This has become more evident in the past years: technological advances have enabled us to scrutinize biological systems at an unparalleled level of detail, and this has generated a real data deluge (Howe et al., 2008). The size and heterogeneity of the data impose challenges in processing, storage, visualization and interpretation. It all stems from the inherent complexity of living organisms.

This complexity, however, builds on simple principles: emergence, modularity, nonlinearity, synchronization (Koch, 2012; Mazzocchi, 2008). Emergence refers to a hierarchical level of organization: local interactions between proteins/cells give rise to higher level organization, global patterns and novel behavior (de Haan, 2006). This is strongly liked to modularity, which denotes a division of complex processes into smaller units of execution and finds examples in periodic phenomena like circadian rhythms or the cell cycle (Saez-Rodriguez et al., 2005). Nonlinearity highlights the dynamic, evolvable nature of biological systems and their stochasticity (Mosconi et al., 2008). Synchrony is another key property that ensures precise orchestration of transcriptional, translational and interaction events in the cell (Ramakrishnan et al., 2010).

These principles of organization confer an inherent flexibility to the system, such that it can adapt to new external or internal conditions (Adami et al., 2000). Adaptability is based on a structured variability that lies at the core of the genotype-phenotype relation and renders system robustness (Toussaint and von Seelen, 2007). It was the Danish botanist, physiologist and geneticist Wilhelm Johannsen who first coined the distinction between genotype (an organism's hereditary material) and phenotype (the observable outcomes) in 1911 (Johannsen, 1911). However, the concept is much older, dating back to the time of Aristotle, who hypothesized a rather abstract inheritance path: the male would contribute the "form" and the female the "matter" to the development of offspring (Mayr, 1963). While too simplistic, his hypothesis was the first one to bring the inheritance idea into discussion. Charles Darwin's theories of evolution developed the idea further, suggesting that differences between organisms were a consequence of modifications transmitted on a hereditary line (Atallah and Larsen, 2009).

Indeed, phenotypic variation stems from evolutionary principles. Natural selection imposes pressure on the species to adapt and they do this by changing their genetic structure. However, the evaluation of an organism's fitness is done through the phenotype: only organisms with successful phenotypes get to keep and pass on the underlying genetic structure that confers this success. Therefore, phenotypes constitute a variable of organism robustness (Wagner, 2012).

Gregor Mendel's work put the basis to the field of genetics in the 19th century in a systematic analysis of the laws that govern the inheritance of single-gene traits (Mendel, 1965). Only later did the concept of "phenotype" evolve to encompass the results of combining genetic and environmental factors. In the current view, the genotype-phenotype relationship is stratified on different layers: pleiotropy (allelic variation), genetic interactions and gene-environment interactions (Greenspan, 2001). They impact not only physical traits, but also the individual susceptibility to infection and the responses to medical treatment (Sawyers, 2008). In order to understand developmental patterns and diagnose, treat or prevent complex diseases we need to have a good understanding of how complex phenotypes arise as a combination of diverse factors.

Over time, the classical understanding of the phenotype has changed, moving from a qualitative to a more quantitative view. This has also triggered a shift from macrophenotypes (denoting changes in morphology) to microphenotypes (referring to physiological outcomes and transcriptional plasticity). Microarray profiling technologies enabled this for the first time through the measurement of gene expression changes under different conditions (Lander, 1999). Phenotypes became measurable features at detailed molecular level (Nachtomy et al., 2007). Moreover, the high-throughput sequencing revolution has enabled the description of a wide array of genetic determinants for complex phenotypes, by identifying single-nucleotide polymorphisms (SNP) and copy-number variations (CNV) and performing genome-wide association studies (GWAS) (Bush and Moore, 2012; Manolio, 2010; Stankiewicz and Lupski, 2010). Rare and common variant identification complements this strategy in an attempt to build a comprehensive genetic architecture of phenotypes (Marian, 2012). Omics technologies supplement this in a layered approach (Schneider and Orchard, 2011).

All this has brought about a diversification of possibilities in testing hypotheses, but also challenges related to annotation, interpretation and error assessment (Henry et al., 2011; Xuan et al., 2012). To address these challenges, we need a better integration of phenotypic data in the genetic and network context. To this purpose, analysis methods should be complemented by visualization approaches that can connect the sources of phenotypic emergence at different levels of detail and in different biological contexts. In the following sections, I discuss how phenotypic traits can be described on the basis of regulation at different levels of biological organization in a dynamic context. I also examine how visualization can be used as an aid to uncover such regulatory links.

1.1 Moving towards a dynamic view of phenotypes

The evolvability of phenotypes emphasizes their strong dynamic component. Variations in environmental conditions or genetic background impact the phenotypic landscape. The history of phenotypic changes can have relevance in organismal development or subsequently acquired diseases (Hidalgo et al., 2009). Hence, the emergence of phenotypes is strongly time-dependent and should be regarded in this context.

Time plays a major role in regulation at all scales, from molecular to population levels, as shown in Figure 1.1. A relatively recent paradigm shift has imposed a more time-aware, dynamic view on biological systems (Przytycka et al., 2010). This new perspective enables us to study not only mechanisms of action of different enzymes, process flows, but also the development of the organisms in healthy and diseased states.

Epigenetic, transcriptional and translational processes are highly dynamic: chromatin states, mRNA and protein levels fluctuate depending on different internal and external factors (Eden et al., 2011; Ernst et al., 2011; So et al., 2011). Moreover, cellular network functions and fate decisions are governed by spatiotemporal design principles. Modularity and synchronization, as discussed previously, play a crucial role in many processes: circadian rhythms, development, metabolism etc. Networks are not hardwired, but respond dynamically depending on the input and thus different reconfigurations of the pathways will lead to different temporal profiles. This is enabled through positive and negative feedback loops (Kholodenko et al., 2010).

Transient protein interactions modulate temporal processes. The strict time regulation is especially relevant for signalling and regulatory proteins, as a mechanism to achieve fast adaptation in cases of changes in environmental conditions (Stein et al., 2009). Signal propagation is often performed through transient protein binding, and these interactions can generate different signalling profiles depending on the binding frequency or abundance. For instance, monophosphorylated kinases exhibit non-monotonous, rapidly decaying activity profiles, while dually phosphorylated kinases display long, flat plateaus of activity (Kholodenko et al., 2010). Time can act as a constraint on a system as well, like in the case of the tight regulation of *Drosophila* embryo development (Sauer et al., 1996).

Ultimately, changes at all levels, from subcellular (e.g. molecular dynamics (Durrant and McCammon, 2011)) to species-wide (e.g. mutations, genetic drift (Lenormand, 2002; Rifkin et al., 2005)), arise as evolvable properties with the goal of achieving robustness (Ciliberti et al., 2007). Deviations from such an

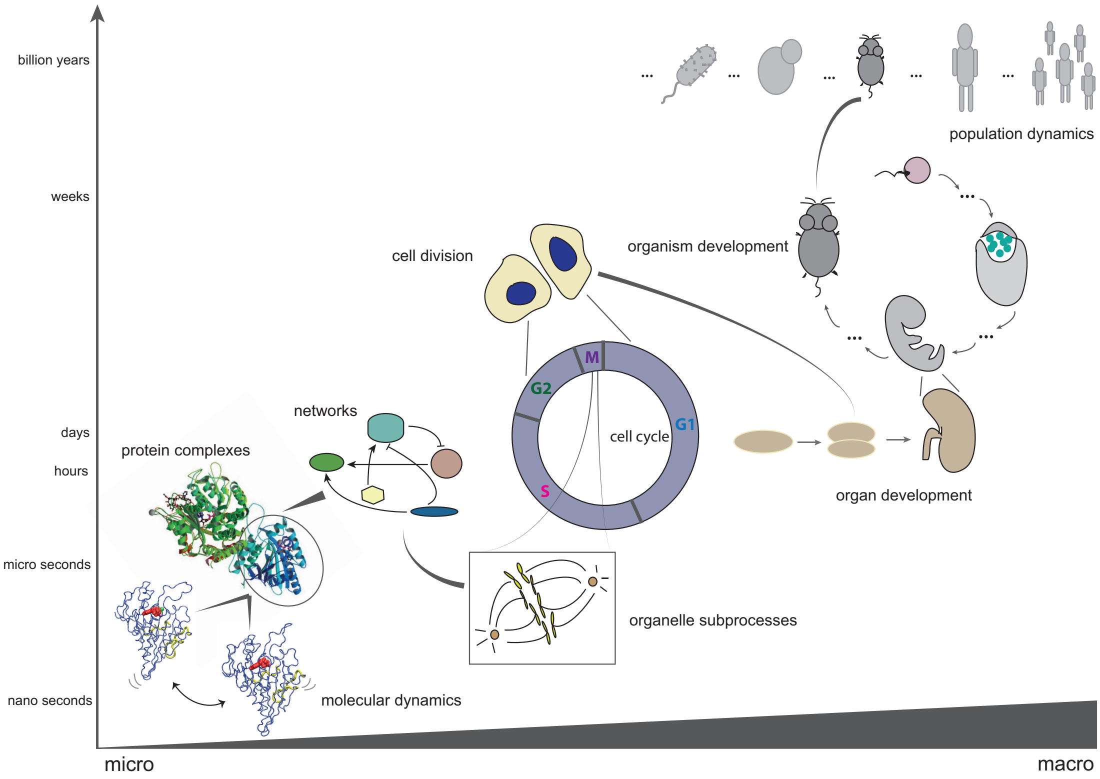


Figure 1.1: The different time scales in biology: from dynamics at the level of molecules (shown: dynamics simulations of kinesin motor protein, as obtained from the DSMM database (Finocchiaro et al., 2003)), to transient protein-protein interactions (shown: a kinesin complexed to microtubule, PDB code 2P4N), network rewiring, organelle dynamics (shown: mitotic spindle formation), periodic processes at the cellular level(shown: the cell cycle and cell division), organ and organismal development, up to population and evolutionary dynamics. Figure taken from (Secrier and Schneider, 2013).

equilibrium state have deep implications in development and disease. One of the best examples of processes where time regulation is essential is the cell cycle. I discuss it in more detail in the next subsection.

1.1.1 The cell cycle

The cell cycle is a fundamental process that governs the development, growth and heredity of living organisms, through cell reproduction (Wilson, 1987). Aberrant cell division, especially when there are defects in checkpoints, leads to accumulation of chromosomal and cellular abnormalities that translate into a wide array of disorders: cancer, cardiovascular diseases, autoimmune and metabolic disorders, viral infections, atherosclerosis, premature aging, etc. (Bicknell and Brooks, 2008; Foster, 2008; Zhivotovsky and Orrenius, 2010). Studying these defects, how and why they arise can provide a platform for finding new therapeutic or preventive targets, which is why the cell cycle has been a major subject of research for many decades (Nurse, 2000).

Temporally organized by clock-like periodicities and switch-like decisions (Tyson and Novak, 2008), the eukaryotic cell cycle is comprised of four main phases: G1 (cell growth), S (replication of the genetic material), G2 (gap phase preceding mitosis) and M (chromosome separation and division into two daughter cells) (Nurse, 2000). These events are orchestrated primarily by a family of proteins called cyclins, first discovered by searching for proteins with fluctuating levels through the cell cycle of marine invertebrates (Evans et al., 1983). They interact with cyclin-dependent kinases (CDKs) and guide the precise spatiotemporal coordination of events. CDKs are universal cell cycle regulators, conserved from yeast to mammals, as was first shown by expressing a human homologue that could rescue the function of a cdc2 mutant in fission yeast (Lee and Nurse, 1987).

Several checkpoints ensure the optimal progression of the cell cycle (Hartwell and Weinert, 1989). The DNA damage and replication checkpoints trigger signal transduction pathways that prevent the onset of mitosis if DNA repair mechanisms are active or if the DNA has not been fully replicated (Niida and Nakanishi, 2006). Complexes like Rad9-Rad1-Hus1 or Rad17RFC trigger downstream signalling of several kinases, among which Chk1 and Chk2. These in turn activate Cdc25, and tumor suppressors Wee1 and p53, that inactivate CDKs and block progression through mitosis (Bulavin et al., 2003; Mir et al., 2010; Shaw, 1996).

The spindle assembly checkpoint (SAC) monitors the association of kinetochores to microtubules and arrests mitosis in cases of improper chromosome orientation, attachment or spindle formation (May and Hardwick, 2006). This process ensures that the genetic material is properly segregated between the daughter cells. At the core of this checkpoint lie the Mad and Bub proteins, whose combined action helps maintain cohesion between sister chromatids (see Figure 1.2). At the beginning of anaphase, a caspase called securin will cleave the Scc1 cohesin subunit after the latter has been phosphorylated by Polo kinase, thereby triggering chromatid separation. The timing of this process is controlled by the

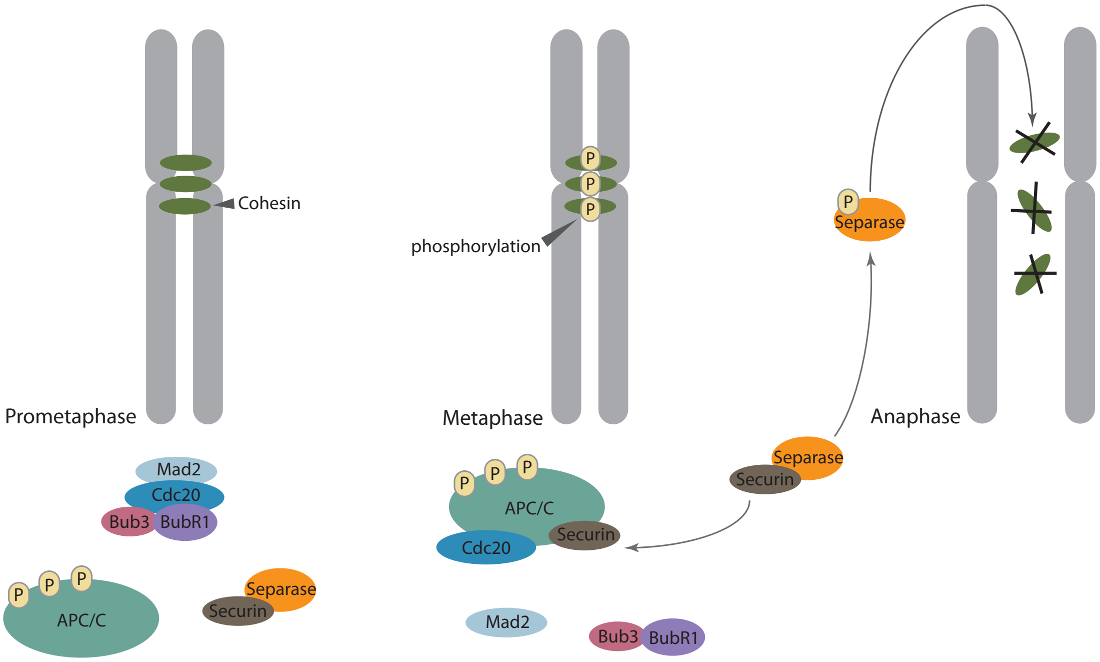


Figure 1.2: Mitotic progression steps. In prometaphase, the Mad-Bub complex prevents Cdc20 from binding the APC/C and thus keeps the cohesion between sister chromatids. In metaphase, cohesins get phoshphorylated. Cdc20 binds the APC/C and recruits securin, thus freeing separase. In anaphase, separase then gets phosphorylated and cleaves cohesins, inducing the separation of genomic material. Figure adapted after Figure 1 from (Musacchio and Hardwick, 2002).

anaphase-promoting complex (APC/C) and its accessory proteins Cdc20 and Cdh1. These target different cell cycle regulators for degradation. Cdc20 attaches to the APC/C and triggers the ubiquitylation of securin. This releases separase and allows it to cleave cohesin and initiate anaphase (Musacchio and Hardwick, 2002).

Failure of checkpoint execution in the cell cycle leads to genomic instability, loss of tumor suppressor functions, oncogene activation and structural chromosome rearrangements. All these are steps in tumor development (Kastan and Bartek, 2004; Vogelstein and Kinzler, 2004).

A lot of research, both experimental and computational, has been performed to minutely dissect the details of the different events occurring within the cell cycle. The knowledge derived from this is extremely complex, and perhaps one of the best impressions of this complexity is given by the reconstruction of the cell cycle network in the budding yeast, illustrated in Figure 1 of the paper (Kaizu et al., 2010), comprising 880 molecular interactors (genes, proteins, RNA) and 772 interactions. In this paper the authors also discuss a large array of feed-forward, inhibitory and feedback mechanisms that ensure cell cycle robustness. However, despite all this available knowledge, bridging the gap between genes and regulation, on the one hand, and disease phenotypes, on the other hand, remains a difficult task.

Neumann *et al.* have shown that defective cell division can lead to a variety of observable phenotypes, which they categorize according to their morphology (Neumann et al., 2010), as shown in Figure 1.3. These phenotypes are descriptive of errors in the regulation of mitotic subphases, from prophase to anaphase. The malfunctioning of the cell division process underlies disruptions in and potential rewiring of protein networks. The dynamics of these interactions may give us preliminary clues about the genetic determinants of some disease phenotypes. However, the influence of these interaction mechanisms on phenotypic outcomes is not clear.

In the following chapters I present different analysis and visualization approaches to leverage the understanding of how such phenotypes arise based on the genetic background, how they succeed each other in cell populations and how they relate to each other. I employ the dataset from (Neumann et al., 2010) extensively to study different aspects of dynamic regulation and how they can be captured by visualization and analysis. Besides identifying global patterns in cell cycle regulation, I also investigate how transient protein-protein interactions shape the array of cellular outcomes in the mitotic context. Since proteins and their interactions are the building blocks of complex processes like the cell cycle, exploring these interactions in more detail should give us clues into how robustness is achieved. Globular and disordered domains at the interface of protein binding enable a dynamic, yet stable, coordination of processes. I look at the role of disordered regions, particularly short functional motifs, in mediating protein interactions to render viable outcomes of cellular phenotypes. More details about these regions are described in the next section.

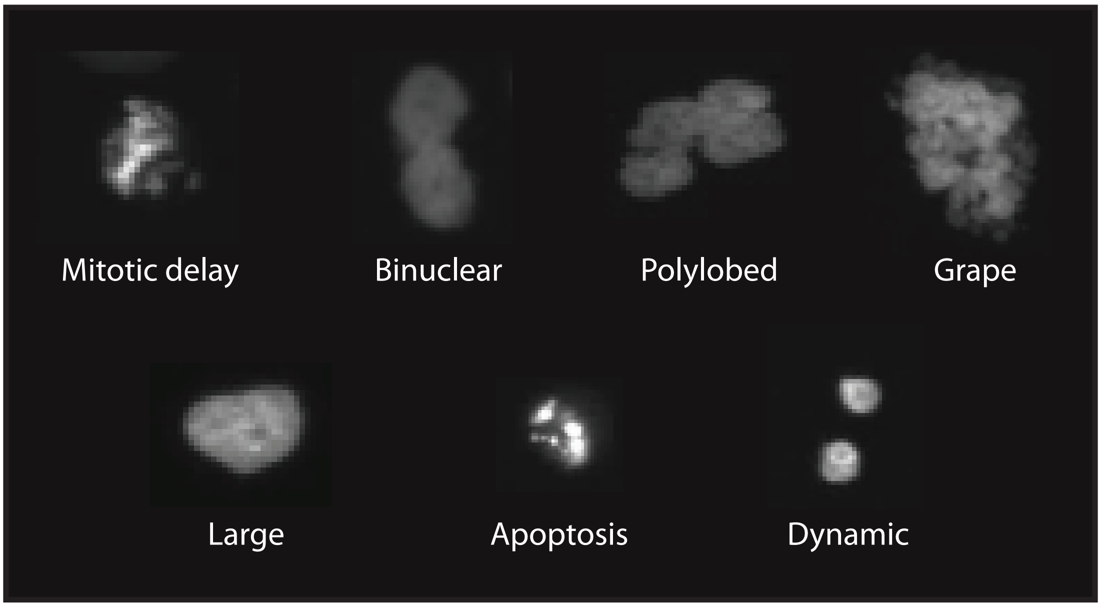


Figure 1.3: The seven main phenotypes obtained by knocking down genes essential to the cell cycle, according to (Neumann et al., 2010). Examples of defective cell morphologies are shown in every case. The images were taken from movies available in the database at http://mitocheck.org/. Mitotic delay: cells are arrested in division because of prometaphase or metaphase alignment problems. Binuclear: cytokinesis defects cause imperfect division, resulting in cells with two nuclei. Polylobed: similar cytokinesis failure gives rise to cells with multilobed nuclei. Grape: cells have multiple micronuclei. Large: division halt causes abnormally big cells. Apoptosis: cells die. Dynamic: abnormall cell division phenotypes that do not fit in any of the previous categories.

1.1.2 The role of disordered regions and post-translational modifications in network dynamics

Synergies between structured and disordered regions of proteins greatly expand the repertoire of functional flexibility and dynamics within the proteome. The modular architecture of proteins enables a lot of diversity in the interaction landscape, where many surfaces, globular or disordered, act as molecular switches. p53 is one of the best examples for this, being comprised of a single globular module and several disordered modules that sometimes overlap (Gibson, 2009). This gives it the ability to interact with many other proteins and accomplish key roles in crucial biological processes.

Starting with the solving of the first protein structure, that of myoglobin, in 1959 (Kendrew et al., 1958), it was widely believed for a long time that the structured, folded regions of proteins (e.g. alpha helices, beta sheets etc.) were necessary to accomplish function (Wright and Dyson, 1999). The structure-function paradigm has recently expanded, as it is becoming increasingly clear that intrinsic disorder is widespread in the human proteome (at least 30%) and plays a key role in shaping the interaction landscape (Mosca et al., 2012). These unstructured protein regions have been shown to arise through convergent evolution and to be more tolerant to mutations (Dunker et al., 2008). Therefore, besides enabling system dynamics, they also confer robustness to cells: disordered regions have a higher capacity of network rewiring and this confers an evolutionary advantage.

Among these disordered regions, short functional peptides termed ”linear motifs" are particularly frequent as mediators of transient interactions (Puntervoll et al., 2003). These functional regions are universally present in eukaryotic proteomes and have significant contributions in many cellular processes, due to their flexibility and evolutionary plasticity (Neduva and Russell, 2005). For instance, they contain key residues in the binding of proteins involved in signal transduction (e.g. SH3 domains bind the motif PxxP) or DNA replication (e.g. the DNA polymerase delta cofactor PCNA binds the motif QxxxxxFF) (Neduva et al., 2005). Transcriptional factors, alternative splicing and post-translational modifications regulate these regions to confer an increased versatility that is often tissue-specific. However, the lack of structure and low specificity renders these motifs highly promiscuous. They need to be tightly regulated in order to prevent disease phenotypes (Davey et al., 2012b).

As mentioned before, many of these linear motifs at the interfaces of protein binding are regulated by post-translational modifications, a crucial process through which diversity at the protein level is achieved. With more than 200 types of modification types already described and probably many still to be discovered, the view of regulation has moved far beyond the classical kinase-enabled phosphorylation that drives enzymatic activity (Deribe et al., 2010). It is now widely understood that these dynamic and most of the times reversible modifications alter the physico-chemical properties of proteins in a variety of ways and enable fast message propagation and heterogeneity of signalling.

However, the precise mechanisms by which linear motifs and post-translational modifications couple to modulate protein binding and regulation are not yet clearly understood. In the context of a dynamic process like the cell cycle, better assessment of the contribution of disorder, in general, and short binding interfaces, in particular, to maintaining organism fitness is required. I address the question of how linear motifs and post-translational modifications are linked to disease phenotypes that are cell cycle-related in the second chapter of the thesis.

1.2 The role of visualization in biology

To research complex processes like the ones previously described, both analysis and visualization tools are necessary. Analysis and visualization are invariably linked in the exploration of biological problems. While analysis gives us the mathematical, statistical or informatics methodology on which to base and verify assumptions, visualization converts observations into patterns that are easily processed by the human brain and aid interpretability. In the context of high volume and heterogeneity of data, visualization becomes essential for understanding relationships and making conclusions through abstraction and simplification.

One of the earliest examples of visualization in biology was Robert Hooke's ”Micrographia”, a 17th century book that presented drawings of microscopic organisms at an unprecedented level of detail. This was one of the first books to draw the attention of scientists as well as of the general public to the world of microbiology, and visualization played a key role in the process. The drawings reproduced observations of life forms that were not visible to the naked eye and helped disseminate the knowledge. Ever since, visualization methods have progressed from fully manual (hand-drawn) towards increased automation, with the help of advances in technology both on the experimental side (better microscopes, high-throughput techniques, sequencing etc.) and on the computational side (the invention of the computer and subsequent revolutions in data processing and visualization). Especially in the sequencing and omics area, where huge amounts of data are produced, visualization is essential to compress information and extract patterns (Gehlenborg et al., 2010). One example is the human genome: no sense could be made of the succession of ”A”s, ”T”s, ”G”s and ”C”s without visualization. We expect that improvements in graphics devices and computer power will take the visualization efforts a step further in the direction of data integration at different biological scales.

1.2.1 Visualizing time-related data in biology

With the move from a static towards a dynamic view of biological systems, visualization also gradually shifted to incorporate a time-oriented perspective. The visualization of time dates back to the classical era (Feeney, 2007) and was first invented to depict historical events. Taking inspiration from the geographical mappings of lands and territories, time visualization became a science by itself under the name of "chronography" (Rosenberg and Grafton, 2010). One of the first recorded sources is the chronography effort of the doctor, botanist and philologist Jacques Barbeu-Dubourg, who in 1753 plotted a linear chronologic history of events on a 16.5 meters-long chart (Ferguson, 1991). Later on, in 1765, Joseph Priestley introduced the use of lines to represent life spans and of dots to indicate uncertainty (Davis et al., 2010).

Using mechanical methods for both information handling and interpretation of events was promoted as a huge success in data exploration, even though it was highly controversial at the time. It may seem trivial now, but viewing time as analogous to spatial measurements could not even be conceived before the 17th century. Mechanical uniformity, i.e. division of space/time into segments of equal dimensions, was in fact one of the innovations of the 18th century and paved the way to a better structured organization in many fields (Davis et al., 2010). However, the inconsistencies in information availability for different time periods (with less events known for earlier periods of mankind) introduced a problem in accommodating a uniform representation in the cartographic space. This led to the adoption of non-linear representations and grouping of events through time (Strass, 1849).

More recently, perhaps the most famous graphical depiction of timed events is Charles Joseph Minard's flow map of Napoleon's campaign in Russia, published in 1869 (Friendly, 2002). The map is already much more complex compared to its predecessors, embedding many variables to describe events visually.

Stemming from geospatial and historical precedents, the visualization of time in biology gradually converted the historical view of events to an evolutionary perspective of life on Earth and later moved on to dissect more detailed dynamics at different scales. One of the first examples of visualizing time-related data was the ”tree of life”. First proposed by Charles Darwin in the famous book ”On the origin of species”, it depicted relationships between different biological lineages (Doolittle, 1999). This representation has changed considerably over time to incorporate new insights on the evolution of species and their taxonomic classification, but it remains widely used in phylogenetic analysis (Pavlopoulos et al., 2010). The fields of mathematics and physics have also inspired the analysis and visualization of dynamics in a series of biological processes, ranging from enzyme and substrate activities underlying biochemical reactions (Chen et al., 2010) to quantitative physiological models of entire organs (Hester et al., 2011).

Most biological visualizations rely on the main graphical elements used for depicting time series: lines, bar charts, heat maps, dendograms and layered views (as shown in Figure 1.4). Splines, contour plots, bifurcation diagrams and other more complex plots are also used depending on the case (Marwan et al., 2007).

Time-related visualization in biology is still rather limited compared to the deluge of visualization tools used for other purposes. Nevertheless, the versatility in representing time-related data accommodates different analysis approaches at different biological scales, from the molecular to the species level. Figure 1.5 depicts only a few examples of graphical methods used in visualizing timed data. Molecular dynamics visualizers rely on animations or trajectory traces to depict molecular motion. Gene expression or metabolic changes can be visualized using a combination of line charts, clustering and network embedding. Dynamics at the level of tissues, organs or populations are simulated and plotted using non-linear dynamics methods (Strogatz, 1995). Evolutionary relationships are often conveyed using multiple sequence alignment and phylogenetic tree depictions, but also circular genomic views for conservation analysis, e.g. Circos plots (Krzywinski et al., 2009).

Without going into detail about the advantages and pitfalls of these various visualization approaches, I would like to point out a few of the gaps that need to be filled in the visualization field. First, there are ongoing challenges in representing large-scale heterogeneous data. The ability to integrate different biological variables while reducing dimensionality and accounting for noise in

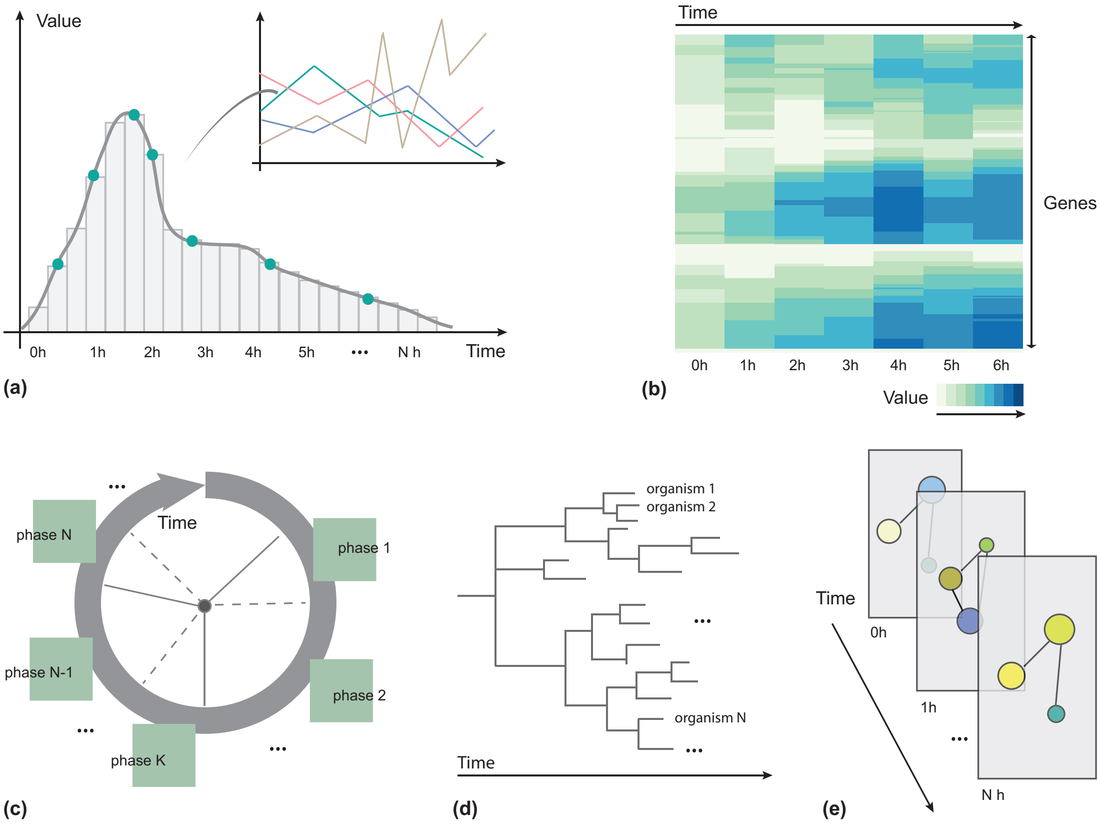
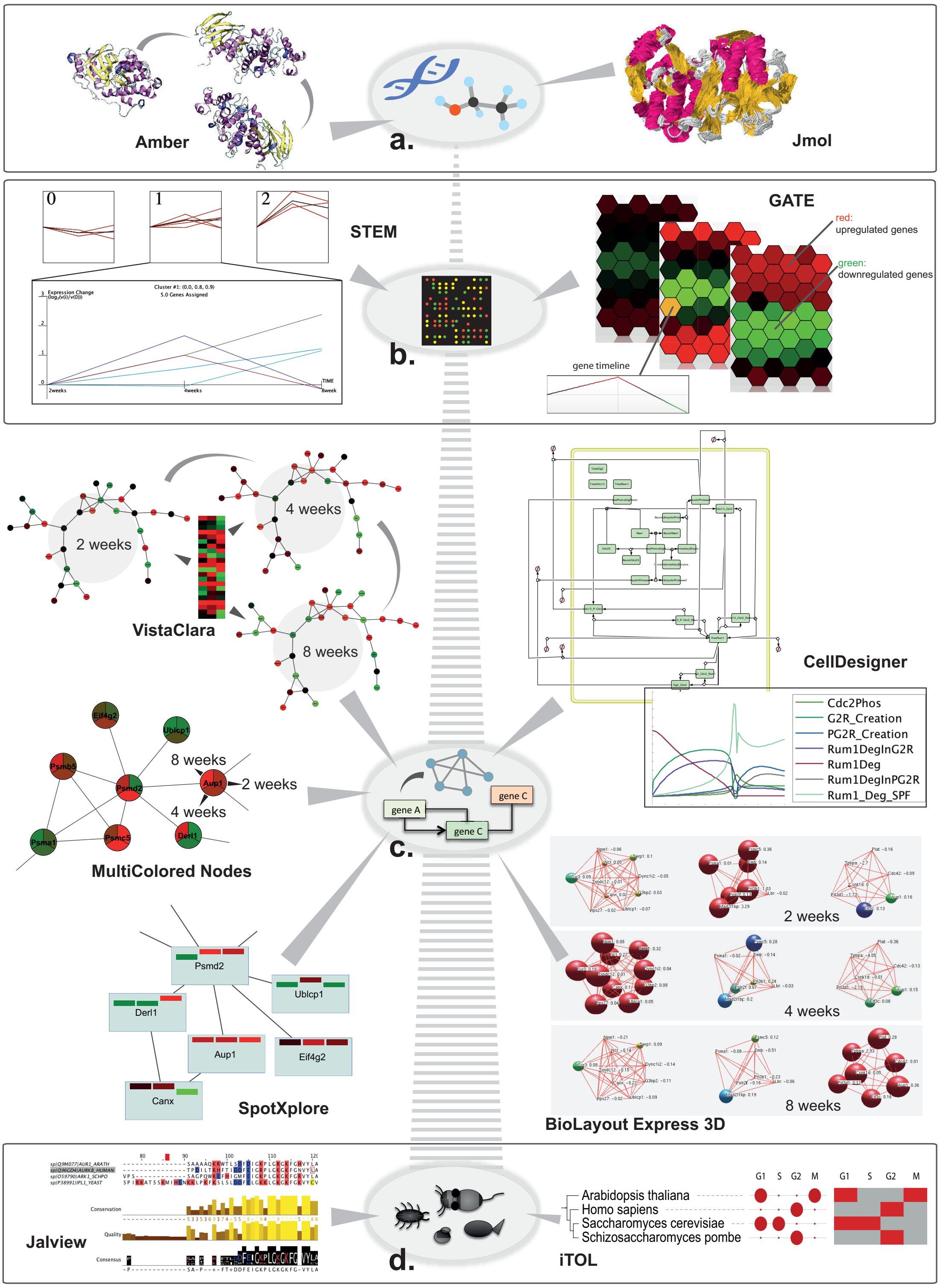


Figure 1.4: The classical representations of time in biology: (a) Linear representations of time course profiles of e.g. gene expression can be depicted individually or all together (in a parallel coordinate plot); (b) Heat maps cluster genes/proteins by their similarities in the associated time series profiles; (c) Circular views can describe divisions in recurring processes like the cell cycle; (d) Dendograms are often used to depict phylogenetic relationships and can be used to indicate the evolutionary distance between species; (e) Layered representations enable comparison of gene, network or tissue states at different time points. Figure taken from (Secrier and Schneider, 2013).

the data is limited. Related to the heterogeneity aspect, visualization tools that are able to connect different layers of information are still scarce. Second, dynamic behavior needs to be better incorporated in visual representations. What is largely missing is the ability to interpret phenotypes in a temporal context. Furthermore, comparing or linking phenotypes based on genetic determinants is not easily achievable and usually requires the use of several tools.

In the third and fourth chapters of the thesis I present different visualization approaches developed to better address some of these challenges. In particular, I focus on depicting relationships between genes and phenotypes and on the


Amber
a.
Jmol
0
1
2
STEM
Expression Change (log2(v(i)/v(0)))
Cluster #1: (0.0, 0.8, 0.9)
5.0 Genes Assigned
TIME
2weeks
4weeks
8week
b.
GATE
red: upregulated genes
green: downregulated genes
gene timeline
2 weeks
4 weeks
8 weeks
VistaClara
CellDesigner
Cdc2Phos
G2R_Creation
PG2R_Creation
Rum1DegInG2R
Rum1Deg
Rum1DegInPG2R
Rum1_Deg_SPF
Eif4g2
Ublcp1
Psmb5
Psmd2
Aup1
8 weeks
2 weeks
4 weeks
Psma1
Psmc5
Derl1
MultiColored Nodes
gene A
gene C
gene C
c.
Psmd2
Ublcp1
Derl1
Aup1
Eif4g2
Canx
SpotXplore
2 weeks
4 weeks
8 weeks
BioLayout Express 3D
Conservation
Quality
Consensus
Jalview
d.
Arabidopsis thaliana
Homo sapiens
Saccharomyces cerevisiae
Schizosaccharomyces pombe
iTOL
G1 S G2 M
G1 S G2 M

temporal context.

1.3 Aims of the PhD project

As discussed before, one of the biggest challenges in biology at the moment is the ability to manage "big data" and extract informative patterns from it. Studying gene-phenotype relationships can pave the way to better strategies for disease prevention or treatment, but untangling the complexity behind these relationships is not straightforward and is complicated by the size and heterogeneity of the data. Visualization becomes a key aspect in the analysis process, as it can alleviate these problems and bring out patterns that would otherwise be difficult to discover. However, efficient visualizations for linking and comparing phenotypes and their genetic determinants in a systematic manner are largely missing. Adding to this, the temporal component is often ignored and not properly represented. In my

Figure 1.5 *(preceding page)*: Depiction of time-related processes at different scales, along with a selection of tools that can be used for this purpose. (a) At the molecular level, tools like Amber (Case et al., 2005) or Jmol (Herraez, 2006) can be used to visualize the movements of macromolecules, as they result from molecular dynamics simulations. In this example taken from the MoDEL library (Meyer et al., 2010b), the molecular movements of the MAP kinase P38 are visualized using animations with Amber or trajectory traces with Jmol. (b) At the gene level, gene expression changes with time under different conditions can be visualized using linear depictions and clustering (e.g. with STEM (Ernst and Bar-Joseph, 2006)) or grouping along a hexagonal grid (e.g. with GATE (MacArthur et al., 2010)). The figure shows the gene expression profiles in the small intestine resulting from a high fat diet in mouse (dataset GDS3357 from Gene Expression Ominbus (Edgar et al., 2002)). (c) At the network level, changes in node color are often used to depict changes in gene expression or other network parameters. These changes can be either visualized in an animation using tools like VistaClara (Kincaid et al., 2008), in chronological segments of a pie chart using MultiColored Nodes (Warsow et al., 2010) or in bar charts embedded within the node using SpotXplore (Westenberg et al., 2010). All three are Cytoscape plugins (Shannon et al., 2003). Metabolic fluxes through pathways can be simulated and visualized using software like CellDesigner (Funahashi et al., 2008, 2003). BioLayout Express 3D (Theocharidis et al., 2009) additionally uses node size expansion to depict changes in the network in three dimensions. The data source is the same as in (b). (d) At the species level, relationships between different genes/proteins/organisms can be investigated using a variety of multiple sequence aligners (e.g. Jalview (Waterhouse et al., 2009)) and phylogenetic tree builders (e.g. iTOL (Letunic and Bork, 2007)). Shown: alignments and trees for aurora kinase B orthologs in four species. iTOL can visualize additional discs, heat maps and other charts adjacent to the dendogram (shown: phases of the cell cycle where aurora kinase B has a periodic peak of transcription, taken from Cyclebase (Gauthier et al., 2008)). Figure adapted after (Secrier and Schneider, 2013).

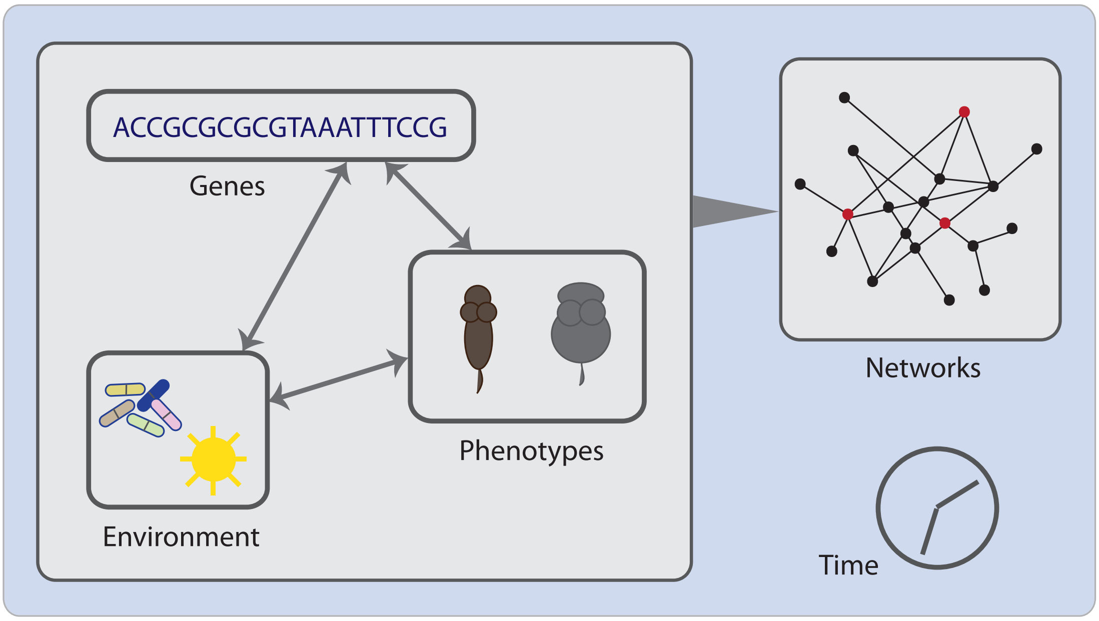


Figure 1.6: Genes, environment and phenotypes are deeply linked. The projects presented in the thesis describe approaches to investigate the connections between some of these factors, in a network context and taking into account the time component (the dynamics of biological processes).

thesis, I addressed these challenges of analyzing and visually representing complex and dynamic gene-phenotype relationships.

More specifically, I have focused on elucidating time-course phenotypic responses derived from perturbations introduced in biological systems. The purpose was to consistently link phenotypes and the underlying genetic background, looking at how phenotypic traits can evolve successively from previous traits and how networks come into play in this progression (see Figure 1.6). A major part of the work builds on experimental data that correlates gene knockdown events and phenotypes of defective cell division as described in (Neumann et al., 2010).

The initial biological question that drove the analysis efforts was the following: how can we systematically compare phenotypes of cell division defects, understand the network context in which they occur and their evolution in the cell populations? I used various strategies to answer these questions.

First, to gain a better understanding of the dynamics of regulation during cell division, I investigated how phenotypes of cell cycle malfunction can arise as a result of perturbations in the underlying network. The physical interactions between proteins essential to the cell cycle are often transient and mediated by short linear motifs. I analyzed the contribution of these linear motifs to cell cycle regulation and to the different phenotypic outcomes.

Second, to complement the analysis approach and gain further insights into comparative aspects of phenotypic emergence, I explored different visualization strategies to compare phenotypes in a network context or to link them based on genetic factors. Furthermore, I focused on time-related visualization as a special challenge in representing phenotypic outcomes. I introduce two visualization tools, Arena3D and PhenoTimer, which implement these strategies.

Hence, the aims of my PhD have been twofold: (1) to elucidate mechanistic details of regulation throughout the cell division process and reverse engineer phenotypes; and (2) to devise better visualization strategies for comparing and linking phenotypes in a dynamic regulatory context. The details of these projects are described in the following chapters.

Chapter 2

Motif-mediated interactions and their role in cell cycle phenotyping

2.1 Description

Complex processes like the cell cycle organize events in a tightly time-regulated manner to ensure the proper functioning of the organism (Silvia D. M. Santos, 2008). Failure in the regulation of these processes leads to severe developmental defects, cancer or other diseases. Since such complex phenotypes are essentially the result of disruptions in the network structure, examining the regulatory connections between proteins allows us to understand where and how rewiring occurs in cases of stress.

As illustrated in Figure 2.1, malfunctioning of a protein essential in the cell division process will disrupt a whole array of interactions that are important for cell cycle regulation and cause cell division defects. Since linear motifs are key mediators of such interactions, I wanted to investigate to what extent they play a role in this process. More specifically, I asked how disruptions of linear motif-mediated interactions might determine different phenotypic outcomes depending on the protein's motif content.

In order to answer this, I used a series of statistical, data integration and

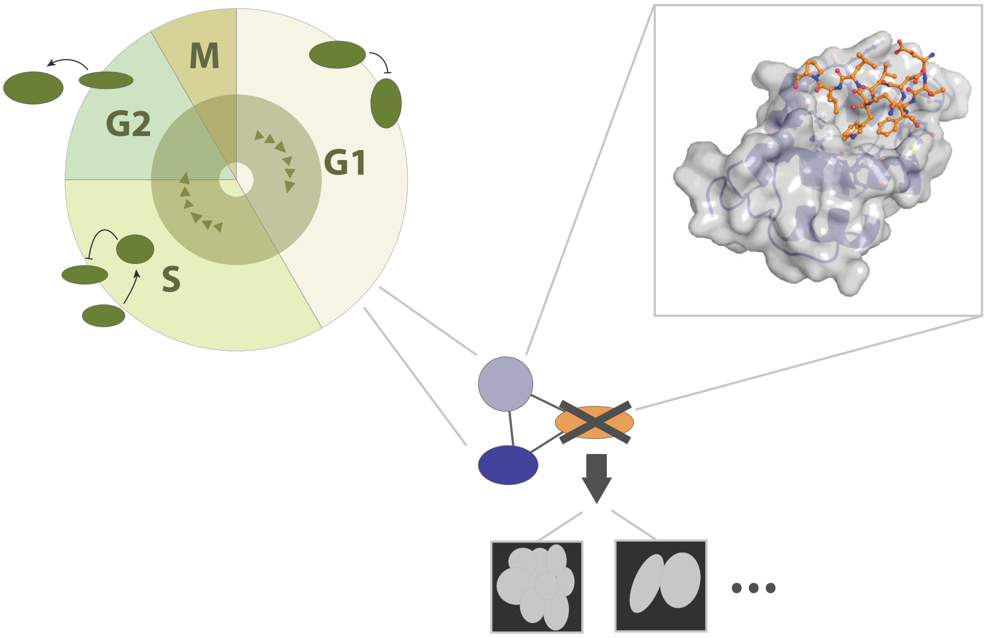


Figure 2.1: The cell cycle (upper left) is regulated through a myriad of protein-protein interactions. Some of these interactions, especially transient ones, are mediated by linear motifs. An example is shown in the upper right box: the short linear motif mediating the binding of MDM2 (purple, grey surface) to P53 (orange) undergoes an induced fit (PDB code [1YCQ], represented using PyMOL). Single gene knockdown results in the disruption of the protein network (center) and of the motif-mediated interactions, which leads to phenotypes of cell division defects observed in the cell populations (bottom).

visual inspection methods. In collaboration with the Gibson group in EMBL, we looked at differential enrichment of linear motifs in groups of proteins belonging to different phenotypic categories, as classified in (Neumann et al., 2010). We found linear motif patterns that occur significantly more often in proteins associated with a particular cell division defective phenotype. We also investigated their role as mediators of protein-protein interactions (PPIs), the post-translational modification neighbourhood of the binding interfaces and their associations with diseases, as described in the following sections.

Altogether, these findings help explain in further detail how short motifs within proteins contribute to the dynamic regulation of the cell cycle and may open paths to the discovery of new therapeutic targets.

2.2 Methods

2.2.1 Phenotypic profiling of cell division defects

For this analysis, I employed a dataset coming from a study on cell division defects obtained from single gene knockdowns, as described in (Neumann et al., 2010). A whole-genome RNA interference screening was performed in HeLa cells to discover genes essential for mitosis. Upon knockdown of such a gene, the cell populations were imaged for several hours. The succession of defective morphologies that the cells underwent was recorded. These morphologies were classified into the following phenotypic classes: ”mitotic delay”, ”binuclear”, ”polylobed”, ”grape”, ”large”, ”dynamic”, ”apoptosis”.

The time-course phenotypic data underlying the analysis performed in (Neumann et al., 2010) was supplied by Jean-Karim Heriché. For every gene that was knocked down, a vector of scores at every time point was specified for each phenotype. The scores were formulated as described in the paper and assess the penetrance of every knockdown event in the imaged cell population, taking into account several morphological features of the cells. The total number of genes in the dataset is 1067. The measurements were done every half an hour for approximately two cell cycles (i.e. 48 hours), so the total number of time points is 96. To filter phenotypes with significant scores, the following thresholds have been applied, as specified in the paper: 0.04 - ”mitotic delay”, 0.092 - ”binuclear”, 0.11 - ”polylobed”, 0.03 - ”grape”, 0.0676 - ”large”, 0.06197 - ”dynamic”, 0.072 - ”apoptosis”. For the analysis in this chapter, I take into account only the first phenotype obtained in the cell population after a gene knockdown.

2.2.2 Linear motifs

Short linear motifs (SLiMs) are defined as short peptides, generally 3-10 residues in length, that are related to a molecular function (Davey et al., 2012b). These microdomains have been found to mediate transient interactions and, as such, play an important role in many biological processes. They reside in conserved and disordered regions of proteins. The Eukaryotic Linear Motif (ELM) database (Dinkel et al., 2012) categorizes SLiMs into four main class types, based on their function: ligand binding sites (LIG), targeting sites (TRG), post-translational modification sites (MOD) and cleavage sites (CLV).

2.2.3 Enrichment of linear motifs in cell division-essential proteins by phenotype

Before performing any enrichment calculations, we needed to map Ensembl gene identifiers as provided in the dataset from (Neumann et al., 2010) to protein UniProt identifiers. This step was performed by Venkata Satagopam.

Subsequently, the enrichment was performed by Norman Davey using SLiMFinder (Davey et al., 2010) and it consisted of two steps: (1) proteins in different phenotypic classes were scanned for SLiM occurrences based on SLiM-specific regular expressions, and (2) for every SLiM class, enrichment in every phenotypic group was calculated based on comparison of its frequency within the group versus the background. The following subsections elaborate on the background choice, as well as on the subsequent filtering that was applied to reduce the number of false positives.

For motif discovery, the library of SLiM-associated regular expressions available at the ELM database was used. Motif scoring employed the methods described in (Chica et al., 2008; Davey et al., 2012a, 2010, 2011).

Importantly, the linear motifs analyzed in this chapter are predicted, rather than experimentally verified: all possible instances of SLiMs occurring in a protein are taken into account based on the characteristic sequence pattern. Therefore, all results should be regarded in the context that most SLiMs found enriched have not been experimentally validated, so there is the chance of false positives in the dataset.

2.2.3.1 Background considerations

Linear motif enrichment calculations were performed with two background datasets: (1) the rest of the targetable HeLa proteome; (2) the rest of the proteins whose disruption caused a phenotype in the experiment by (Neumann et al., 2010).

In the former case, the targetable HeLa proteome refers to all protein-coding genes and transcripts from Ensembl version 66 that were tested in the experiment, i.e. all those that could be targeted using siRNAs (irrespective of the obtained phenotype). The list comprised approximately 95% of the human protein-coding genome when considering Ensembl v66 as reference, and was provided by Jean-Karim Heriché.

In the latter case, for instance, if one wants to compute the enrichment of linear motifs for the "mitotic delay" phenotype, the background would consist of all proteins in the other phenotypic groups: "binuclear", "polylobed", "grape", "large", "dynamic" and "apoptosis".

I will term the two background categories "background 1" and "background 2", respectively, in the rest of the chapter.

2.2.3.2 Filtering enriched motifs

Initially, it was debated whether we should perform SLiM filtering based on expression levels to eliminate off-target effects. This alternative was discarded, in order to avoid that important genes that have very low expression should be eliminated.

After the enrichment analysis, I filtered out SLiMs found in extracellular regions, endoplasmic reticulum and Golgi apparatus. The motifs of extracellular proteins are less studied and many are false positives. We wanted to avoid that such SLiMs should appear as hits in intracellular proteins, since most proteins essential to the cell division process are localized in the nucleus and cytosol. This filtering measure was therefore aimed at reducing the number of false positives in the enriched SLiM dataset. The only SLiM that was eliminated from the list using this criterion was TRG_ER_KDEL_1. Since the enrichment analysis assessed whether each linear motif was overrepresented in a phenotypic group individually, the fact that I eliminated this motif after the enrichment calculations does not affect the rest of the results.

True motifs are most often found in conserved and disordered regions of proteins (Chica et al., 2008; Fuxreiter et al., 2007), so I also filtered the enriched motifs for conservation and disorder. Conservation was calculated according to a tree-based scoring method that scores the local conservation of residues in the context of constraints imposed by the adjacent regions in the protein, as described in (Davey et al., 2012a). The score varies from 0 to 1 to indicate relative conservation: 0 - conserved motif surrounded by non-conserved regions; 0.5 - the motif has exactly the same conservation as the surrounding residues; 1 - non-conserved motif surrounded by conserved regions. I selected for further analysis only motifs with a conservation score of less than 0.5.

The disorder was calculated using the prediction algorithm employed by IUPred (Dosztnyi et al., 2005). This algorithm estimates the degree of disorder in a region by calculating the interresidue interaction energy in an amino acid sequence. This value is normalized on a range from 0 to 1, corresponding to increasing degrees of disorder. I selected only motifs with a score greater than 0.3, the same threshold as described in (Davey et al., 2012a).

Only motifs that were enriched with a p-value < 0.05 were taken into account in subsequent analysis steps.

2.2.4 Linear motifs mediating protein-protein interactions

A list of pairs of proteins that were shown to interact in human was compiled from the following databases: IntAct (physical interaction or direct interaction) (Kerrien et al., 2012), MINT (yeast two-hybrid, Co-IP, pull down, affinity chromatography or affinity purification) (Ceol et al., 2010), MIPS (yeast two-hybrid, Co-IP or co-purification) (Pagel et al., 2005), STRING (experiments) (von Mering et al., 2003, 2005), BioGRID (all sources) (Chatr-aryamontri et al., 2013), DIP (all sources) (Salwinski et al., 2004), HPRD (all sources) (Keshava Prasad et al., 2009) and Reactome (direct complexes) (D'Eustachio, 2011). The list was provided by Jean-Karim Heriché.

Interactions derived from orthologous relations rather than from experiments in human cell lines were not taken into consideration because the cell cycle becomes considerably rewired in other organisms compared to the one in human, the farther away the species is from human in the taxonomic classification.

The data were downloaded from the databases in February 2012 and the following versions were used:

- MINT downloaded 06/02/2012
- BioGRID version 3.1.85
- DIP version 20111027
- HPRD release 9 (13/04/2010)
- IntAct downloaded 24/02/2012
- MIPS downloaded 24/02/2012
- Reactome downloaded 24/02/2012
- STRING v9.0

I only searched for linear motifs that were predicted to mediate the interaction between two protein partners. The list of all SLiMs predicted to mediate such interactions was supplied by Holger Dinkel. Out of these, I filtered those corresponding to the dataset of interest. The disorder filter (>0.3) was also applied.

2.2.5 Post-translational modification sites around enriched SLiMs

The post-translational modification (PTM) sites around the linear motifs in proteins associated to different phenotypes were mapped by Norman Davey from the PhosphoSitePlus (Hornbeck et al., 2012) and Phospho.ELM (Dinkel et al., 2011) databases. Regions of 10 residues before and after the motif were scanned. Subsequent selection and analysis was performed by me.

To assess the enrichment of PTM sites in different phenotypic groups, as well as of PTM-SLiM associations, I calculated the odds of a particular PTM or a PTM-SLiM instance occurring in one phenotypic group compared to all the other defined phenotypic groups. The higher the odds ratio, the more likely it is that the specific instance is overrepresented in the respective group compared to the other groups. The significance of the enrichment was calculated using the Fisher exact test, on a 95% confidence interval. Only significantly enriched categories were considered.

2.2.6 Linear motifs mutated in disease

The list of linear motifs that are mutated in different diseases, as well as their naturally occurring variants, was supplied by Bora Uyar. The mutations were scanned from the OMIM (Hamosh et al., 2005), COSMIC (Forbes et al., 2011), dbSNP (Sherry et al., 2001) and 1000 Genomes (The 1000 Genomes Project Consortium, 2010) databases. We only considered mutations that overlapped with the motifs of interest, depending on the SLiM localization in the proteins associated to different phenotypes. The results from the OMIM and COSMIC databases contained the following information: (1) protein identifier, (2) SLiM, (3) start and end position of the motif within the respective protein, (4) the position of the mutation, (5) the mutated residue and the new residue, (6) the disease in which the mutation occurs. The results from the 1000 Genomes and dbSNP databases specified naturally occurring, rather than disease-specific, mutations in the SLiMs and were used to assess natural variation of the motifs of interest.

2.2.7 Other considerations

All statistics plots presented in the following section have been produced using R. All networks were constructed using Cytoscape Shannon et al. (2003).

2.3 Results

2.3.1 SLiMs enriched in phenotypic groups

In order to investigate the existence of phenotype-specific linear motifs, we searched for motifs enriched in groups of proteins that are associated to different phenotypes. The most abundantly enriched motifs were ligand binding and PTM sites, with an average of more than three SliMs/protein and more than 20% of the proteins containing at least two classes of enriched SLiMs (see Table 2.1 for statistics of linear motif content).

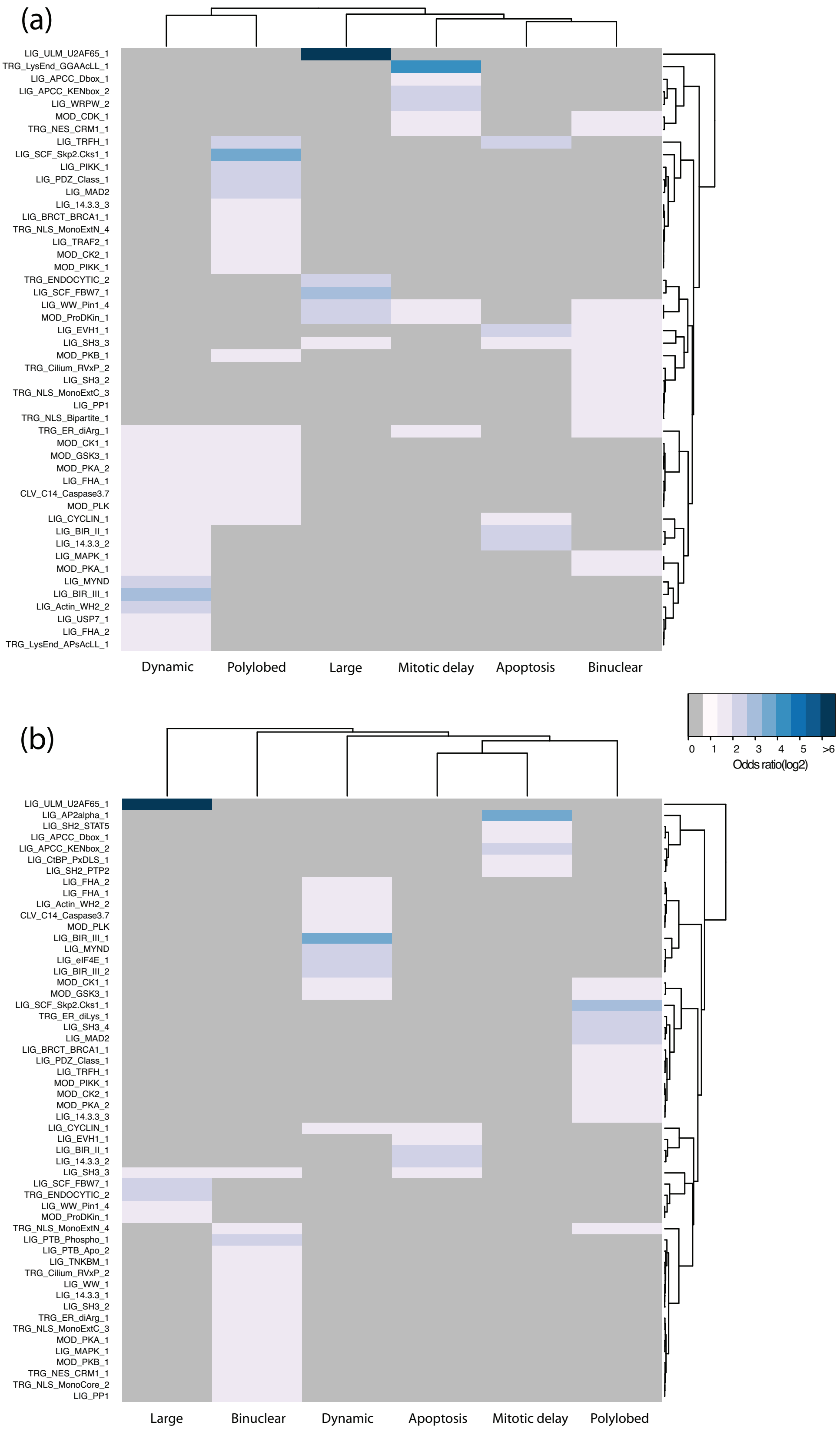
	Background 1	**Background 2**
Number of proteins	654	579
Number of SLiMs	48	54
Average enriched SLiMs/protein, by type		
Total	4.465	3.206
LIG	1.950	1.522
TRG	0.783	0.622
MOD	1.644	1.026
CLV	0.089	0.047
Proteins with		
2 types of SLiMs	38.53%	20.38%
3 types of SLiMs	6.57%	0%
Average proteins/enriched SLiM	60.83	34.37
Median proteins/enriched SLiM	33	17

Table 2.1: Enrichment calculations statistics for the two reference backgrounds.

2.3.1.1 Enrichment analysis reveals phenotype-specific motifs

After filtering according to the criteria detailed in section 2.2.3.2, background-dependent enriched groups of SLiMs were obtained, as shown in Figure 2.2. The heat maps show the odds ratio values for linear motifs enriched in different phenotypes, a higher odds ratio indicating a stronger specificity of that motif for the respective phenotypic category compared to its distribution throughout the whole proteome (Figure 2.2a) or within the other phenotypic groups of proteins (Figure 2.2b). There were no linear motifs enriched for the ”grape” phenotype in either category, so this morphology was excluded from subsequent analysis.

The linear motif content of the six phenotypic categories is more clearly defined and with less overlap when we consider background 2 for enrichment. On the one hand, this suggests that proteins whose disruption causes cell division defects might have a higher degree of similarity in terms of their linear motif composition compared to proteins not involved in cell cycle processes. On the other hand, these proteins seem to present more subtle differences in motif load, depending on the mitosis defect they associate with. It is these latter differences I am mostly interested in, since I want to be able to explain whether linear motifs play a role in differentiating phenotypic outcomes when the cell cycle malfunc-


(a)
LIG_ULM_U2AF65_1
TRG_LysEnd_GGAAcLL_1
LIG_APCC_Dbox_1
LIG_APCC_KENbox_2
LIG_WRPW_2
MOD_CDK_1
TRG_NES_CRM1_1
LIG_TRFH_1
LIG_SCF_Skp2.Cks1_1
LIG_PIKK_1
LIG_PDZ_Class_1
LIG_MAD2
LIG_14.3.3_3
LIG_BRCT_BRCA1_1
TRG_NLS_MonoExtN_4
LIG_TRAF2_1
MOD_CK2_1
MOD_PIKK_1
TRG_ENDOCYTIC_2
LIG_SCF_FBW7_1
LIG_WW_Pin1_4
MOD_ProDKin_1
LIG_EVH1_1
LIG_SH3_3
MOD_PKB_1
TRG_Cilium_RVxP_2
LIG_SH3_2
TRG_NLS_MonoExtC_3
LIG_PP1
TRG_NLS_Bipartite_1
TRG_ER_diArg_1
MOD_CK1_1
MOD_GSK3_1
MOD_PKA_2
LIG_FHA_1
CLV_C14_Caspase3.7
MOD_PLK
LIG_CYCLIN_1
LIG_BIR_II_1
LIG_14.3.3_2
LIG_MAPK_1
MOD_PKA_1
LIG_MYND
LIG_BIR_III_1
LIG_Actin_WH2_2
LIG_USP7_1
LIG_FHA_2
TRG_LysEnd_APsAcLL_1
Dynamic
Polylobed
Large
Mitotic delay
Apoptosis
Binuclear
0
1
2
3
4
5
>6
Odds ratio(log2)
(b)
LIG_ULM_U2AF65_1
LIG_AP2alpha_1
LIG_SH2_STAT5
LIG_APCC_Dbox_1
LIG_APCC_KENbox_2
LIG_CtBP_PxDLS_1
LIG_SH2_PTP2
LIG_FHA_2
LIG_FHA_1
LIG_Actin_WH2_2
CLV_C14_Caspase3.7
MOD_PLK
LIG_BIR_III_1
LIG_MYND
LIG_eIF4E_1
LIG_BIR_III_2
MOD_CK1_1
MOD_GSK3_1
LIG_SCF_Skp2.Cks1_1
TRG_ER_diLys_1
LIG_SH3_4
LIG_MAD2
LIG_BRCT_BRCA1_1
LIG_PDZ_Class_1
LIG_TRFH_1
MOD_PIKK_1
MOD_CK2_1
MOD_PKA_2
LIG_14.3.3_3
LIG_CYCLIN_1
LIG_EVH1_1
LIG_BIR_II_1
LIG_14.3.3_2
LIG_SH3_3
LIG_SCF_FBW7_1
TRG_ENDOCYTIC_2
LIG_WW_Pin1_4
MOD_ProDKin_1
TRG_NLS_MonoExtN_4
LIG_PTB_Phospho_1
LIG_PTB_Apo_2
LIG_TNKBM_1
TRG_Cilium_RVxP_2
LIG_WW_1
LIG_14.3.3_1
LIG_SH3_2
TRG_ER_diArg_1
TRG_NLS_MonoExtC_3
MOD_PKA_1
LIG_MAPK_1
MOD_PKB_1
TRG_NES_CRM1_1
TRG_NLS_MonoCore_2
LIG_PP1
Large
Binuclear
Dynamic
Apoptosis
Mitotic delay
Polylobed

tions. Thus, the results in the following sections are analyzed in the enrichment framework where the background was the set of proteins forming the other phenotypic groups (background 2).

For the "mitotic delay" phenotype, the KENbox (LIG_APCC_KENbox_2) and Dbox (LIG_APCC_Dbox_1) motifs appear enriched. These are degradation motifs recognized by the APC/C (Glotzer et al., 1991; Pfleger and Kirschner, 2000). The delay in mitosis occurs because degradation is not triggered and the cell division process is halted at the point before anaphase. The Dbox and the KENbox are consistently enriched for "mitotic delay" irrespective of the background used. This gives us higher confidence in asserting that they are true motifs specific for this phenotype. Several other SLiMs enriched for "mitotic delay" function in clathrin coat assembly, transcriptional repression and signalling pathways.

The "binuclear" category features motifs involved in signalling and nuclear localization, cAMP metabolism, responses to stress, telomere homeostasis, as well as cell cycle-related processes like protein degradation (LIG_WW_1) and exit from mitosis (LIG_PP1). Disruption of some of these motif interactions can lead to severe diseases like ciliopathies, Huntington, Alzheimer, cancer, asthma, cherubism (Berson, 1996; Guettler et al., 2011; Passani et al., 2000).

Proteins belonging to the "polylobed" phenotype are enriched in several motifs with roles in the cell cycle, either in regulating the transition from G1 to S phase (LIG_SCF_Skp2_Cks1_1), the mitotic spindle checkpoint (LIG_MAD2) or DNA damage responses (LIG_BRCT_BRCA1_1, MOD_PIKK_1, LIG_TRFH_1). Many of these motifs or their binding partners have been linked to cancer (Clapperton et al., 2004), (Nakayama and Nakayama, 1998). Others have roles in protein transport, signalling and cell growth.

The "apoptosis" phenotype presents enrichment in cell cycle motifs like LIG_CYCLIN_1, a cyclin recognition site, or LIG_BIR_II_1, a caspase suppressor that acts as an inhibitor of apoptosis. Given the central roles of different cyclins in the cell cycle, as well as the other functions linked to motifs enriched for this phenotype (G1 phase regulation, microtubule organization, cell growth, signal

Figure 2.2 *(preceding page)*: Odds ratios of SLiM enrichment, by phenotype. The enrichments were calculated with respect to: (a) Background 1; (b) Background 2.

transduction), the destructive nature of the phenotype obtained upon malfunctioning is expected.

The "dynamic" category is also enriched in the cyclin recognition motif LIG_CYCLIN_1. Besides this, there are a series of motifs related to DNA repair (LIG_FHA_1/2, MOD_CK1_1), apoptosis (LIG_BIR_III_1/2, CLV_C14_Caspase3-7), actin binding, transcription, translation and phosphorylation.

Finally, the "large" phenotype groups motifs with roles in cyclin destruction inhibition in the G1 phase (LIG_SCF_FBW7_1), checkpoints for the start of mitosis and the metaphase-anaphase transition (LIG_WW_Pin1_4), as well as signal sorting, endocytosis and cell growth. The highly enriched motif LIG_ULM_U2AF65_1 mediates interactions between splicing factors.

Frequently encountered motifs throughout the cell cycle are in the family of 14-3-3 domains. Interestingly, three such types of motifs appeared enriched, all in different categories: "binuclear", "polylobed" and "apoptosis".

Furthermore, Figure 2.2b shows quite distinct enriched motif groups for the "binuclear" and "polylobed" categories. These are two phenotypes that most of the times succeed each other in the experiment: if the cells undergo a "binuclear" morphology, they will keep dividing and eventually form an aggregate that is termed "polylobed". Considering this, one would expect similar enrichment of motifs in the two phenotypes, but here we see the results are quite distinct. The "polylobed" phenotype that is instantiated without a "binuclear" phase in the cells is therefore probably triggered by some different mechanisms of division failure compared to "binuclear". It is perhaps a different category altogether from the "polylobed" morphology observed after the "binuclear" transition.

2.3.1.2 "Binuclear", "polylobed" and "dynamic" motifs are more prevalent in the dataset

The most frequent motifs in the entire protein dataset according to the enrichment with background 2 are LIG_SH3_3 (a motif involved in several PPIs mediated by the SH3 domain), TRG_ER_diArg_1 (a membrane motif that drives localization to ER), MOD_GSK3_1 and MOD_CK_1, two phosphorylation sites widespread in vertebrates (Figure 2.3). Even though the order differs when taking the rest

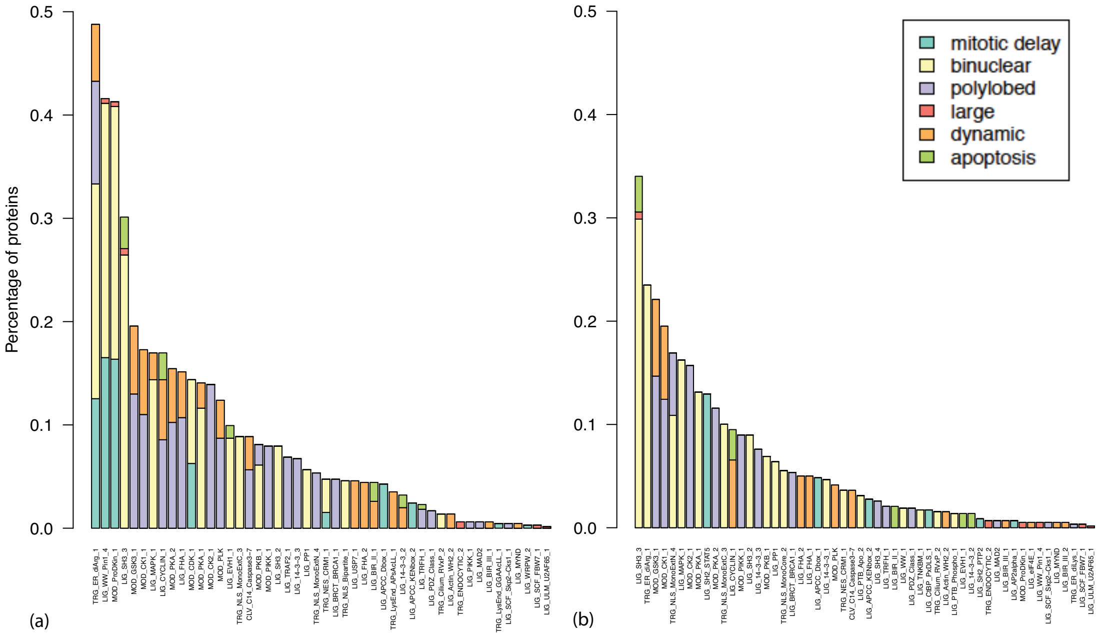


Figure 2.3: Relative protein counts with different enriched SLiMs, by phenotype, according to calculations using: (a) background 1; (b) background 2.

of the proteome as background for the enrichment, these motifs rank high in both cases. A more prevalent occurrence of motifs specific for the ”binuclear”, ”polylobed” and ”dynamic” phenotypes is observed in the case of background 2 as compared to ”binuclear” and ”mitotic delay” for background 1.

2.3.2 Linear motifs mediating protein-protein interactions

Gene knockdown events trigger a cascade of disruptions in PPIs, which leads to the malfunctioning of pathways and eventually to defects in cell division. The observable outcomes are the six phenotypes discussed, but in order to gain a better understanding of the causes of malfunction, we need to investigate which protein interactions are disrupted and how. In the human PPI network compiled from different databases, as described in subsection 2.2.4, I looked at those interactions that were affected by the knockdowns performed in the experiment. More specifically, I filtered those interactions where at least one partner's function was suppressed by the RNA interference procedure. The resulting network underwent a second filtering for interactions that are mediated by SLiMs. The outcome is shown in Figure 2.4. This network depicts the linear motifs that mediate interactions where at least one partner is associated to a phenotype.

2.3.2.1 The motif-mediated protein network contains several hubs with central roles in cell division

The linear motifs mediate a large landscape of interactions that are relevant for the phenotypic outcomes of defective cell division. This suggests that several of them might have an important role in the process. The network of motif-mediated interactions is overall well connected and some structure is evident from the distribution of the interactions and of the mediating SLiMs. Several hubs stand out in the network, and the largest are central to proteins LCK, CDC20, CDK1 , CCNB1, TRAF3 and NCF2. In most hubs, there is a single SLiM that is mediating all interactions between the central protein and all the others.

CDK1 is a key modulator of the cell cycle, promoting G1 progress, and G1-S and G2-M transitions (Fourest-Lieuvin et al., 2006). It phosphorylates the APC/C and keeps it deactivated. CCNB1 also controls the G2-M transition (Jackman et al., 2003). CDC20 is required for the activation of APC/C (Ge et al., 2009). The three hub proteins share many interaction partners, but their interactions are mediated by different types of motifs, enriched either in "binuclear" (CDK1), "mitotic delay" (CDC20) or "apoptosis" (CCNB1). All these interactions affect normal progression through the M phase and the linear motif enrichment reflects this.

PPP1CB (PP1B_HUMAN) and NEDD4L (NED4L_HUMAN) are other two smaller protein hubs where "binuclear" motif interactions converge. The former is a subunit of PP1, an essential cell division regulator that controls chromatin structure and progression in the later mitotic stages (Lee et al., 2010). NEDD4L is involved in protein degradation (Zhou and Snyder, 2005). In both cases, malfunction in anaphase leads to "binuclear" morphologies in the dividing cells.

In contrast to most other hubs, the interactions of the LCK hub are more diverse, being mediated by several linear motifs, enriched in the "large", "binuclear", "mitotic delay" or "apoptotic" phenotypes, as well as by some other mo-

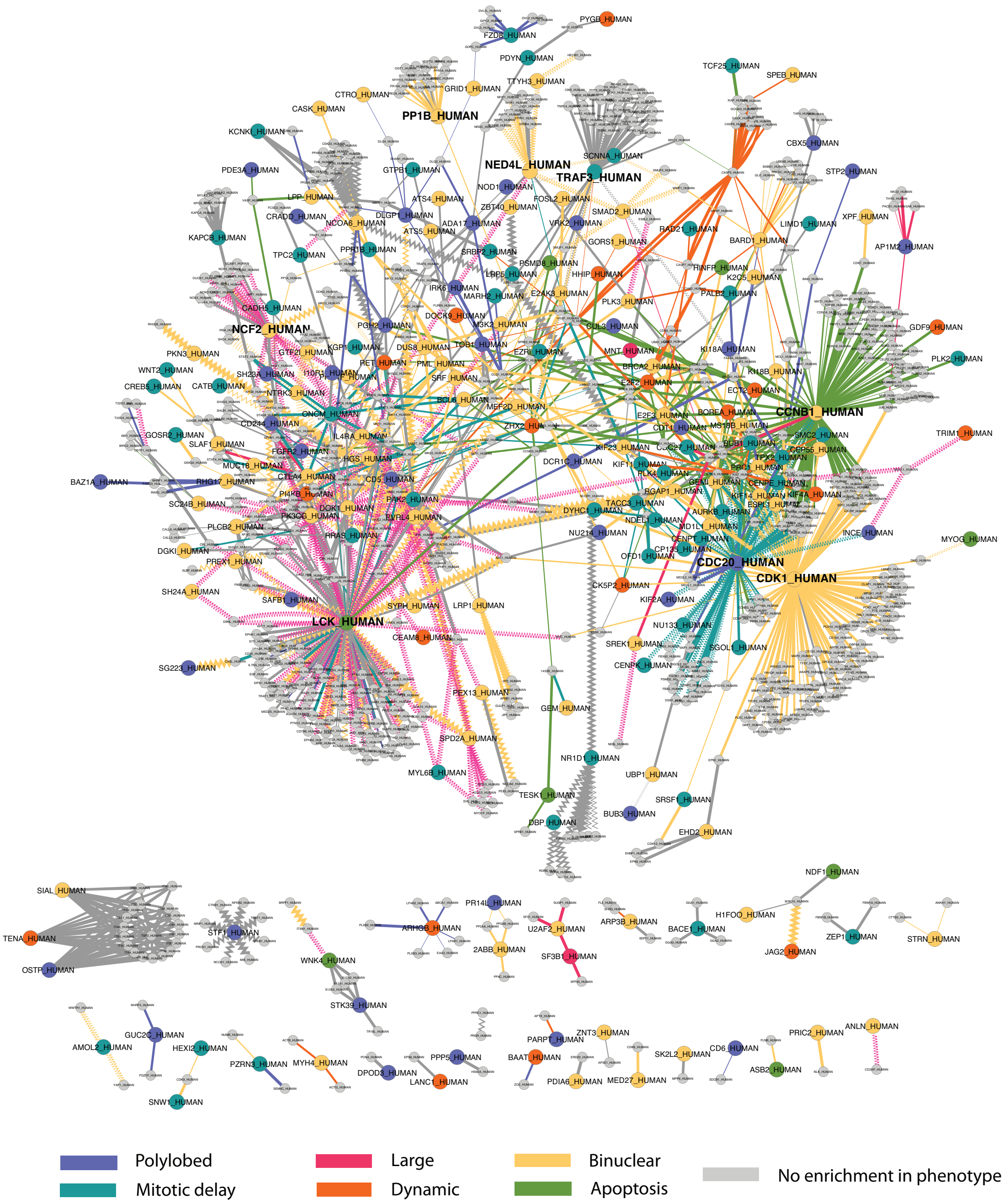


Figure 2.4: SLiMs-mediated protein interactions where at least one partner is associated to a phenotype. The nodes denote proteins and they are linked if there is evidence of an interaction between them. The color of the nodes corresponds to the phenotype obtained upon knockdown of the respective gene. The color of the links denotes the phenotype where the SLiM mediating the interaction is enriched. The SLiM classes that mediate interactions are indicated using different shape and color combinations for the links.

tifs with no enrichment in any phenotypic group. Interestingly, LCK knockdown leads to apoptosis, but the SLiMs mediating its interactions to other proteins are predominantly specific to the ”large” phenotype. This suggests that while LCK suppression is lethal for the cell, many of its interaction partners having less central role will be associated to non-lethal, albeit severe, defects in cell division. LCK is a T cell-specific protein tyrosine kinase, with roles in the maturation of developing T-cells, their proliferation and function, especially in signal transduction pathways (Palacios and Weiss, 2004). It also phosphorylates a series of microtubule-associated proteins (Scales et al., 2011), which explains its central role in the cell cycle.

Overall, the structure of the network reflects central elements in the regulation of cell division. This highlights the importance of SLiMs as mediators of interactions in essential cell cycle subpathways. Linear motifs enriched in the ”binuclear”, ”large”, ”mitotic delay” and ”apoptotic” phenotypes dominate.

2.3.2.2 Agreement of link and node enrichment varies by phenotype

Out of the interactions mediated by SLiMs that were found enriched in phenotypes, the percentage of cases when the phenotype of the SLiM agrees with the phenotype of at least one protein partner is as follows: 71.89% for ”binuclear”, 22.95% for ”polylobed”, 16.6% for ”mitotic delay”, 9.88% for dynamic and less than 2% for ”apoptosis” and ”large”. This distribution may be biased by the different sizes of the datasets, but I found that the order is kept roughly the same also after normalizing by the size of the phenotypic group. Hence, overall agreement of node- and link-associated phenotypes (i.e. having the same link and node color) is rather low in the network (22.76% of the links), with varying distributions for different phenotypes. The ”binuclear” phenotype appears to be largely consistent in proximal network participants, though.

Ostaszewski et al. have previously shown that there is a relationship between protein proximity in the network and their phenotypic similarity in the mitotic hits dataset (Ostaszewski et al., 2012). Proteins with lower network distance also had a significantly closer phenotypic distance. From the previous calculations, I could not conclude a similar relationship in the network of SLiM mediated inter-

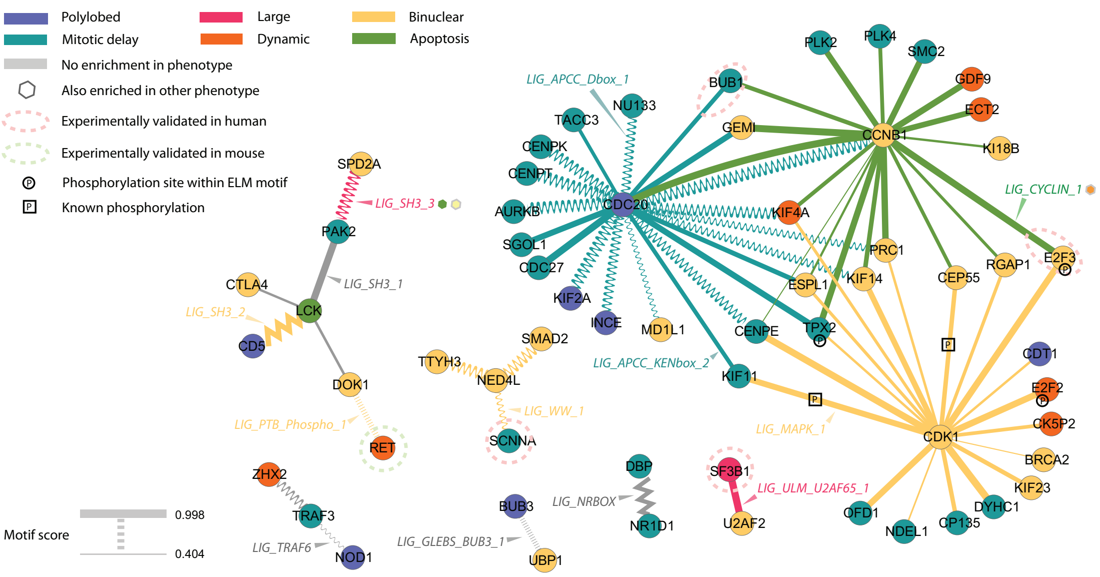


Figure 2.5: SLiMs mediating physical PPIs in proteins with a phenotypic profile. Nodes denote proteins, links denote linear motifs mediating the interaction. The color of the nodes and of the links indicates the phenotype associated to the protein or SLiM, respectively. The SLiM classes that mediate interactions are indicated using different shapes for the links, as well as in writing. The thickness of the link is proportional to the motif score. Experimentally validated motif-domain interactions are circled. Confirmed phosphorylation events from Phospho.ELM are also marked on the network.

actions. However, the observation might be confounded by the large number of proteins and interactions with no phenotype associated. Indeed, when eliminating these proteins from the network, as shown in Figure 2.5, more agreement in proximal phenotypes can be observed.

2.3.2.3 The reduced motif-mediated network is more uniform in phenotypic coverage

Focusing only on SLiM-mediated interactions where both partners are associated to a phenotype reduces the network considerably (Figure 2.5). Many of the linear motifs are enriched in the same or similar phenotypes as the proteins whose interaction they are predicted to mediate.

The three main hubs essential for cell division, CCNB1, CDK1 and CDC20 are kept. Interactions with CCNB1 are mediated by the cyclin motif LIG_CYCLIN_1. CDK1 interactions are carried out by LIG_MAPK_1 motif phosphorylation, even though phosphorylation events have been experimentally confirmed only for protein partners KIF11, CEP55 and E2F2 (according to the Phospho.ELM database).

Since most of the interactions in Figure 2.5 are only predicted, and not experimentally tested, I propose them as good candidates for future validation.

2.3.3 Post-translational modification sites around linear motifs

Protein modifications induced by post-translational regulation alter the physicochemical properties of proteins and influence their conformation and binding activity (Deribe et al., 2010). They are thus important for protein function. PTM sites around SLiMs influence the protein interaction landscape by promoting or hindering protein contact sites. Moreover, several studies have shown that some disruptions of PTM sites are linked to disease (Li et al., 2010). For instance, the Wnt/β-Catenin pathway was found to be affected both by gain and loss of phosphorylation sites in cancer (Radivojac et al., 2008). Thus, investigating the types of modifications found around SLiMs enriched in different phenotypes might give us additional clues to phenotype-specific factors of protein regulation failure and might prove relevant for therapeutic purposes.

I found that proteins belonging to the "large" phenotype had the highest abundance, on average, of modification sites around SLiMs (Figure 2.6a). The second most abundantly present were PTMs in proteins belonging to the "polylobed" group, followed by "binuclear" and "mitotic delay". Proteins in the apoptotic category contained less than two modifications around the linear motif sites.

Even though for classes like "polylobed", "binuclear" and "mitotic delay" the average number of modifications around a SLiM is similar, the types of modifications and their order differ quite a lot, as shown in Figure 2.6b. Only 19 modification patterns are shared, while "binuclear" has five fold more modifications specific only for this category.

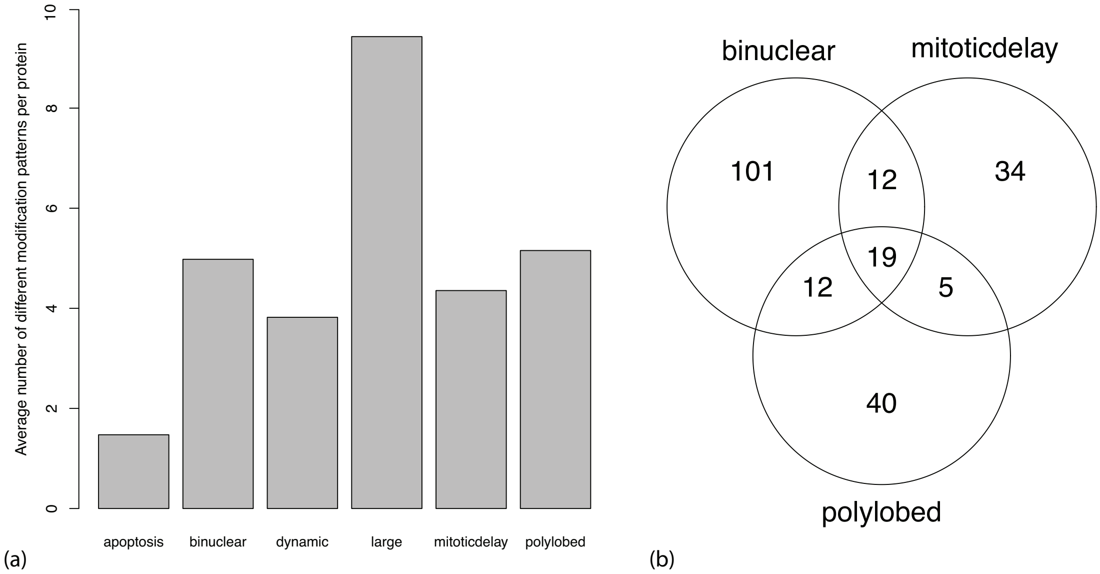


Figure 2.6: Modification site count distribution, by phenotype. (a) Average number of modification sites around linear motifs enriched in different phenotypic groups. (b) Counts of common and specific modification patterns around SLiMs enriched in the ”binuclear”, ”mitotic delay” and ”polylobed” categories.

2.3.3.1 Phosphoserine and phosphothreonine modifications are frequent in all phenotypic groups

Given these differences, I decided to further investigate the distribution of specific PTM patterns, by phenotype. The most common one in the entire dataset is a phosphoserine modification site, followed by double phosphoserine site (Figure 2.7). These are rather uniformly distributed among phenotypes, even though the former is less specific to ”mitotic delay” proteins and the latter is completely missing in ”large” ones. A phosphothreonine modification site is the third most frequent, found most prominently in the ”large” group. Figure 2.7 highlights a series of modification site patterns that are found almost exclusively in the ”large” morphology: (1) triple phosphothreonine; (2) phosphoserine followed by double phosphothreonine; (3) double phosphothreonine; (4) quintuple phosphothreonine; and a series of other patterns, mostly containing combinations of phosphoserine and -threonine sites.

The PTM class distribution is dominated in all phenotypic groups by phosphoserine and phosphothreonine sites, the latter being significantly more abundant

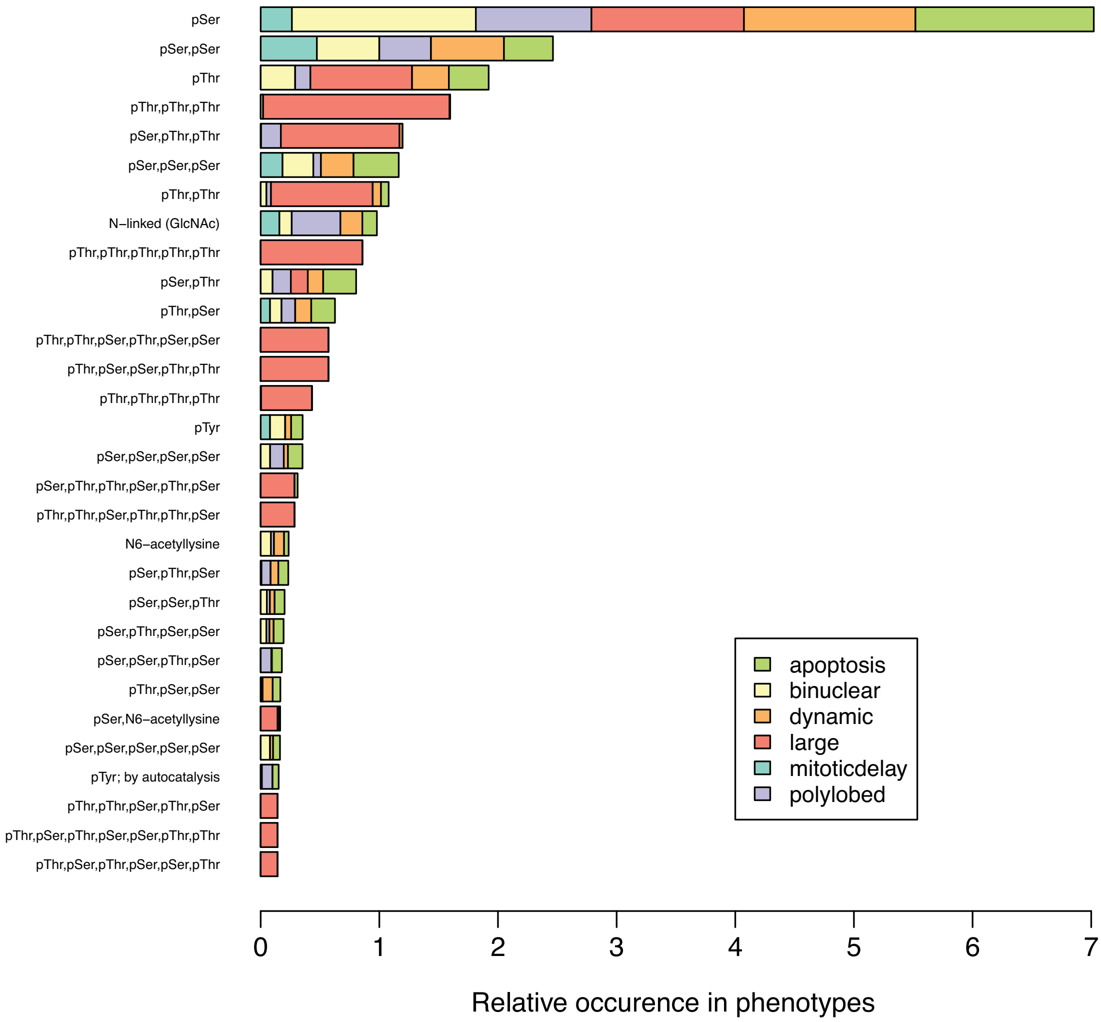


Figure 2.7: Top most abundant PTM patterns around linear motifs appearing in different phenotypic groups, relative to group size. Abbreviations: pSer = phosphoserine; pThr = phosphothreonine; pTyr = phosphotyrosine.

in the ”large” phenotype, as shown in Figure 1 of Appendix A. Phosphorylation events are most abundant overall, followed by glycosylation and acetylation events (see Figure 2 of Appendix A).

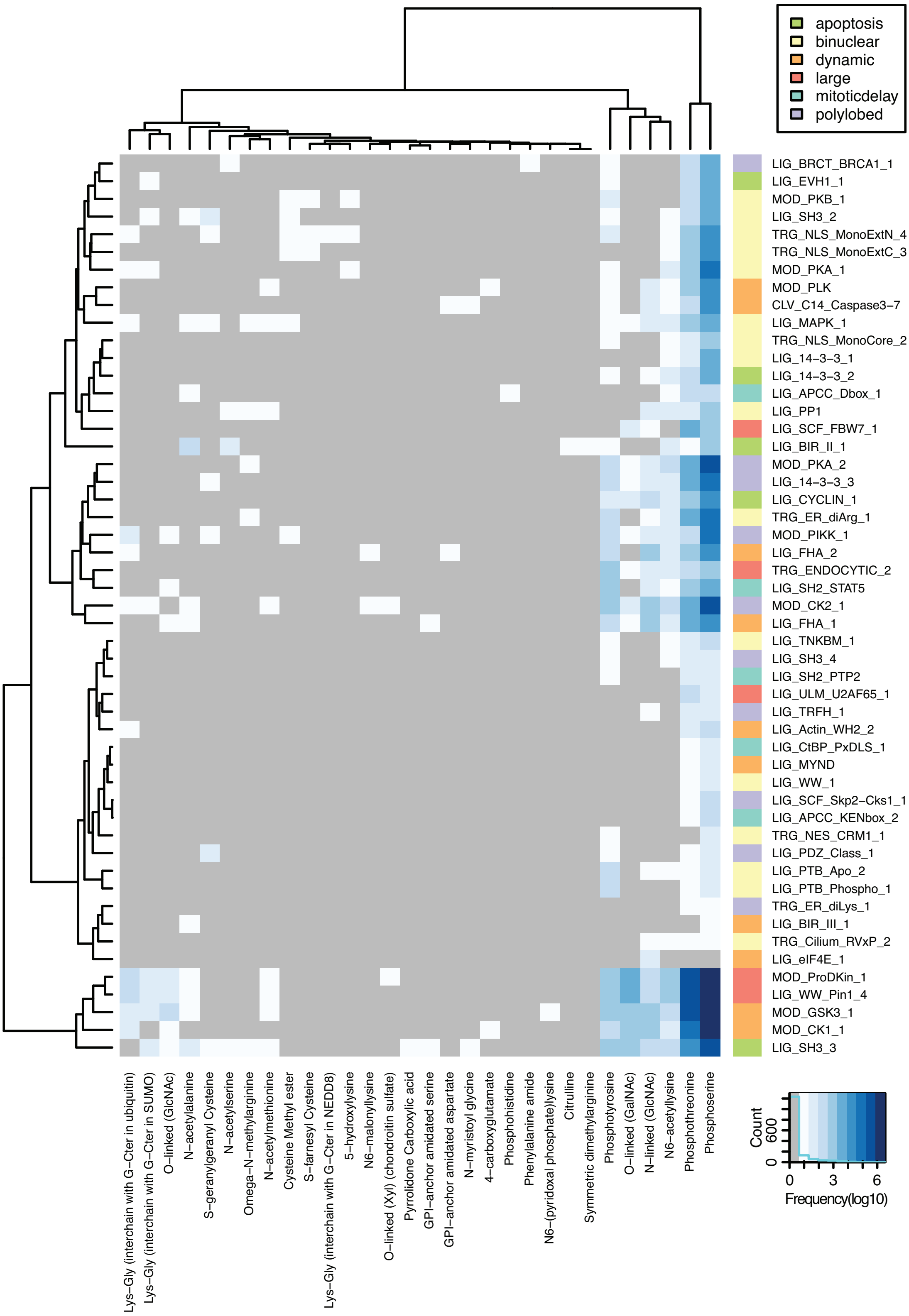
2.3.3.2 PTM class enrichment reveals distinct patterns for phenotypic groups

Calculating the enrichment of different PTM classes in phenotypic groups reveals distinct categories that are phenotype-specific (Figures 4 and 3 of Appendix A). N-linked glycosylation, appears enriched in the ”dynamic” and ”apoptosis” groups, while O-linked glycosylation is overrepresented only in ”binuclear”. N-acetylalanine, an acetylation site, appears only in ”mitotic delay”. The ”large” phenotype maintains the phosphothreonine specificity, similarly to ”polylobed”. Since the detailed patterns of PTM succession around SLiM sites are also rather distinct (Figure 4), these results suggest different mechanisms of binding among the different phenotypic groups of proteins.

2.3.3.3 SLiM-PTM associations suggest phenotype-specific regulation

Investigating SLiM-PTM coupling patterns can provide further indications about the mechanism of action of different linear motifs. For this purpose, I looked at the frequency of occurrence of these kinds of patterns for SLiMs enriched in different phenotypic groups. Figure 2.8 shows a high prevalence of phosphoserine and phosphothreonine sites around many SLiMs. In general, phosphorylation and glycosylation events are the most common, with the tightest cluster formed by the SLiMs that were previously shown to be the most frequent in the entire dataset (see also Figure 5 of Appendix A). There does not seem to be a clear clustering of these coupled occurrences by phenotype. Nevertheless, some phenotypic clusters with similar PTM patterns could be observed when performing enrichment analysis (as shown in Figure 6 of Appendix A.)

When scanning for enriched PTM classes around SLiMs, rather than specific patterns, the list becomes much shorter, as seen in Figure 2.9. Strong enrichment of phosphotyrosine sites can be observed for LIG_PTB_Phospho_1 and LIG_PTB_Apo_2, two motifs enriched in the binuclear group. This is expected, as these SLiMs are ligand sites for phosphotyrosine binding domains. LIG_BIR_II_1 stands out as the only SLiM with a strong and significant association to the N-acetylalanine class. It would be interesting to validate these strong SLiM-PTM associations by mutating the sites and observing the phenotypic outcomes in the


apoptosis
binuclear
dynamic
large
mitoticdelay
polylobed
LIG_BRCT_BRCA1_1
LIG_EVH1_1
MOD_PKB_1
LIG_SH3_2
TRG_NLS_MonoExtN_4
TRG_NLS_MonoExtC_3
MOD_PKA_1
MOD_PLK
CLV_C14_Caspase3-7
LIG_MAPK_1
TRG_NLS_MonoCore_2
LIG_14-3-3_1
LIG_14-3-3_2
LIG_APCC_Dbox_1
LIG_PP1
LIG_SCF_FBW7_1
LIG_BIR_II_1
MOD_PKA_2
LIG_14-3-3_3
LIG_CYCLIN_1
TRG_ER_diArg_1
MOD_PIKK_1
LIG_FHA_2
TRG_ENDOCYTIC_2
LIG_SH2_STAT5
MOD_CK2_1
LIG_FHA_1
LIG_TNKBM_1
LIG_SH3_4
LIG_SH2_PTP2
LIG_ULM_U2AF65_1
LIG_TRFH_1
LIG_Actin_WH2_2
LIG_CtBP_PxDLS_1
LIG_MYND
LIG_WW_1
LIG_SCF_Skp2-Cks1_1
LIG_APCC_KENbox_2
TRG_NES_CRM1_1
LIG_PDZ_Class_1
LIG_PTB_Apo_2
LIG_PTB_Phospho_1
TRG_ER_diLys_1
LIG_BIR_III_1
TRG_Cilium_RVxP_2
LIG_eIF4E_1
MOD_ProDKin_1
LIG_WW_Pin1_4
MOD_GSK3_1
MOD_CK1_1
LIG_SH3_3
Lys-Gly (interchain with G-Cter in ubiquitin)
Lys-Gly (interchain with G-Cter in SUMO)
O-linked (GlcNAc)
N-acetylalanine
S-geranylgeranyl Cysteine
N-acetylserine
Omega-N-methylarginine
N-acetylmethionine
Cysteine Methyl ester
S-farnesyl Cysteine
Lys-Gly (interchain with G-Cter in NEDD8)
5-hydroxylysine
N6-malonyllysine
O-linked (Xyl) (chondroitin sulfate)
Pyrrolidone Carboxylic acid
GPI-anchor amidated serine
GPI-anchor amidated aspartate
N-myristoyl glycine
4-carboxyglutamate
Phosphohistidine
Phenylalanine amide
N6-(pyridoxal phosphate)lysine
Citrulline
Symmetric dimethylarginine
Phosphotyrosine
O-linked (GalNAc)
N-linked (GlcNAc)
N6-acetyllysine
Phosphothreonine
Phosphoserine
Count
0
600
0
3
6
Frequency(log10)

cells, to check whether the same morphological defects are obtained.

All these results help construct a more detailed image of protein binding regulation by SLiMs and PTMs and how this regulation drives phenotypic outcomes.

2.3.4 Linear motifs mutated in diseases

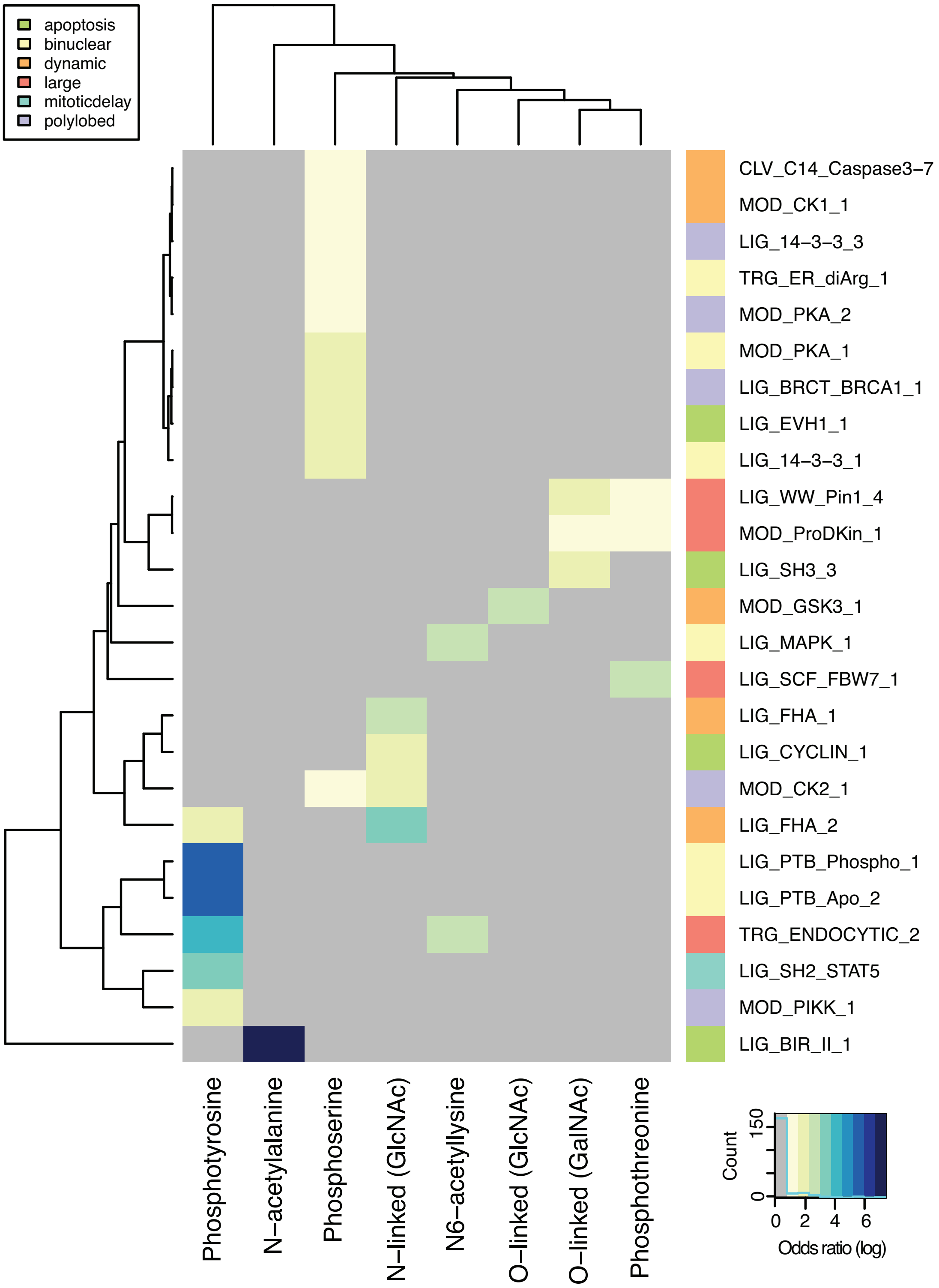
It is estimated that more than 20% of missense mutations in disease affect intrinsically disordered regions of proteins and interfere with their function, often by inducing disorder to order transitions (Vacic et al., 2012). Moreover, several SLiMs have been shown to be mutated in different diseases (Deretic et al., 1998; Eikenboom et al., 1996; Weil et al., 2003), which emphasizes their role in mediating critical PPIs. Since the morphologies observed upon mitosis malfunction may denote disease phenotypes, I investigated how these phenotypes might relate to currently classified diseases and how SLiMs play a role in the process.

2.3.4.1 Mapping SLiM-disease associations

Figure 2.10 shows the network of SLiMs enriched for different phenotypes and the diseases in which they are mutated, according to OMIM. One might expect that SLiMs enriched in the same phenotype would be mutated in the same diseases, but this does not seem to be necessarily the case. Some loose grouping by phenotype can be observed, but the network is too small to allow for generalization. Several polylobed and dynamic-related SLiMs are associated with disease.

One should also consider that the motifs shown in this network are predicted, and not experimentally validated to have functional role in the respective proteins. The fact that they are mutated in certain diseases may have relevance for the disease, but this is not implied. In this network, the only motif that has been annotated in the literature to have functional importance in disease is MOD_GSK3_1. A serine mutation at the site S165F within this motif in junctophilin-2 was shown to cause familial hypertrophic cardiomyopathy type

Figure 2.8 *(preceding page)*: PTM patterns around the sites of SLiMs enriched in different phenotypic groups. The heat map shows the logarithm of the frequency value for each PTM-SliM combination. Colored rectangles in front of the SLiM names indicate the phenotypes where these SLiMs are enriched.


apoptosis
binuclear
dynamic
large
mitoticdelay
polylobed
CLV_C14_Caspase3−7
MOD_CK1_1
LIG_14−3−3_3
TRG_ER_diArg_1
MOD_PKA_2
MOD_PKA_1
LIG_BRCT_BRCA1_1
LIG_EVH1_1
LIG_14−3−3_1
LIG_WW_Pin1_4
MOD_ProDKin_1
LIG_SH3_3
MOD_GSK3_1
LIG_MAPK_1
LIG_SCF_FBW7_1
LIG_FHA_1
LIG_CYCLIN_1
MOD_CK2_1
LIG_FHA_2
LIG_PTB_Phospho_1
LIG_PTB_Apo_2
TRG_ENDOCYTIC_2
LIG_SH2_STAT5
MOD_PIKK_1
LIG_BIR_II_1
Phosphotyrosine
N−acetylalanine
Phosphoserine
N−linked (GlcNAc)
N6−acetyllysine
O−linked (GlcNAc)
O−linked (GalNAc)
Phosphothreonine
Count
0
150
0
2
4
6
Odds ratio (log)

17 (CMH17). The mutation diminishes the phosphorylation of this protein at the respective site, which leads to perturbations of calcium signalling in skeletal muscles and marked cardiomyocyte hyperplasia (Landstrom et al., 2007). The downregulation of the junctophilin gene gives rise to a polylobed phenotype in the cell populations. The associated motif, MOD_GSK3_1, is also enriched in the polylobed phenotype.

Whether the mutations in the predicted motifs are functionally relevant in the associated diseases should be tested experimentally.

2.3.4.2 Reconstructing networks of SLiM-mediated PPIs relevant in cancer

Focusing on cancer, I investigated mutation events in the linear motif segments found in proteins in different cancer types, as recorded in the COSMIC database. The results are shown in Figure 2.11. The most frequently mutated linear motifs are MOD_GSK3_1, LIG_SH3_3 and MOD_ProDKin_1, which are also some of the most frequent motifs found overall in the proteins of the given dataset. By phenotype, the ”polylobed”, ”large” and ”dynamic” categories contain most often mutations in cancer. The least frequent cancer-related mutations occur for ”apoptosis” in linear motifs LIG_CYCLIN_1, LIG_EVH1_1, LIG_14-3-3_2 and LIG_BIR_II_1. This is not normalized by the overall frequency of these motifs in the dataset, which would eliminate some promiscuous SLiMs and is planned for the future.

The diseases most frequently associated with mutations in linear motifs from the analyzed dataset were large intestine carcinoma, skin malignant melanoma and ovary carcinoma. The high number of mutated sites in the entire dataset associated to some type of cancer (as shown in Figure 7 of Appendix A) emphasizes the potential therapeutic relevance of this study.

The frequency of mutations in cancer is less than the frequency of polymor-

Figure 2.9 *(preceding page)*: PTM classes enriched around linear motifs specific to different phenotypic groups. The heat map shows the natural logarithm of the odds ratio (increasing values on a gradient from yellow to dark blue), or 0 for no enrichment (grey tiles). Colored rectangles in front of the SLiM names indicate the phenotypes where these SLiMs are enriched.

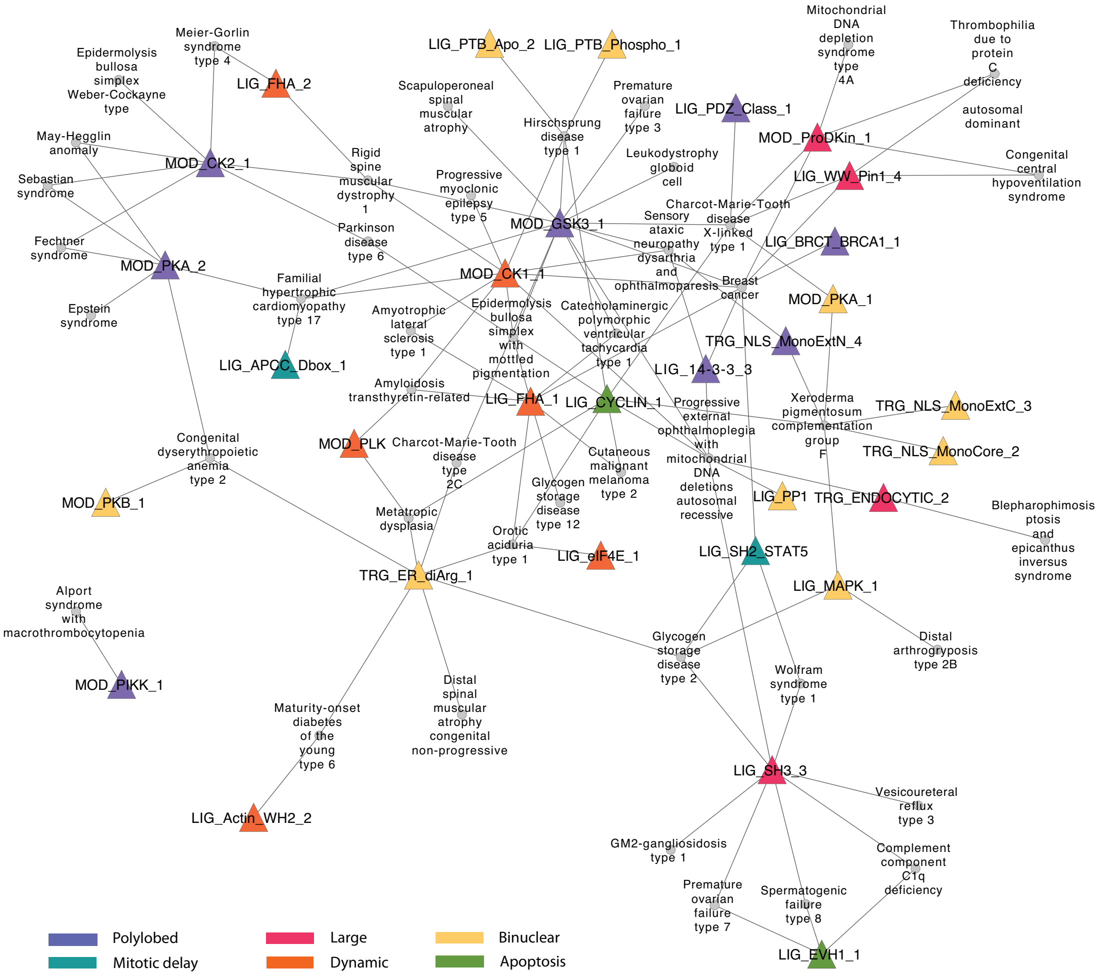

Mitochondrial DNA depletion syndrome type 4A
Thrombophilia due to protein C deficiency
Meier-Gorlin syndrome type 4
Epidermolysis bullosa simplex Weber-Cockayne type
LIG_PTB_Apo_2
LIG_PTB_Phospho_1
LIG_FHA_2
Scapuloperoneal spinal muscular atrophy
Premature ovarian failure type 3
LIG_PDZ_Class_1
autosomal dominant
Hirschsprung disease type 1
MOD_ProDKin_1
May-Hegglin anomaly
Rigid spine muscular dystrophy 1
Leukodystrophy globoid cell
Congenital central hypoventilation syndrome
MOD_CK2_1
Progressive myoclonic epilepsy type 5
LIG_WW_Pin1_4
Sebastian syndrome
Charcot-Marie-Tooth disease X-linked type 1
Sensory ataxic neuropathy dysarthria and ophthalmoparesis
MOD_GSK3_1
Parkinson disease type 6
LIG_BRCT_BRCA1_1
Fechtner syndrome
MOD_PKA_2
MOD_CK1_1
Familial hypertrophic cardiomyopathy type 17
Breast cancer
MOD_PKA_1
Epstein syndrome
Epidermolysis bullosa simplex with mottled pigmentation
Catecholaminergic polymorphic ventricular tachycardia type 1
Amyotrophic lateral sclerosis type 1
TRG_NLS_MonoExtN_4
LIG_APCC_Dbox_1
LIG_14-3-3_3
Amyloidosis transthyretin-related
LIG_FHA_1
LIG_CYCLIN_1
Progressive external ophthalmoplegia with mitochondrial DNA deletions autosomal recessive
Xeroderma pigmentosum complementation group F
TRG_NLS_MonoExtC_3
Congenital dyserythropoietic anemia type 2
MOD_PLK
Charcot-Marie-Tooth disease type 2C
Cutaneous malignant melanoma type 2
TRG_NLS_MonoCore_2
Glycogen storage disease type 12
LIG_PP1
TRG_ENDOCYTIC_2
MOD_PKB_1
Metatropic dysplasia
Blepharophimosis ptosis and epicanthus inversus syndrome
Orotic aciduria type 1
LIG_eIF4E_1
LIG_SH2_STAT5
TRG_ER_diArg_1
Alport syndrome with macrothrombocytopenia
LIG_MAPK_1
Glycogen storage disease type 2
Distal arthrogryposis type 2B
MOD_PIKK_1
Distal spinal muscular atrophy congenital non-progressive
Wolfram syndrome type 1
Maturity-onset diabetes of the young type 6
LIG_SH3_3
Vesicoureteral reflux type 3
LIG_Actin_WH2_2
GM2-gangliosidosis type 1
Complement component C1q deficiency
Premature ovarian failure type 7
Spermatogenic failure type 8
LIG_EVH1_1
Polylobed
Mitotic delay
Large
Dynamic
Binuclear
Apoptosis

Figure 2.10: Network of SLiMs and the diseases where they are mutated, according to OMIM. Diseases are depicted with grey discs. The colored triangles are the different SLiMs, with color indicating the phenotype where the linear motif is enriched. SLiMs and diseases are connected if the disease is associated to a mutation in the respective SLiM.

phisms at the SLiM sites for all motifs, according to evidence from the 1000 Genomes project and DBSNP (see Figure 2.12). The same observation is made after normalizing by the number of occurrences of each motif in the entire dataset (not shown). This normalization filters out SLiMs that occur frequently in proteins and whose mutations don't necessarily have a functional impact. However, after normalization the following SLiMs were found to have a higher rate of mutation in cancer compared to the rate of natural variation with a functional impact: LIG_APCC_KENbox_2, LIG_PP1, LIG_SH2_STAT5, LIG_SH3_2,

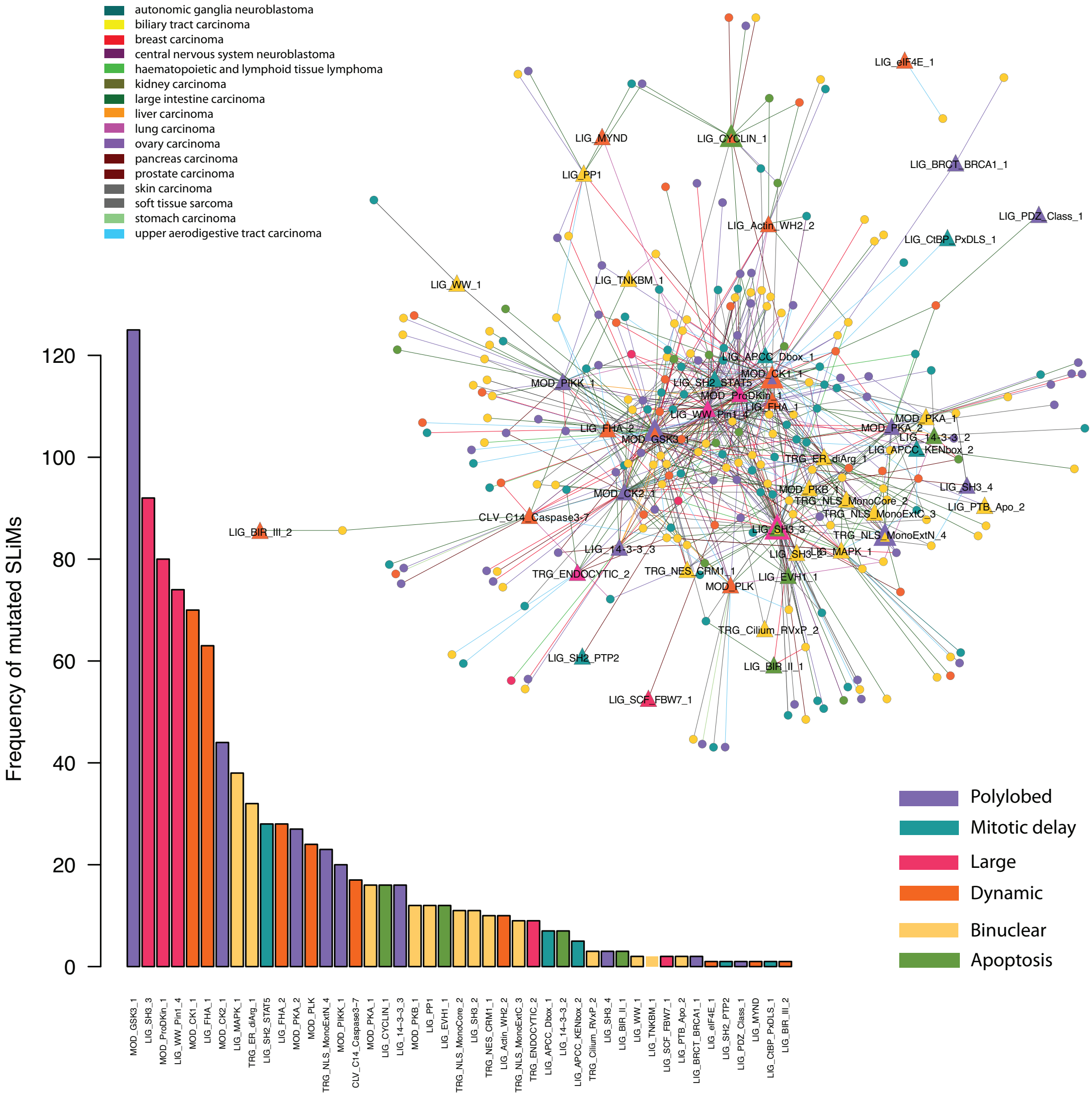


Figure 2.11: Network of SLiMs that appear mutated in proteins in different types of cancer, according to the COSMIC database. Circles in the networks denote proteins; triangles denote SLiMs. Proteins and SLiMs are connected if the linear motif is mutated in that protein in a specific disease. The color of the nodes indicates the phenotype to which the protein or the SLiM belongs. The color of the links indicates the type of cancer in which the mutation occurs. The bar chart in the lower left corner displays the distribution of the most frequently mutated SLiMs in cancer, by phenotype.

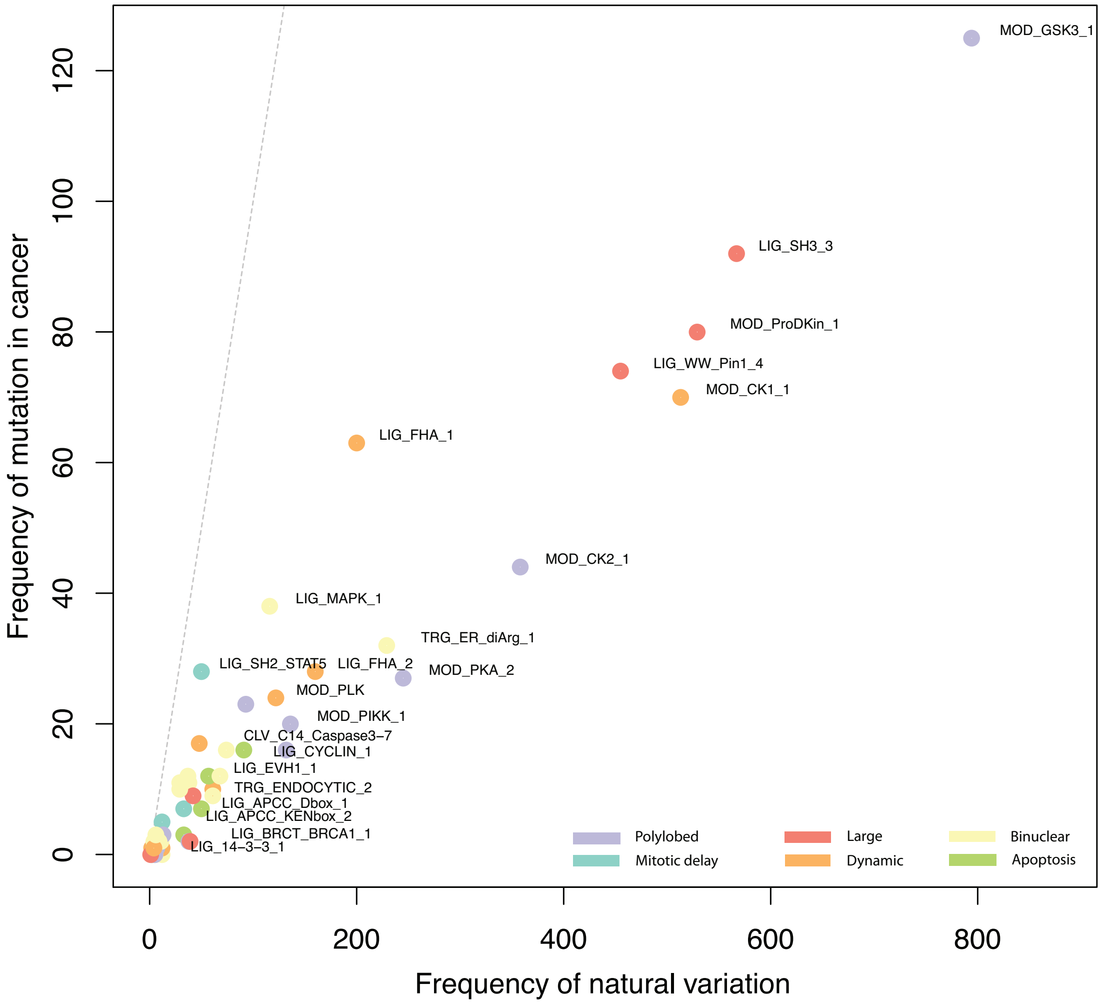


Figure 2.12: Frequency of SLiM mutations in cancer versus natural variations (polymorphisms), by enriched phenotypic group. The grey dotted line indicates equal rates of cancer-related mutations and natural variation. Anything above the line has higher rates of mutation in cancer; anything below the line has higher frequency of polymorphisms.

TRG_NES_CRM1_1, TRG_NLS_MonoCore_2. These motifs had between 1.1 and 1.6 fold higher rate of mutations in cancer. A higher rate of mutation in cancer compared to polymorphisms that occur randomly in proteins could suggest potential functional consequences of disrupting that SLiM in disease. Therefore, the highlighted motifs could be good candidates for a more detailed study into SLiM mutations in cancer.

The SLiM with the highest comparative mutation rate is the KENbox. Considering its crucial role during anaphase, the mutations identified might imply a gain or loss of function within the degradation pathway, which could trigger disease. All motifs with higher mutational rates in cancer belong either to the ”mitotic delay” or to the ”binuclear” phenotype. This indicates that these two defective cellular morphologies might be more relevant in a cancer context.

Further analysis is needed to understand the extent of mutation effects in SLiM regions in the context of cell division and the potential relevance of these factors to disease.

2.4 Discussion

2.4.1 Summary of results

Several genes essential for cell division identified in the phenotypic screen by (Neumann et al., 2010) are relevant in disease, with some phenotypes, e.g. ”polylobed”, being more prevalently associated to disorders than others (see Figure 2.13). Furthermore, the corresponding protein products are targets to a wide variety of drugs. Figure 2.14 shows these proteins and associated drugs, as extracted from the STITCH (Kuhn et al., 2010) database. The network is highly connected and protein hubs for many drugs can be easily spotted. This suggests the potential for identifying targets for drug repurposing or for elucidating side effect sources. Thus, obtaining more details about the mechanism of action of these proteins and their interactions can have medical and pharmaceutical applications.

From the analysis, I was able to distinguish linear motifs specific for different phenotypic classes. Several motifs already known to be linked to cell cycle processes appeared enriched in many phenotypic categories. This suggests that the enrichment method used performs well at distinguishing motifs specific for cell division compared to random motifs in the proteome.

Network reconstruction allowed me to infer putative novel motifs mediating interactions between proteins involved in cell cycle regulation. These should be tested experimentally. Analyzing PTM-SLiM associations for different phenotypes enabled the discovery of phenotype and SLiM-specific patterns, which

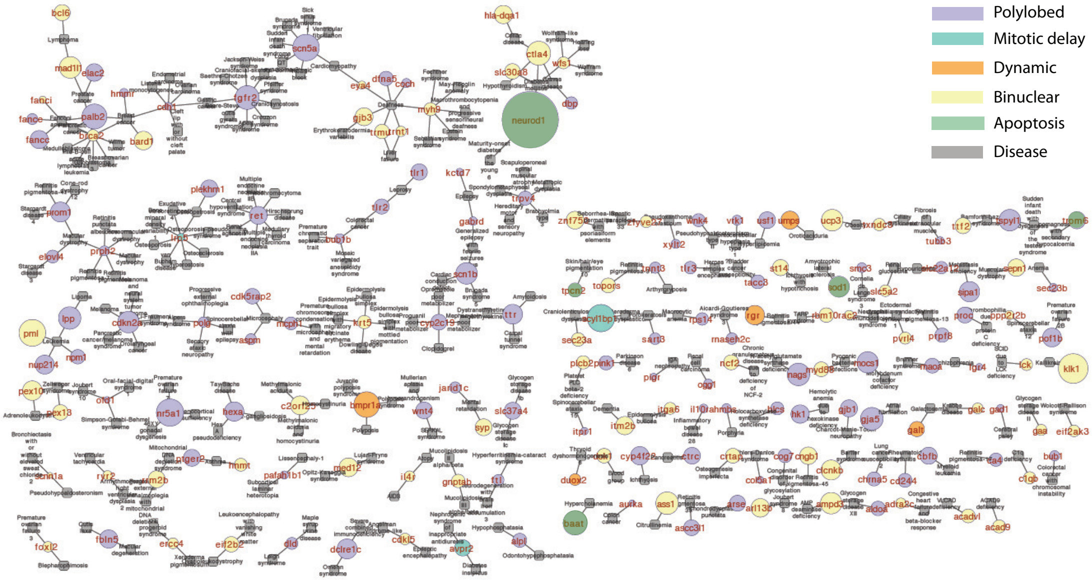


Figure 2.13: Diseases associated to genes knocked down in the experiment, by phenotype. The colored nodes depict the genes, the grey nodes the diseases linked to them as extracted from OMIM. The color of the genes depicts the first phenotype obtained upon knockdown. The size of the node is proportional to the average score associated to each gene for the respective phenotype in the cell population.

indicates the possibility of a phenotype-specific protein binding regulation. This is particularly evident for the ”large” phenotype. Further investigation might enable a better characterization of this morphological outcome.

The most common PTM classes found overall were phosphorylation and glycosylation. These have been shown to be the most abundant modification types in the proteome (Minguez et al., 2012). While the PTM-specific grouping suggests some protein features that might help distinguish phenotypic outcomes, it would be worthwhile to check how the PTM enrichment landscape in these groups changes when comparing to the rest of the proteome. This would filter out common PTM-SLiM associations and emphasize the ones that are specific for mechanisms of cell division regulation.

Mutations of SLiMs in disease were found to be often rarer than natural variation at these sites, which lowers the confidence in functionally relevant mutations. Nevertheless, the multitude of diseases these SLiMs are associated to motivates further research into the topic, for finding potential novel candidates for drug

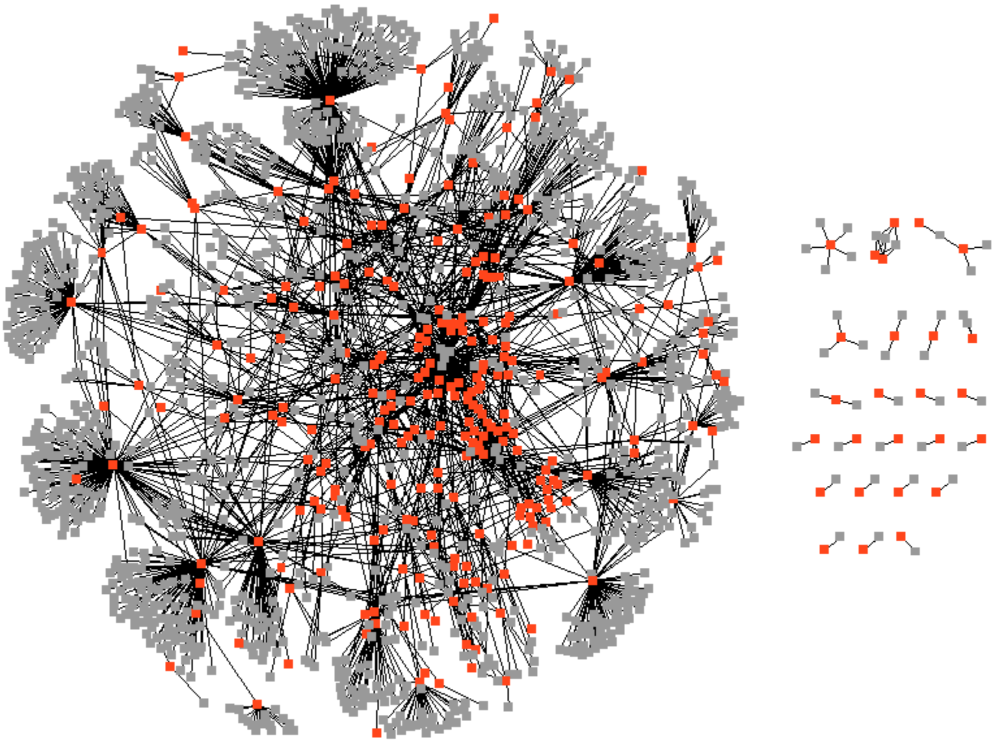


Figure 2.14: Drugs targeting protein products in the dataset of genes essential for cell division. The drugs were extracted from STITCH (Kuhn et al., 2010) and are depicted by gray nodes. The red nodes correspond to proteins.

design. Besides the fact that they play crucial roles in many processes, the small interaction surface provided by these motifs makes them better targets for intervention by small molecule compounds (Petsalaki and Russell, 2008). The connections between SLiM-mediated interaction regulation, post-translational modifications and mutations in different diseases should be further investigated.

2.4.2 Challenges

One of the main challenges of the project is the interpretability of the results. Most inferences were made on predicted, rather than annotated linear motifs, such that the functional relevance of these motifs is only assumed, but not proven. Since most linear motifs are degenerate, i.e. they occur stochastically in proteins, distinguishing the functional sites from random occurrences in the proteome is far from trivial. We have less confidence in inferences made about promiscuous motifs, e.g. LIG_SH2_STAT5, which have a high likelihood of matching protein sequences by chance. While filters like conservation or disorder score are designed to alleviate this issue, experimental validation should eventually be performed to confirm the most interesting hits.

Likewise, the disease-associated mutations are not always of functional importance. Some of these are only passenger mutations that accumulate over time throughout the disease progression, but do not effectively alter the phenotype of the cell (Haber and Settleman, 2007). While the current knowledge about passenger mutations in disease is not vast enough to allow us to confidently eliminate these non-function altering hits, some estimations are planned for the future in order to reduce these confounding factors.

2.4.3 Future directions and conclusions

Besides experimental validation of the proposed SLiM-mediated interactions, a series of computational methods are planned for future analysis. Gene ontology and pathway enrichment analysis of the linear motif clusters will be performed to obtain indications about SLiM cooperativity in fulfilling particular functions. Given that most linear motifs yield very weak phenotypes when mutated (Gibson, 2009), these complementary approaches can help us better define the biology of linear motifs in the context of the cell cycle. This is also why investigating cooperative effects of SLiMs or SLiM-PTM coupling is essential in making better predictions on regulatory mechanisms that might affect this process.

Correlations between cancer-associated mutations, linear motif content and modifications around the SLiM sites will be further investigated. For disease and natural variant analysis, we also plan to integrate resources from other databases, e.g. Protein Mutant Database (Kawabata et al., 1999) and Swiss-Prot (Bairoch and Apweiler, 2000). Moreover, structural bioinformatics techniques may be employed to investigate the binding interfaces of certain SLiMs in more detail.

A more detailed investigation of the mechanisms of domain-motif binding will enable us to understand how short peptides shape the transient interaction landscape and how robustness is built within the cell cycle. This, in turn, will allow for better inferences about the link between system disequilibrium and disease instantiation.

In the next chapters, I discuss visualization approaches that can complement the type of analysis that tries to compare or link phenotypes like the ones presented in this chapter, in a dynamic manner and taking into account the genetic, network and/or environmental context.

Chapter 3

Temporal phenotypic profiling: visualizing system-level differences with Arena3D

3.1 Description

Projecting the dynamic genetic context on the phenome has lately become of major interest in biomedicine, as gene-phenotype connections are essential to dissecting inheritance patterns, developmental outcomes, susceptibility to disease and different reactions to treatment. The temporal aspect introduces additional inferences about process evolution. However, the size and heterogeneity of the data imposes severe limitations on its interpretability. In this context, visualization tools become imperative as they leverage the understanding of complex topologies. Pattern identification helps synthesize outcomes into comprehensible forms, make new observations and hypotheses. Despite the recent deluge of time-resolved phenotypic studies, though, software that merges temporal factors and phenotypic outcomes in a network context is scarce and usually focused on a narrow range of biological data.

In the previous chapter, I have presented an analysis on regulatory differences behind different phenotypic outcomes of cell division defects. Comparing these outcomes in a systematic manner was, however, limited, and it did not take into account the temporal dimension of the process. To leverage this, the current chapter presents novel 3D visualization approaches for handling time-resolved phenotypic data as an outcome of systemic perturbations. I introduce Arena3D, a tool for 3D multilayered visualization of biological networks, and the features I have built into version 2.0 of the software. The new functionality allows the tracking and analysis of temporal patterns for different phenotypes through animation, clustering, peak highlighting, individual gene tracing, correlation display and similarity scoring. It enables users to examine genotype-phenotype relationships at different levels of depth, from molecular to tissue or entire organisms. It is therefore applicable to any perturbation analysis datasets with multiple phenotypic outcomes. Furthermore, the novel features considerably enhance the interpretation of small to medium-sized and even large datasets with a temporal component. This tool allows easy integration of different levels of information for a better understanding of how time-resolved genetic regulation reflects into phenotypic changes.

In the following sections, I describe the visualization concepts employed for time-driven phenotypic profiling. I also illustrate the effectiveness of this approach on data coming from two knockdown studies, on pluripotency factors in embryonic stem cells and on human cell division essential regulators. I used Arena3D to investigate how systemic perturbations propagate from epigenetic to translational processes, as well as to compare phenotypic patterns of cell division defects through time, as a continuation of the analysis in the previous chapter. Further details can be obtained from the published paper (Secrier et al., 2012).

3.2 Implementation

Arena3D implementation is based on the concept of multilayered graphs that are visualized in three-dimensional space. Networks of different biological entities are displayed, each on a separate layer, and these layers are connected according to the correspondence or relationships between genes, proteins, structures, diseases etc.

Networks can be displayed using several layouts: grid, circular, spherical, hierarchical, random. Moreover, different clustering algorithms are available:

affinity propagation, Markov clustering, k-means, neighbor-joining, hierarchical, UPGMA, force directed (Fruchterman-Reingold), distance geometry. Clustering enables grouping of similar entities together, where "similarity" is defined depending on the context and the biological relevance is to be established by the user. Clustering can also be performed between layers, and not only within a single layer.

The initial implementation of Arena3D, including all the features described above, was done by a previous PhD student in the group, Georgios Pavlopoulos (Pavlopoulos et al., 2008). I improved and added to the functionality of the software for the purpose of visualizing and analyzing time-resolved phenotypic responses to system perturbations. In order to manage time-resolved data, as well as compare phenotypes, I have implemented several new features into the application: time series and single gene tracking, layered clustering by gene expression, correlation calculations and statistical methods for scoring similarities and comparing phenotypes. As a result, Arena3D now captures dynamic changes in the system using several visuals: color, dynamic clustering, node enhancement, dynamic linking, node-associated graphics. These are described in detail in the following subsection.

3.2.1 Graphical methods

3.2.1.1 Dynamics captured using color

Changes in gene expression, protein concentration and other type of variations in the network can be visualized time-wise through changes in node color, where the color gradient maps to the gene/protein-associated value range for the entire time-series. The color gradient extremities will map to the lowest, respectively the highest value in the network at all time points. The minimal and maximal value is determined for each layer separately, as there may be cases when the values on different layers are not of the same magnitude or comparable (e.g. a layer of genes, with expression values associated, versus a layer of diseases, with disease severity scored on a scale). The default color gradient is yellow-to-blue, but one can switch to different predefined gradients, including color-blind safe, as well as define a custom gradient (as in Figure 3.1). Zero is denoted by gray.

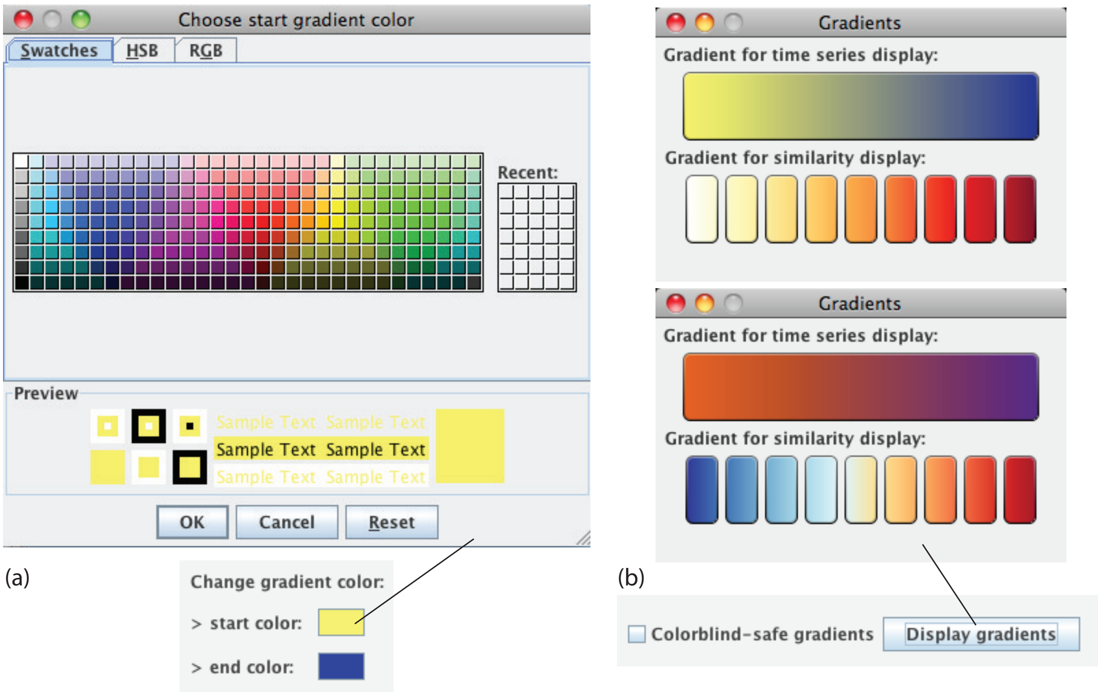


Figure 3.1: Color gradient options in Arena3D. (a) The user can change the default yellow-to-blue gradient by selecting other colors in the ”General” panel of the software. (b) Two default color gradients are shown for time series and scoring similarity feature encoding. The user can opt for one of these predefined color gradients in the ”Time-course data analysis” panel.

3.2.1.2 Clustering

Clustering on different layers according to the gene-associated values is possible, and the clustering changes dynamically at every time point. This allows the user to follow groups of genes/proteins that behave similarly through time, as well as assess the overall phenotypic differences for different conditions, tissues, or any other biological parameter that a layer represents.

The default algorithm used for clustering on a specific layer is based on distance geometry of the values associated to the genes/proteins on that layer, as described in (Crippen and Havel, 1988). First, a distance matrix between all points is calculated. Then, the distance geometry algorithm generates the coordinates of each point in 3D space. The nodes with shorter scoring distance are placed closer to each other. This algorithm does not require a predefined number of clusters into which to group the genes, but uses the distance matrix to position them in close proximity. This clustering is only performed to optimize visualization and to suggest genes with related time course profiles. It has no effect on subsequent analysis. Furthermore, one is not limited to using the default clustering, but can choose a different type of clustering among those available at any point during the analysis.

Genes/proteins are linked among layers to emphasize that the specific instance has the highest phenotypic impact at the particular time point. I.e., the gene/protein with the highest associated value for the particular time point relative to others will be connected among all layers. The top three entities are emphasized in this manner.

3.2.1.3 Individual gene tracking

If a user is interested in a specific gene/protein and wishes to follow its changes through time in the context of the network of partners and in different conditions, Arena3D enables this. Individual genes can be tracked by simple selection of the specific gene. Upon this action, the node corresponding to the selected gene will be emphasized through an increase in node size. This makes the gene easy to track visually for all time points, because even when it changes its position it can be quickly detected because of its higher volume. One can then follow the changes in expression or other time-associated values, as well as how it clusters with other genes and how this varies on different layers.

3.2.2 Statistical methods

3.2.2.1 Correlations

Similarities between gene/protein profiles can be inferred by calculating the correlations of their time-resolved vectors. For visual display, I connect the nodes on a particular layer if there is a positive (yellow links) or negative (red links) correlation between the time course gene expression profiles.

For this purpose, two correlation algorithms are available: Pearson and Spearman. The former assumes that the data is normally distributed, while the latter does not make this assumption. Arena3D does not check for this internally, so it is up to the user to decide which is the best measure to use according to the data they want to analyze.

To assess the significance of the correlation calculated using the Pearson algorithm, I use the Pearson product-moment correlation coefficient (PMCC) table of critical values. This table lists the minimal values of the Pearson correlation coefficient for a certain level of significance according to the number of degrees of freedom.

Assessing the significance of the non-parametric alternative, Spearman rank correlation, is done using the following formula:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.1}$$

where r is the correlation significance and n is the number of time points in the series. This has an approximate Student's t distribution with $n-2$ degrees of freedom under the null hypothesis.

One can set a threshold for the p-value of the calculated correlation such that only correlations with a p-value less than the threshold are displayed graphically. The choices are 0.1, 0.05, 0.02 and 0.01. The default setting is 0.05.

Importantly, when using this feature one should be aware of the limits of correlation statistics for time course experiments. The different samples in the time series data are not independent, an assumption generally made by this type of statistics calculations, so one should interpret the results with care. The correlation feature offered by Arena3D uses very simplified assumptions and it is only meant to provide a first rough indication on how similar genes are based on their time course profiles. In the future, we plan to extend these calculations to non-parametric association measures that take into account the dependency between columns (Kruglyak and Tang, 2001; Masry, 2011), as well as incorporate multiple testing corrections (e.g. Benjamini-Hochberg false discovery rate (Benjamini and Hochberg, 1995)).

3.2.2.2 Similarity scores

For overall phenotypic comparisons from the data, nodes can also be colored according to the similarity of their profiles throughout the entire time course. Each of the associated gene vectors are assigned a score that is then normalized to a range from 0 to 10. According to this the genes are placed into different bins that map to a color gradient. Afterwards, the nodes will be colored correspondingly on this scale. Identical or neighboring colors will indicate similarity in the overall time-series profile. A color gradient from white to red is used for this purpose, but other color-blind friendly gradients are also available.

Two schemes are available for scoring the genes: (a) vector value averages; and (b) the lower bound of the Wilson score confidence interval for a Bernoulli parameter, as in the following equation:

$$S(g_i, \alpha) = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}) + \frac{z^2_{\alpha/2}}{4n}}{n}}}{1 + \frac{z^2_{\alpha/2}}{n}}, t \in \{0..N\} \tag{3.2}$$

This is calculated for every gene g_i, with $i \in \{1..M\}$, where M is the total number of genes. n is the number of ratings, among which p denotes the fraction of positive ones, and $z_{\alpha/2}$ represents the $1 - \frac{a}{2}$ quantile of the Gaussian distribution (Agresti and Coull, 1998; Wilson, 1927).

Scoring scheme (a) is straightforward and offers a very general assessment of the similarity, not taking into account skewed distributions or the number of observations in the experiment. In contrast, scoring scheme (b) should balance the proportion of positive ratings with the uncertainty of a small number of observations (time points).

3.2.2.3 Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that confers a visual estimation of the number of clusters a given set of observations can be grouped into. It is based on an orthogonal transformation of potentially correlated variables that results in a division into a set of linearly uncorrelated variables denoted principal components. These principal components are ordered according to the amount of variance they explain. PCA reduces the effective dimensionality of the data without significant loss of information through the described change of basis for the vectors such that the signal-to-noise ratio (SNR) is maximized. Therefore, complex datasets are projected onto a reduced space which captures the most variable components, and thus the ones of highest interest (Jolliffe, 2002).

In Arena3D, PCA provides an additional method for the user to check for structure in the data both timewise and for different phenotypes, as well as to confirm whether the gene modules obtained by other methods reflect the real divisions of the data (Quackenbush, 2001). Before performing the PCA, the input vectors are centered by subtracting the average across all experiments from each data point. Analysis was performed using R and the JFreeChart library (http://www.jfree.org/jfreechart/) within Arena3D.

3.2.3 File formats

3.2.3.1 Input files

Arena3D accepts files in a specific format for parsing: a tab-delimited file that specifies how many layers the user wants to visualize and then goes on to describe the contents of each layer. If connections are available for the data, they are also described, specifying which layers the connection partners originate from. There are also optional variables that can be specified, like URLs for nodes or edges, connection strength etc. In the case of time course data, the user will load a slightly different file, where a series of values is specified after each node on each layer. These values correspond to the measurements for the particular entity at every time point. The total number of time points also needs to be specified for each layer. See Figure 8 in Appendix B for input file format specifications. More examples are available on the website.

In the future, users will also be able to read SBML files (Hucka et al., 2003) into the application. I have already experimented with this feature, but it is disabled in the current version until the parsing procedure is optimized.

3.2.3.2 Export files

Arena3D can export the results in the following formats: txt, jpg, Pajek, Medusa, VRML. Deciding the export file formats was done with the consideration of achieving tool cross-compatibility.

3.2.4 Summary: improvements compared to previous version

The main improvements of Arena3D have been in the direction of visualizing and analyzing time-resolved data and comparing different phenotypic outcomes. To this purpose, new functionality implemented enables:(a) tracking changes in gene/protein expression profiles throughout the time course; (b) emphasis of high-impacting genes on the phenotype; (c) clustering of genes according to values at each time point; (c) tracking of a particular gene/protein of interest; (d) comparison among genes or phenotypes through correlation analysis and similarity scoring. All these features combined enable thorough analysis of time-series datasets both on the general as well as the detailed level. One can detect or zoom into time points of interest, focus on genes of interest, as well as make comparisons among phenotypes corresponding to different conditions, perturbations, tissues etc. Table 3.1 lists in more detail the features that I have added in the new version of the software. All these features can be accessed within the Arena3D application as shown in Figure 3.2.

It is important to note that Arena3D functionality is not restricted to time course gene expression data, but is applicable to a wide-range of time course profiles coming from biological experiments, as will be demonstrated in the next subsection.

3.2.5 Extra features and future development

3.2.5.1 SBML parsing and display

Another feature I experimented with was reading and displaying biological models in the SBML file format. SBML is a standard format for describing biological processes that are summarized using ordinary differential equations (Hucka et al.,

Functionality	**Previous versions**	**Arena3D 2.0**
Input		
Network data	✓	✓
Time course data		✓
Layouts		
Circular	✓	✓
Grid	✓	✓
Spherical	✓	✓
Hierarchical	✓	✓
Clustering		
Fruchterman - Reingold	✓	✓
Distance Geometry	✓	✓
Affinity Propagation	✓	✓
Markov Clustering	✓	✓
K-Means	✓	✓
Neighbor Joining	✓	✓
UPGMA	✓	✓
Interaction		
Move nodes	✓	✓
Move/scale/spin layers	✓	✓
Time course data analysis		
Time slider		✓
Cluster by gene expression		✓
Highlight peaks		✓
Cluster by top expression changes		✓
Play animation		✓
Individual gene tracking		✓
Pearson/Spearman correlation		✓
Similarity scoring		✓
Time course line plot		✓
PCA plot		✓
Choose color scheme		✓
Colorblind-safe color scheme		✓
Network export		
Medusa format	✓	✓
Pajek format	✓	✓
VRML format	✓	✓
JPEG format	✓	✓

Table 3.1: Arena3D functionality in 2.0 and previous versions.

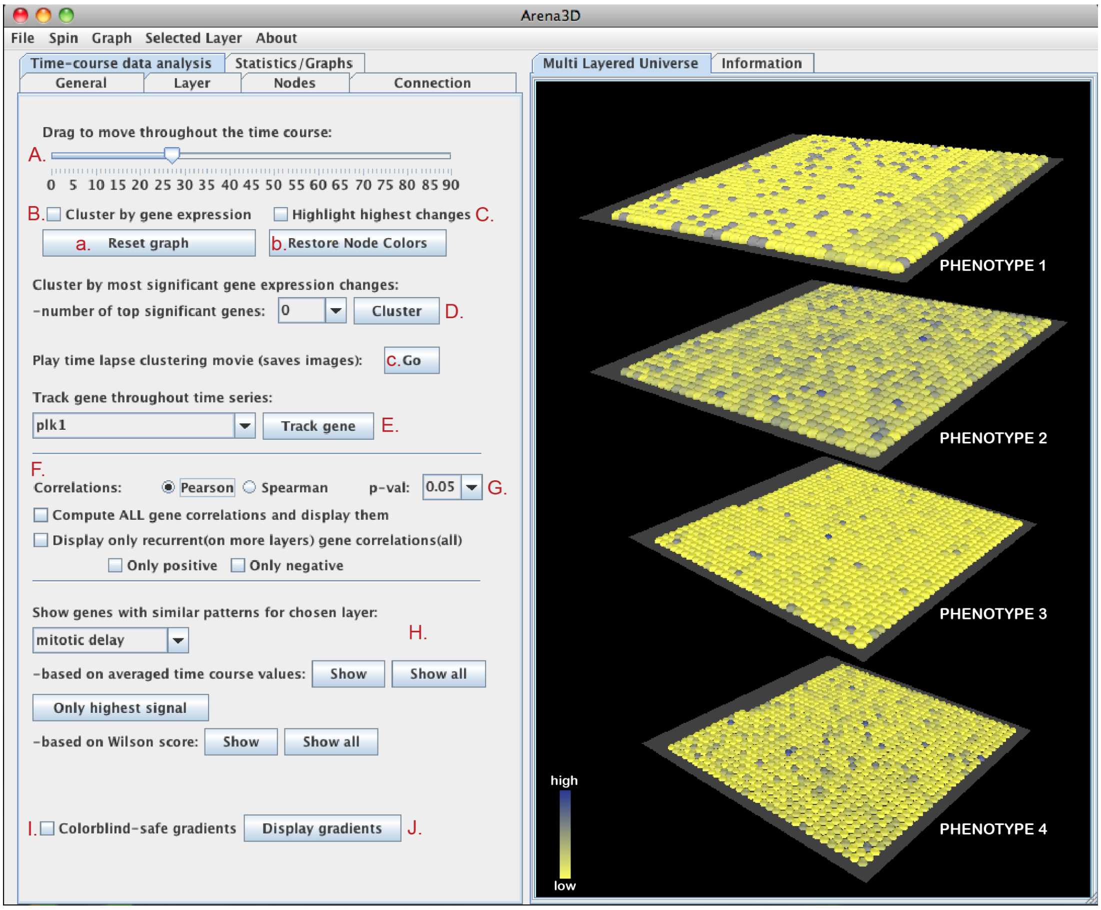

Arena3D
File Spin Graph Selected Layer About
Time-course data analysis
Statistics/Graphs
General
Layer
Nodes
Connection
Multi Layered Universe
Information
Drag to move throughout the time course:
A.
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
B. □ Cluster by gene expression
□ Highlight highest changes C.
a. Reset graph
b. Restore Node Colors
Cluster by most significant gene expression changes:
-number of top significant genes: 0
Cluster D.
Play time lapse clustering movie (saves images):
c. Go
Track gene throughout time series:
plk1
Track gene E.
F.
Correlations: ● Pearson ○ Spearman
p-val: 0.05 G.
□ Compute ALL gene correlations and display them
□ Display only recurrent(on more layers) gene correlations(all)
□ Only positive □ Only negative
Show genes with similar patterns for chosen layer:
mitotic delay
H.
-based on averaged time course values:
Show
Show all
Only highest signal
-based on Wilson score:
Show
Show all
I. □ Colorblind-safe gradients
Display gradients J.
PHENOTYPE 1
PHENOTYPE 2
PHENOTYPE 3
PHENOTYPE 4
high
low

Figure 3.2: The "Time-course data analysis" panel offers access to the newly implemented features for analyzing temporal multiple-phenotype data: (A) Time slider used to move through the data points. Every time it is moved the network visualization is updated; (B) Option that enables clustering according to gene-associated values for every time point and every layer separately; (C) Option that emphasizes genes/proteins with highest change in value between consecutive time points by connecting the corresponding nodes throughout all layers; (D) This highlights the most significant events (i.e. peaks, valleys in gene expression timeline), along with the respective time point and gene (text is displayed along with the graph to indicate these factors); (E) Individual gene tracking option (will track node by volume expansion); (F) Correlation panel, with several options for calculation methods and display; (G) Threshold for the correlation significance can be set here; (H) Similarity scoring panel, with several options for layers, methods and display; (I) Option that switches to colorblind-safe gradients; (J) Displays color gradient legend. Besides the main functionality, buttons are also available for: (a) reseting the graph (brings all nodes to initial position); (b) restoring node color (recolors the nodes according to initial color assigned by default in the application); (c) saving a series of images for all time points, that can then be used to reconstruct a movie of the time-lapse changes. The figure is reproduced from (Secrier et al., 2012).

2003). It is commonly used by a variety of software for modelling. I wanted to add this functionality to Arena3D such that biological models can be visualized in three dimensions. Parsing was done with the help of the libSBML library (Bornstein et al., 2008).

SBML files essentially describe biochemical processes of conversion of different biological "species" into "products". The rates of the reaction, as well as how the conversion takes place are specified using differential equations. Furthermore, for each species or product we know the cellular compartment where it is located. Using this information, I chose to divide the components on different layers when displaying the model: a layer of species, a layer of products and a layer of reactions that mediate the conversion. Optionally, a layer of cellular compartments can be shown. Connections to this layer would indicate the localization of reaction participants. Proper alignment of layers and rotation in 3D space enables an easy and thorough investigation of the reactions describing a specific process. Figure 3.3 shows how visualization in 3D can relieve overcrowding in 2D space that many times impedes good visualization and interpretation of results. This task will become more cumbersome in the future, with bigger and more complete maps and pathways of biological processes being produced. For instance, the comprehensive yeast cell cycle pathway described in (Kaizu et al., 2010) and mentioned in the *Introduction* currently has no feasible solution for investigation or even just display. Using tools like Arena3D might help reduce the complexity by structuring it into different layers that can then be more easily understood.

The next step would be to use the information given in the system of ordinary differential equations (ODEs) to perform a simulation and correspondingly map the changing concentrations of species and products through time, similar to what is shown in Figure 9 of Appendix B. This step has not been implemented yet, but is planned for the future. There are several limitations in simulating the SBML files: some of them deviate from the standard format and cannot be parsed properly, some are too big for simulating in a feasible time frame, some would even be too big for feasible display or would exceed the memory capacity of the application. These limitations, nevertheless, can be overcome by restricting the simulation and display of results only to models of reasonable size.

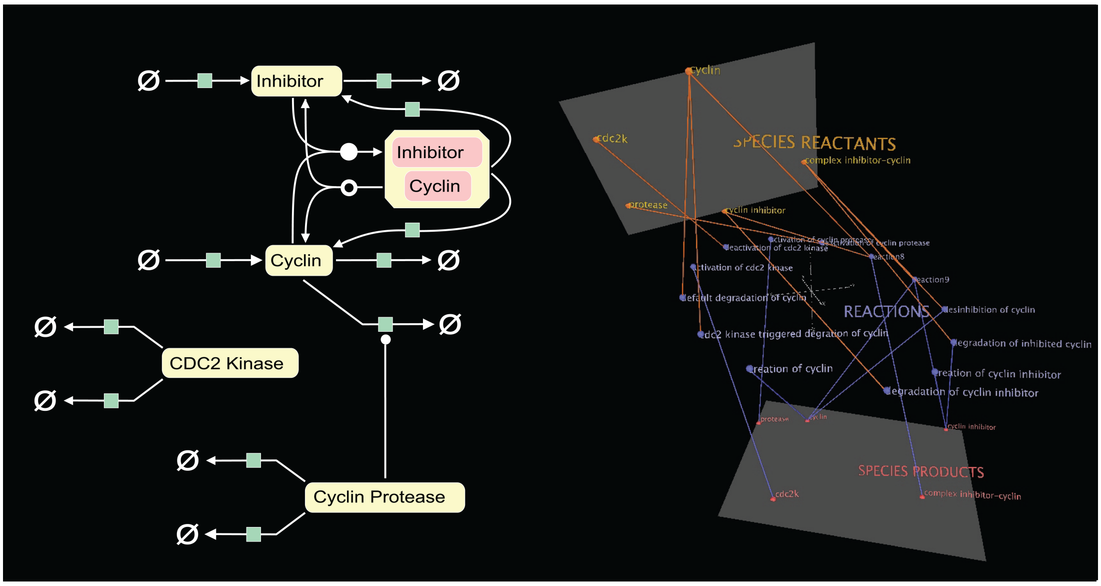


Figure 3.3: Simplified scheme of biochemical reactions occurring through the cell cycle, as described in (Gardner et al., 1998). Left: reaction graph obtained from the BioModels database (Le Novère et al., 2006). Right: layered depiction of reaction flow from reactants to products as displayed by Arena3D. While this example is relatively small and just for illustration purpose, the power of 3D will prove more useful when handling huge pathways.

3.2.5.2 Pairwise vector derivative plots

As an alternative to identifying patterns in time course data, I implemented an experimental feature in Arena3D that consisted in plotting pairwise gene vector paths. This feature was tested only on the dataset coming from (Neumann et al., 2010), as described in section 3.3.2. More explicitly, I plotted the derivative of the measurement vectors for every pair of genes, only for the cases when there is proof of phenotypic impact changes upon knockdown of more than 50% of the previous phenotypic value between two consecutive time points. Noise was introduced to amplify changes greater than 80% so that they are more easily recognizable in the plot. The resulting paths of phenotypic progression are indicative of how strongly a pair of genes individually knocked down would affect the undertaking of a particular phenotype. The phenotypic paths take one of the 8 directions as shown in Figure 3.4 to indicate whether the coupled knockdown exhibits an increase in prevalence of the respective phenotype in the cell populations at the respective time point compared to the previous one, a decrease, or no change. Since the assumptions of this method might have easily misguided an uninformed user to misinterpret the results and since for the studied dataset the observations made were inconclusive, we decided to exclude this feature from the final version of the software.

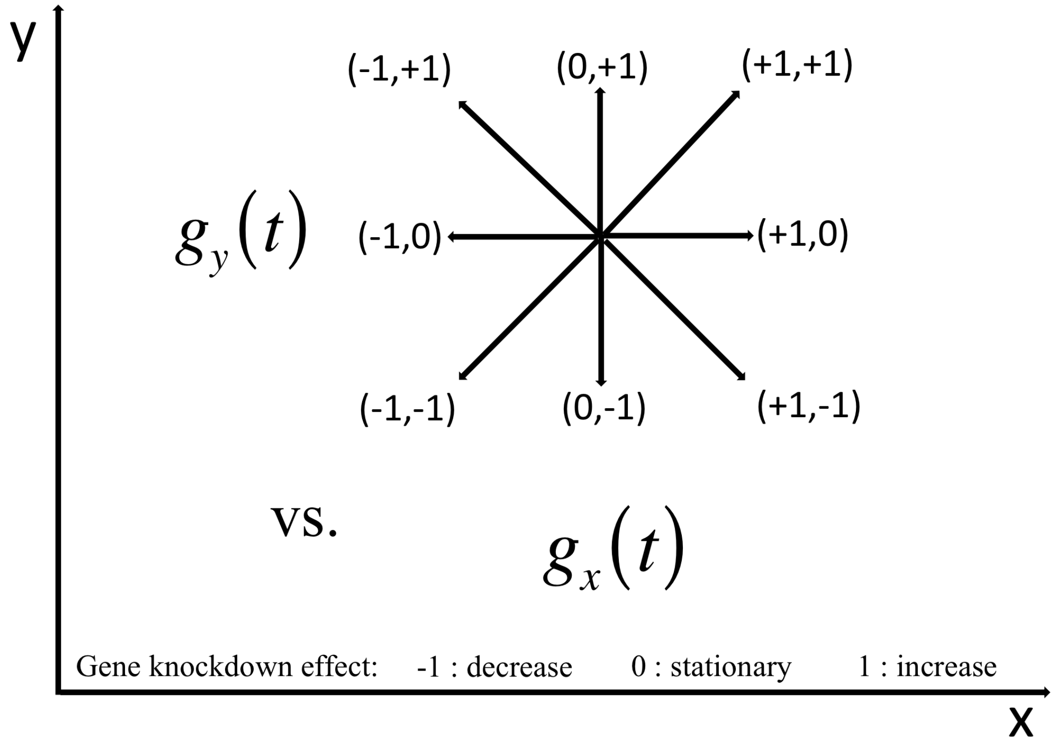


Figure 3.4: Basis for plotting the phenotypic paths derived from coupled gene-associated vectors: increases, decreases and stationary effects of the knockdown at a specific time point determine the direction of the plotted line. $g_x(t)$ and $g_y(t)$ indicate that the values evaluated are the ones of genes g_x and g_y, where $x, y \in \{1..n\}$, with n being the total number of genes in the dataset, and t is the current time point.

3.2.5.3 Integration with the Garuda platform

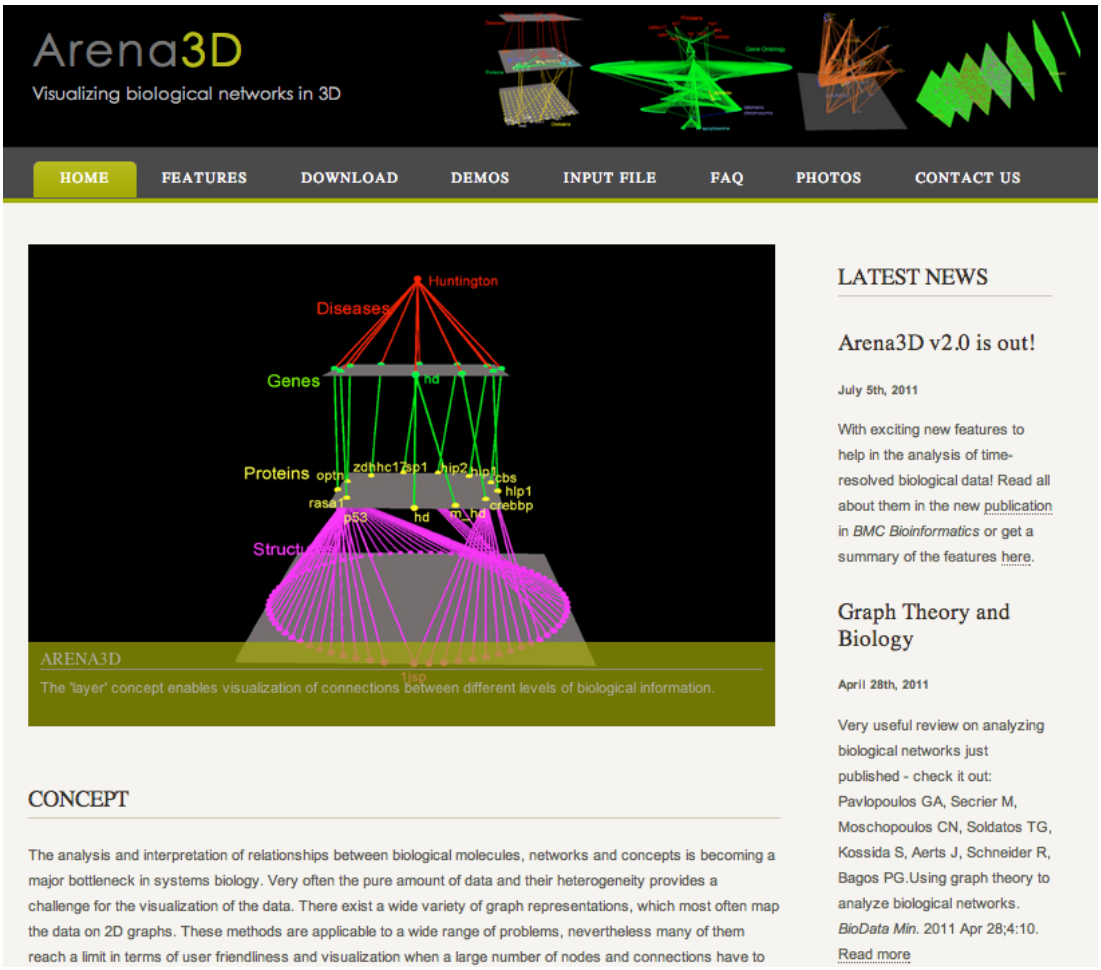
Garuda is an organized effort to develop a common platform that integrates different biomedical software for the use of both the academic and the industrial sector (Ghosh et al., 2011). The main idea behind the project is to have a platform into which different tools can be plugged, such that they communicate with each other and can send results from one to the other. This creates the possibility of having workflows for repetitive tasks, similarly to Galaxy (Blankenberg et al., 2010; Giardine et al., 2005; Goecks et al., 2010) for NGS studies. For the moment, the software suite focuses on applications meant for modelling and visualization in Systems Biology, but it aims to go further than that with the participation of other interested groups after it launches. Tools like CellDesigner (Funahashi et al., 2003), Cytoscape (Shannon et al., 2003), PhysioDesigner (Asai et al., 2012) and others have already been enabled to communicate through Garuda.

As part of this initiative, I have integrated Arena3D into Garuda. Within the platform, it can visualize results coming from any other tool, as long as they are converted into the Arena3D specific file format, as well as visualize SBML files. The integration into Garuda should enable Arena3D to gain access to a wider audience of biologists and it will definitely help popularize the tool.

3.2.6 Technical specifications and availability

Arena3D was implemented using Java (JDK 1.6) and Java3D (1.6.1 API). The JFreeChart library is used for the line plot view of time course values upon node click events, as well as for the PCA plots. The software is freely available for academic use as a standalone platform-independent application downloadable from the website http://arena3d.org/ (see Figure 3.5). The initial website created by Georgios Pavlopoulos has been completely redesigned by me, with some of the contents kept and considerable content added.

The Java Runtime Environment (http://www.java.com/) and Java3D libraries (http://java3d.java.net/) are required for running Arena3D on any operating system and Macintosh users should also install the JOGL libraries from http://opengl.j3d.org/. Simple API implementation for plug-in development is planned for the future. The source code is available for download for users that wish to customize their analysis.



ARENA3D

The 'layer' concept enables visualization of connections between different levels of biological information.

CONCEPT

The analysis and interpretation of relationships between biological molecules, networks and concepts is becoming a major bottleneck in systems biology. Very often the pure amount of data and their heterogeneity provides a challenge for the visualization of the data. There exist a wide variety of graph representations, which most often map the data on 2D graphs. These methods are applicable to a wide range of problems, nevertheless many of them reach a limit in terms of user friendliness and visualization when a large number of nodes and connections have to

LATEST NEWS

Arena3D v2.0 is out!

July 5th, 2011

With exciting new features to help in the analysis of time-resolved biological data! Read all about them in the new publication in *BMC Bioinformatics* or get a summary of the features here.

Graph Theory and Biology

April 28th, 2011

Very useful review on analyzing biological networks just published - check it out: Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG.Using graph theory to analyze biological networks. *BioData Min.* 2011 Apr 28;4:10.

Read more

Figure 3.5: Screenshot of the homepage of the Arena3D website, located at http://arena3d.org/. The website was completely redesigned for presentation purposes.

3.3 Results

Arena3D has been used with several datasets in order to identify patterns in time series outcomes of different biological experiments. Two of the applications are described below.

3.3.1 System-level differences in the epigenetic, transcriptional and translational dynamics of embryonic stem cells

In the first case study I looked at how downregulation of certain factors in the cell propagates from the epigenetic to the organismal level and how phenotypic differences arise as a result. The dataset employed comes from a knockdown experiment of the pluripotency regulator *Nanog* in embryonic stem cells (ESC) (Lu et al., 2009). Upon downregulation, dynamic changes are recorded at three different levels: epigenetic, transcriptional and translational. These are described in measurements of histone acetylation, RNA polymerase II localization, mRNA abundance and protein levels for a set of genes at three time points (days 1, 3 and 5). I used Arena3D to visualize dynamic changes within the core ESC protein-protein interaction network, as defined in (Lu et al., 2009). The newly implemented functionality for time course data handling enabled me to discover patterns not found in the original paper, such as recurrent correlations in perturbation dynamics and potential network rewiring, as discussed later.

First, I parsed and converted the data into the Arena3D specific format so that it could be read by the application. The dynamic changes upon *Nanog* downregulation were represented on four different layers, corresponding to the four biological levels of measurements: histone acetylation, chromatin-bound RNA polymerase II, mRNA levels and nuclear protein abundance, as shown in Figure 3.6. Each layer depicts a network, the ESC core, where nodes correspond to genes/proteins and links between them indicate interaction. The color gradient used for the nodes maps to the gene or protein-associated values for each layer. The color of the node changes at every time point according to how the histone acetylation, polymerase occupancy, mRNA or protein levels increase or decrease. The lowest values are coded in yellow, highest in blue and intermediate according to the gradient. Grey stands for values of zero. The changes in these values through time can then be easily tracked visually by using the time slider provided in the application. Moving the slider will result in an update of the network with the corresponding state at that specific time point, encoded in color and/or other representations.

3.3.1.1 Clustering reveals dampening of perturbation from the chromatin to the protein level and potential fragility points

Clustering the nodes on every layer for consecutive time points paints a dynamic landscape in the ESC core network, that is highly variable at the chromatin level, but rather constant at the protein level. Figure 3.6 shows how the network structure changes considerably upon clustering from day 1 to day 5 from the point of view of histone acetylation amounts and polymerase occupancy. In contrast, the mRNA and protein abundance layers display more stable networks and less change in the measured values as well. Thus, downregulation of *Nanog* seems to have a prominent effect at the epigenetic level, with less perturbation of the transcriptional and translational processes.

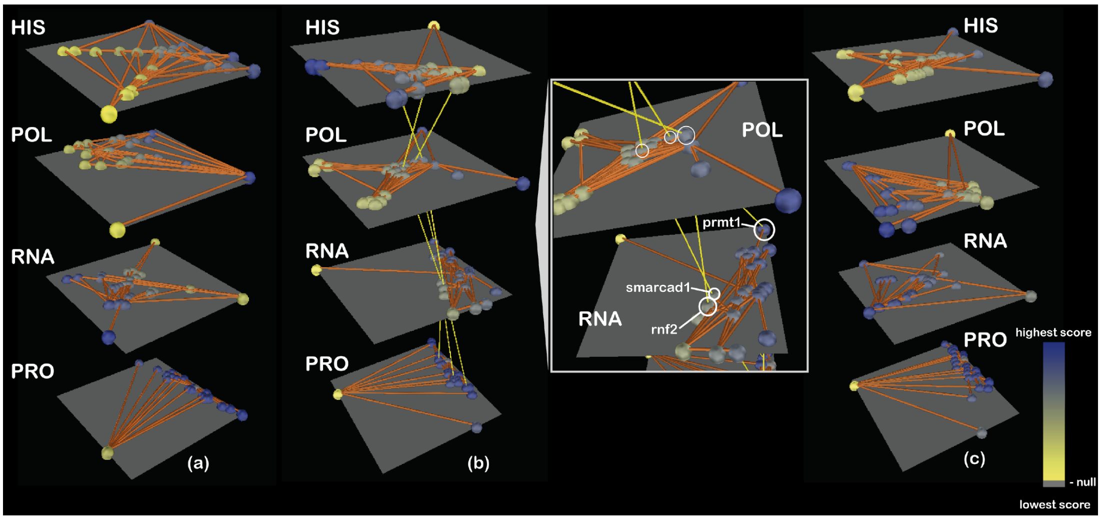


Figure 3.6: Dynamic clustering of layered effects upon system perturbation. Each layer depicts the ESC core network, with the component genes/proteins colored to indicate changes in histone acetylation levels (HIS), RNA polymerase II binding affinity (POL), mRNA production (RNA) and protein levels (PRO) as a result of downregulation of the pluripotency factor *Nanog*. The color of the nodes changes according to the values associated to each time point, on a scale from yellow to blue. Clustering patterns are shown for each phenotypic outcome at three time points: (a) day 1; (b) day 3; (c) day 5. The evolution of gene associations indicated by clustering suggests a more dynamic landscape upon perturbation at the epigenetic rather than translational level. Subfigure (b) highlights genes with the sharpest change in measurements at the respective time points by connecting them with yellow links throughout all layers. The close-up picture reveals the name of these genes: *Prmt1*, *Smarcad1* and *Rnf2*. The figure is reproduced from (Secrier et al., 2012).

Nanog downregulation has highest impact on genes *Smarcad1* [Ensembl: ENSG00000163104], *Prmt1* [Ensembl:ENSG00000126457] and *Rnf2* [Ensembl: ENSG00000121481], as automatically highlighted in Arena3D by linking the nodes on different levels at the second time point. *Smarcad1* is a matrix-associated regulator of chromatin that is actin-dependent, *Prmt1* is an arginine methyltransferase and *Rnf2* is a ring finger protein belonging to the Polycomb group. Their roles in the cell, preceding mRNA synthesis, justify why the recorded signal is higher at the epigenetic levels for these genes. Their strong impact change in the context of the ESC core network (Figure 3.6), where they are peripherally situated, might suggest there is an alternative route from *Nanog* to these genes that makes them highly susceptible to the downregulation of the former factor. The reasons for this network fragility need to be investigated further experimentally.

3.3.1.2 Correlation calculations indicate a high level of heterogeneity, but also recurring patterns between transcriptional and translational levels

I looked at correlations between gene-associated measurements in the ESC core network at the four defined levels, from epigenetic to protein synthesis. I used the Pearson correlation coefficient to determine the genes that are positively and negatively correlated in this example, but the same procedure can be undertaken using the Spearman rank correlations instead. After selecting the algorithm and the threshold for the p-value (0.05 in this case), the results of the calculations were displayed by linking significantly correlated genes in the network on each layer (see Figure 3.7).

Importantly, the user should consider whether the number of data points available from measurements justifies performing this calculation at all: three time points would normally be considered insufficient for obtaining significant correlations (degree of freedom is 1). In this example, however, there are several cases with significant correlations with coefficients greater than 0.997 at a p-value less than 0.05. For illustration purposes we consider this sufficient, but the user should use this feature with careful consideration, on a case-by-case basis.

Figure 3.7 shows the correlations in acetylation patterns, chromosome occu-

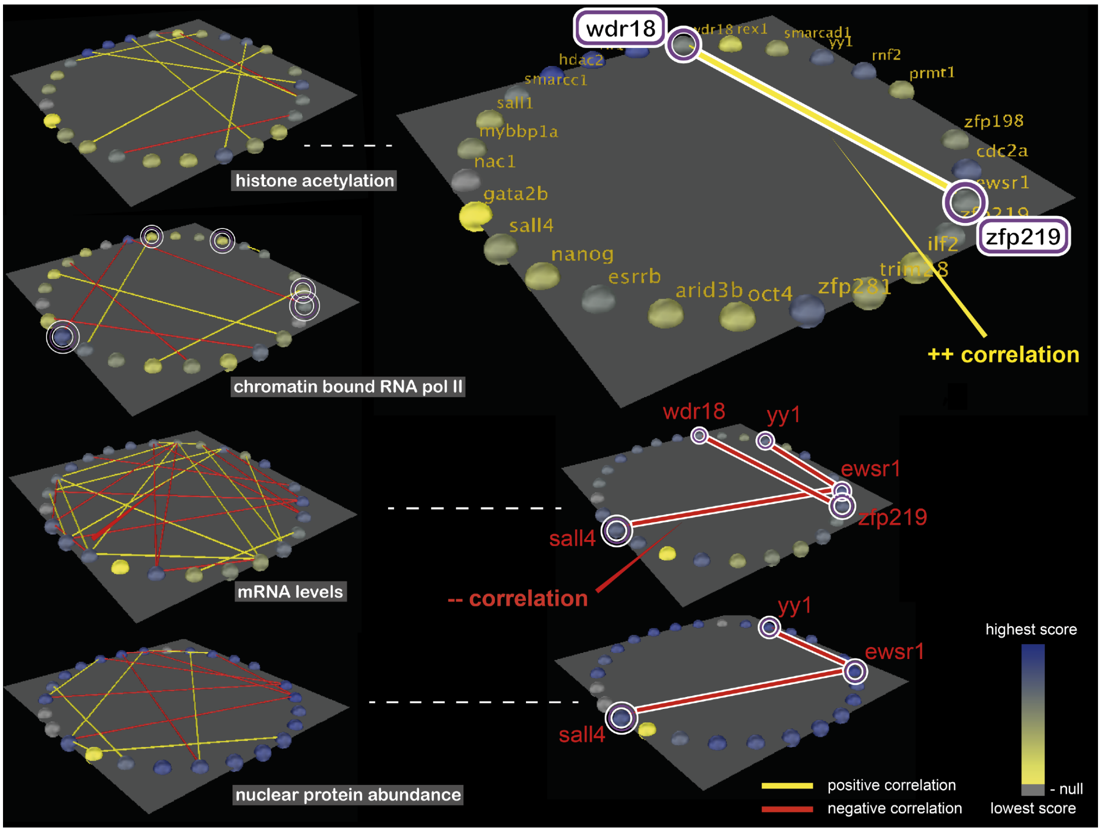


Figure 3.7: Correlations from the epigenetic to the translational level are calculated and displayed as links between nodes for genes in the ESC core network based on their 3-day measurement profiles. Nodes are colored according to the gene-associated value on a yellow-to-blue color gradient. Yellow links indicate positive correlations between the genes corresponding to the respective node pair and red links are negative correlations. The left hand side of the figure shows all significant correlations (p-value< 0.05 and correlation coefficient higher than 0.997) as connections between nodes. The right hand side shows only recurrent correlations, i.e. any correlation between the same pair of genes that holds on at least 2 layers. The layer of RNA polymerase II occupancy is omitted because it contains no recurrent correlations. The figure is reproduced from (Secrier et al., 2012).

pancy, expression or protein amount between genes/proteins in the core ESC network (left). More significant correlations are recorded at the mRNA synthesis level than at the epigenetic level. Also, there seem to be more positive correlations at the epigenetic level and more negative correlations at the protein level. This shows that perturbations in the system affect chromatin processes in a way that does not necessarily reflect subsequent alterations of protein fluxes.

The right hand side of Figure 3.7 focuses on recurrent correlations, i.e. correlations that appear between the same partners on at least two different layers. These recurrent correlations are indications of synergies between epigenetic, transcriptional and translational processes. Genes *Wdr18* and *Zfp19* are positively correlated in terms of acetylation patterns and negatively correlated in mRNA levels. Furthermore, no correlation exists between them for the other two layers. Both *Wdr18* and *Zfp19* are protein-coding genes with unclear function. The correlation at epigenetic and transcriptional levels suggests that they might appear in the same pathways, but with different stoichiometries. These patterns, along with a general scarcity of significant correlations observed between genes in the ESC core network, suggest a high level of heterogeneity from the chromatin level down to protein outcomes.

However, some consistent correlations can also be observed. Gene *Ewsr1* negatively correlates both with *Yy1* and *Sall4* at the level of mRNA and protein abundance changes in time. All three genes are involved in transcription regulation, according to GeneCards (Rebhan et al., 1997). *Yy1* belongs to the class of zinc finger transcription factors. *Sall4* has also been stipulated to belong to the same class. Since *Yy1* can act either as a repressor or an activator of transcription (Shi et al., 1991; Wu et al., 2007) and *Ewsr1* represses transcription by the Polymerase II machinery (Li and Lee, 2000), it could be that *Yy1* and *Ewsr1* function in an exclusive manner to regulate the expression of certain genes. *Sall4* is likely to be involved in similar processes as *Yy1*. Moreover, this recurrence at transcriptional and translational level is in accordance with evidence from the literature that mRNA and protein copy numbers correlate even though their half-lives do not (Schwanhäusser et al., 2011).

3.3.2 Temporal profiles of phenotypic defects in cell division upon single perturbations in the system

In the second case study, I continued the analysis of phenotypic profiles of cells throughout the cell cycle upon single gene knockdown, as described in (Neumann et al., 2010) and introduced in the previous chapter. Time-lapse imaging reveals defective morphologies of cells as a result of deletion of essential genes for cell division. In contrast to the work presented in the first chapter, here I employed not only the first phenotype observed in the population, but the entire time course phenotypic profiles in the analysis. The strategy in this case focused on capturing dynamic global patterns in the dataset and comparing the phenotypes based on these patterns.

I used Arena3D to represent the seven main phenotypes on different layers. The positions and colors of the nodes (genes) indicate how prominent the impact of the respective gene knockdown is in the cell population at the particular moment in time. The results spanned the first 90 time points, or 45 hours of cell life. The application visualized simultaneously the effects of all individual knockdowns in determining the cells to adopt a certain phenotype. Changes in phenotypic penetrance for every gene knockdown were projected through proportional changes in color on a gradient scale analogous to the one in the previous subsection. In this case the values associated to the genes came from the phenotypic scoring scheme based on morphological features extracted from the images of the affected cells, as described in section 2.2.1. Nevertheless, this type of visualization is applicable to any datasets with gene expression, protein concentration or other time course measurements.

3.3.2.1 Cluster dynamics unfold resistant and volatile phenotypes

The full dataset consists of 1067 essential mitotic genes. I selected a subset of genes as discussed in the paper (Neumann et al., 2010) for more detailed analysis (see Table 3.2). Visualization of the impact of suppressing these genes on different phenotypic outcomes reveals morphology-specific changes, as shown in Figure 3.8. Dynamic clustering on each layer allows more effective comparison between phenotypes. One can even distinguish between relatively more resistant

Gene name	Description
anln	anillin, actin binding protein
aurkb	aurora kinase B
bard1	BRCA1 associated RING domain 1
c13orf23	chromosome 13 open reading frame 23
c14orf54	family with sequence similarity 71, member D
cabp7	calcium binding protein 7
cenpe	centromere protein E, 312kDa
ckap5	cytoskeleton associated protein 5
ect2	epithelial cell transforming sequence 2 oncogene
incenp	inner centromere protein antigens 135/155kDa
kif11	kinesin family member 11
kif23	kinesin family member 23
lsm14a	Sm-like protein, SCD6 homolog A (S. cerevisiae)
mfsd3	major facilitator superfamily domain containing 3
myh9	myosin, heavy chain 9, non-muscle
plk1	polo-like kinase 1
prc1	protein regulator of cytokinesis 1
ptger2	prostaglandin E receptor 2 (subtype EP2), 53kDa
rab24	member RAS oncogene family
racgap1	Rac GTPase activating protein 1
rgma	RGM domain family, member A
tor1aip1	torsin A interacting protein 1
tpx2	microtubule-associated, homolog (Xenopus laevis)

Table 3.2: List of potentially interesting mitotic genes, as discussed in the paper (Neumann et al., 2010). Information about the genes has been extracted from the GeneCards database (Rebhan et al., 1997).

and more volatile phenotypes: ”mitotic delay”, ”binuclear” and ”polylobed” display steady clustering patterns through time, whereas the other morphological categories show greater variation in node positioning at different time points. The more ”dynamic” phenotypes may show this behavior because they are transient and thus rapidly succeeded by a different, more stable morphology within the cell population.

”Apoptosis” is included in the latter, more changeable category. This seems counterintuitive at first sight, since it is a final phenotype. However, the measurements reflect prevalence of a specific phenotype within a cell population, not

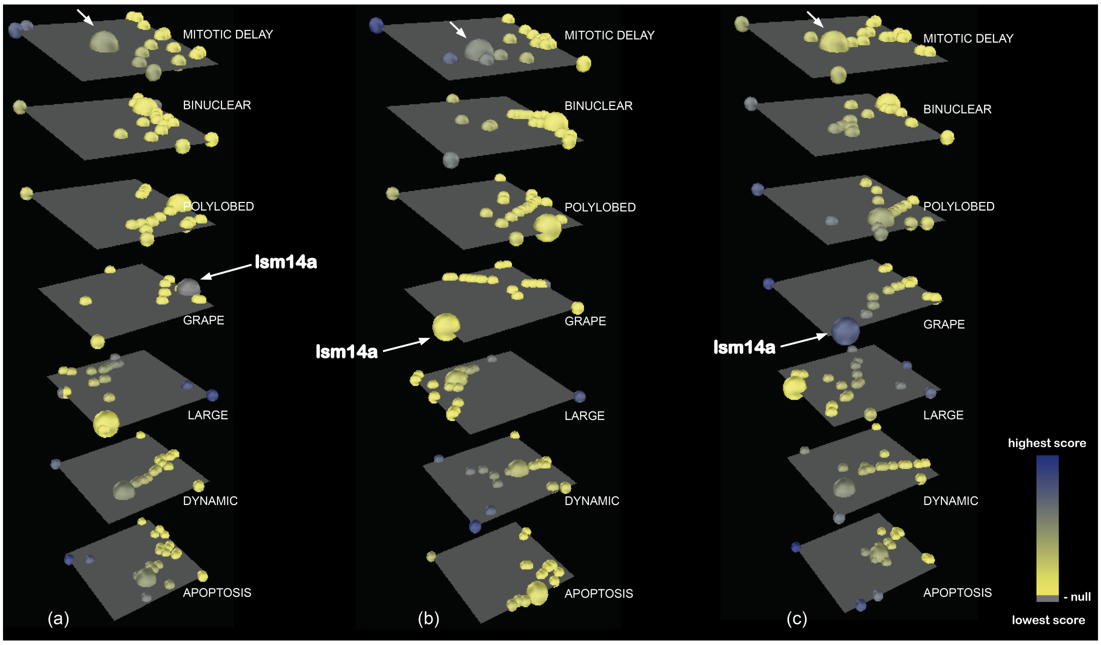


Figure 3.8: Dynamic clustering of phenotypic outcomes as a result of defective cell division. The seven main mitotic phenotypes are represented as different layers, each one containing the subset of essential mitotic genes described in Table 3.2 depicted as nodes. These nodes are colored on a yellow-to-blue gradient depending on the score associated with the impact of the respective gene knockdown on the cell population at every time point. The figure shows the outcome for all phenotypes at three selected time points: (a) t = 2h; (b) t = 7h; (c) t=33h. Nodes are clustered to indicate similarities in the knockdown profiles of several genes along the time course. Comparing clustering profiles reveals more variation in the ”grape”, ”large” and ”dynamic” morphologies compared to the rest, indicating they are more dynamic, whereas ”mitotic delay” and ”polylobed” have more stable genetic effect patterns. Gene *LSM14A* is tracked throughout the time course by node expansion (also highlighted by arrows here for the ”mitotic delay” and ”grape” phenotypes). Suppression of this gene leads to a mild, but increasing effect with time on the former phenotype, and a latent but pronounced impact at the last time point depicted on the latter phenotype. This seems to suggest that the ”grape” morphology is adopted after a period of stagnation during mitosis. The figure is reproduced from (Secrier et al., 2012).

an individual cell. Thus, as consequence of cell turnover, at future time points after apoptosis dominance other phenotypes will take over as new cells develop and start dividing. This is why apoptosis signals can sparsely appear and disappear in the imaged cells, making the phenotype more dynamic. In contrast, "mitotic delay", "binuclear" and "polylobed" phenotypes tend to linger longer in the populations, as cells are arrested in these morphologies without dying.

3.3.2.2 Time course tracking of gene *LSM14A* suggests potential novel roles in cell division

Arena3D can be used to track individual genes of interest. Here I exemplify this feature on gene *LSM14A*, an Sm-like protein thought to be involved in pre-mRNA splicing and P-body formation. It has also been suggested it becomes associated with the mitotic spindle during cell division. Figure 3.8 shows this gene tracked through time by node volume incrementation. The visualization suggests that the knockdown of *LSM14A* determines the cells to latently adopt the "grape" morphology. Interestingly, "grape" is a rare phenotype, because very few cells have been observed throughout experiments to assume this phenotype. Understanding what might cause such morphology is therefore a particularly difficult challenge. Comparative tracking of *LSM14A* on different phenotypic layers uncovers more information: the impact of the gene is rather mild for "mitotic delay" in the first studied time point (Figure 3.8a), becomes stronger in Figure 3.8b, after which it switches pronouncedly to "grape" (Figure 3.8c). This helps reconstruct an ordered phenotypic succession within the cell population for the *LSM14A* knockdown. Given its studied functions in the literature and the previous observation, we might infer novel hypotheses for this gene. Association with the mitotic spindle and the resulting "mitotic delay" morphology upon suppression suggest roles in karyokinesis. The subsequent "grape" phenotype might additionally imply a potential additional role in cytokinesis, since the multiple micronuclei characteristic for this morphology can be the result of defects either in nuclear or cytoplasmic separation. Further experiments should be conducted to establish the subprocesses in which the protein product of *LSM14A* is involved. Additional evidence would enable us to revise the knowledge about this gene's versatility and adaptability.

3.3.2.3 Global phenotypic patterning aids comparison and pinpoints potential interesting targets

Similarity scoring is a feature that can be used to identify genes with analogous time-resolved profiles, as well as easily compare phenotypes and their progression, especially in large-scale datasets. I illustrate the effectiveness of pattern comparison in the full mitotic defects dataset comprising the 1067 genes essential for cell division and the effects of their knockdown. Figure 3.9 shows these genes scored on different phenotypic layers according to the two scoring schemes discussed in subsection 3.2.2.2: (a) averages of knockdown score vectors; and (b) lower bounds of Wilson score confidence intervals. The two scoring schemes allow for different interpretations based on individual statistical assumptions and calculations.

Coloring the nodes according to the former scoring scheme emphasizes relatively few highly scoring genes, whose suppression should have a strong impact on the cell. "Polylobed" appears as the phenotype with most highly scoring genes, and is indeed a prevalent phenotype throughout many of the screens. The relatively low number of strong signals allows preselection of these genes as potentially interesting targets for future experiments.

The scoring scheme employed in Figure 3.9b, on the other hand, eliminates noise caused by low signals in the data. Consequently, it enables better comparison between genes among single phenotypes. Caution is required when comparing phenotypes or interpreting high signals. First, the signal for one gene cannot be matched among different phenotypes because each morphology is uniquely scored, so phenotypes are not comparable. Second, the normalization used in the scoring method has the effect of bringing out many high signal points in pools of low values (for instance, the "grape" phenotype shows many highly scoring genes after normalization, but the phenotype overall is a rare one) and these high signals should not be interpreted as prominent phenotypes. What this scoring scheme does instead is enable true signal discovery within a particular phenotype.

The line plots of the time course scoring measurements for the genes *INCENP* (an inner centromere protein antigen [Ensembl:ENSG00000149503]) and

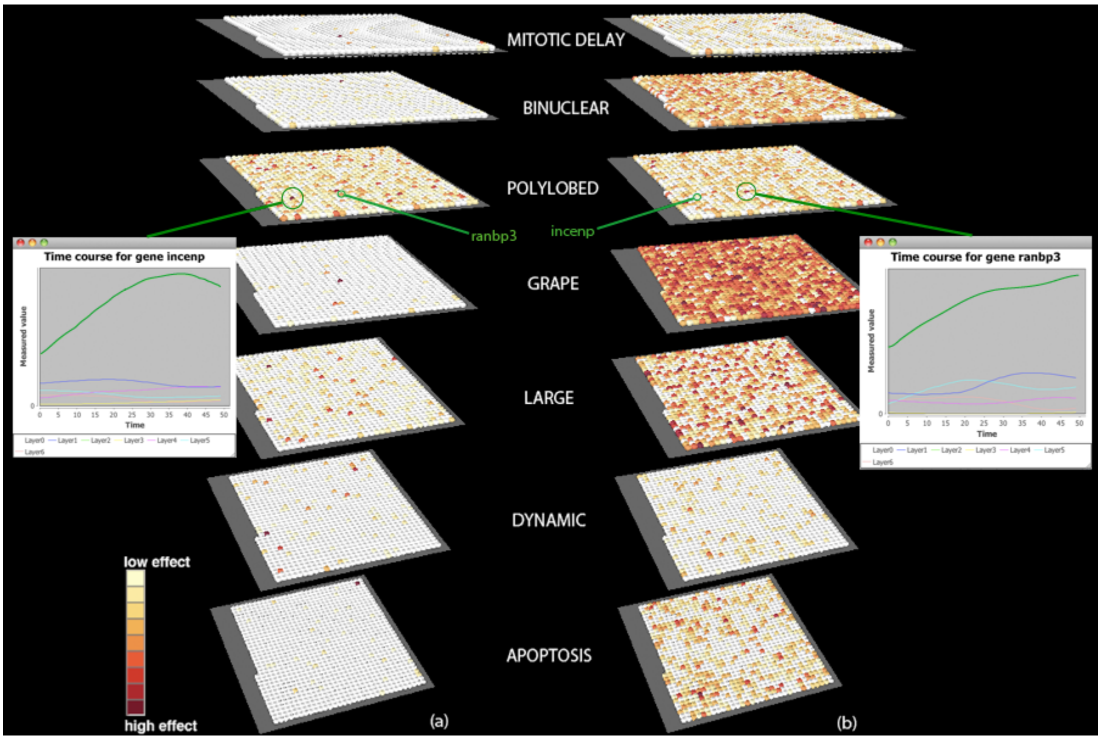
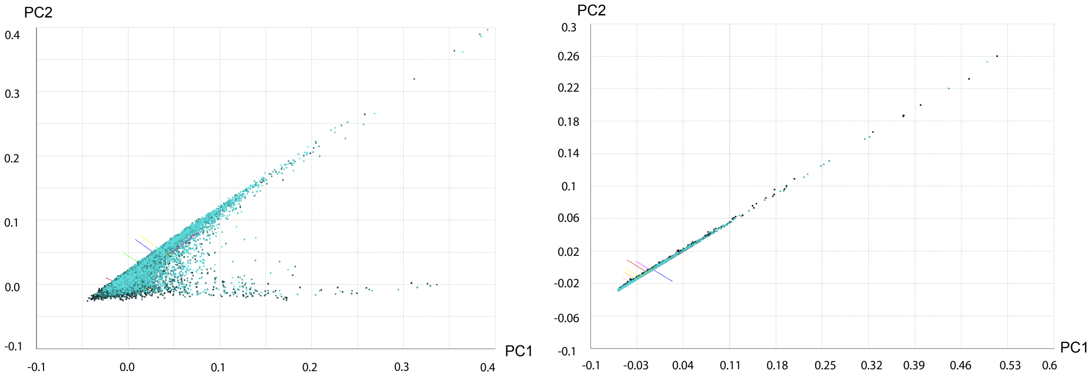


Figure 3.9: Similarity scoring of phenotypic profiles derived from cell division disruption. The impact of single gene knockdown experiments in cell populations is scored for the entire time course (spanning one cell cycle, approximately 50 time points) using (a) the averaging scoring scheme; (b) the lower bound of Wilson score confidence interval method. The nodes are colored according to these two schemes for every phenotypic layer on a white-to-red scale. In this case, we show the full set of 1067 genes and each one is placed in the same position on all layers, in a grid layout. The figure also displays pop-up windows containing the temporal profiles of genes *RANP3* and *INCENP* represented as line plots for all phenotypes (”polylobed” in green), along with the genes’ position on the polylobed layer. These plots can be obtained by clicking on the respective nodes. Both genes display increasing effect on the cell populations adopting the ”polylobed” morphology with time. The figure is reproduced from (Secrier et al., 2012).

RANBP3 (a RAN binding protein [Ensembl:ENSG00000031823]) in Figure 3.9 show that the suppression of both genes results in a high prevalence of the ”polylobed” phenotype. However, the two scoring schemes show higher signal for *INCENP* in (a) and lower in (b) compared to *RANBP3*. A closer look at the time course curves depicting the measurements for the ”polylobed” morphology yields a slower rising curve and a lower average for *RANBP3*. This explains why it scored less according to the averaging scheme in (a), but its higher peaking signal at the end of the time course was recovered by the scoring scheme in (b) when some of the noise was balanced out. The two scoring schemes thus serve well in identifying global patterns in time course datasets, bringing out even subtle differences in measurements between genes, and are best used complementarily.

3.3.2.4 PCA analysis

To get further insight about the structure of the data, I used Arena3D to perform PCA. The resulting populations are shown in Figure 3.10a for superimposed populations corresponding to every time point of the experiment and in Figure 3.10b for superimposed populations corresponding to all phenotypes. The populations are plotted in the directions with the largest variance in the vector space, which contain the dynamics of interest. For the populations of different time points, the signal is strong with high variance, while at the same time containing some noise. The populations depicting distinct phenotypes show very high signal-to-noise ratio individually, but also high redundancy among themselves. There was no distinct grouping in either case, indicating that global patterns in the dataset of all knockdown events are rather uniform and difficult to separate into modules.



(a)

(b)

Figure 3.10: The first two principal components computed for populations corresponding to (a) all time points and (b) all phenotypes, superimposed. Populations are color-coded from black to cyan from time 0 to 50 (a) or in the following phenotypic order: mitotic delay, binuclear, polylobed, grape, large, dynamic, apoptosis (b).

If we look at the top 10 scoring genes according to the average knockdown score for each phenotype and perform PCA on the corresponding vectors, we obtain in all cases that more than 90% of the variance is explained by the first principal component, with the 10 genes having rather similar contributions to this variation (not shown). For the rest of the principal components, however, the proportion of gene contribution changes, such that we can pinpoint genes with higher impact on a particular phenotype accounting for more of the variation within the phenotype. Figure 3.11 shows the PCA results for the ”polylobed” phenotype. Results are similar, alas with different genes, for the other morphologies. Gene *RANBP3*, discussed also earlier, stands out as contributing to a higher proportion of the variance in the third principal component. Genes *C19orf54*, *FAM92B*, *DCLRE1C* recur as impacting several principal components. While the contribution to the overall variation is very small, their recurrence indicates them as potentially interesting targets to further investigate, especially since some of them are of unknown function (e.g. *C19orf54*). Therefore, this method helps dissect the sources of phenotypic dynamics obtained upon single gene perturbations.

3.3.2.5 Pairwise phenotypic changes upon knockdown illustrate paths in phenotypic progression

Figure 3.12 depicts some of the obtained patterns when plotting coupled gene impact trajectories. By analyzing them, one can pinpoint the moments throughout the cell cycle when these phenotypes are more prevalent, as well as infer how they compare in terms of overall appearance and duration in the cell populations. ”Large” appears as a rather variable phenotype, as the trajectories take all directions throughout the time course: it appears sparsely in cells. For many of the cases (upper right quadrant) it has constantly increasing values, indicating persistence in the phenotype. There are, however, several cases when the trajectory goes down for at least one of the gene knockdowns (all the other quadrants). This suggests that the cells adopt the ”large” morphology in the beginning but then another quickly succeeds it. In contrast, ”binuclear” and ”apoptosis” exhibit clear tendencies for increase in the beginning of the time course, which indicates that cell populations adopt these phenotypes early and are arrested in them for a certain period. The upward direction is more consistent in the ”binuclear” case, indicating a strong incipient signal: it is a phenotype that appears early in

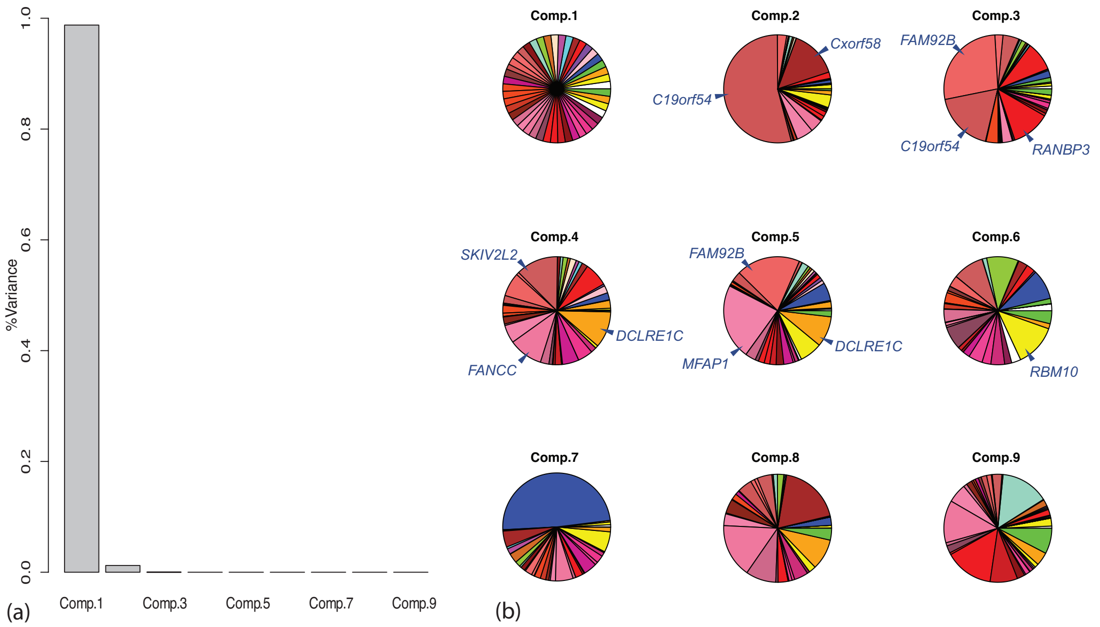
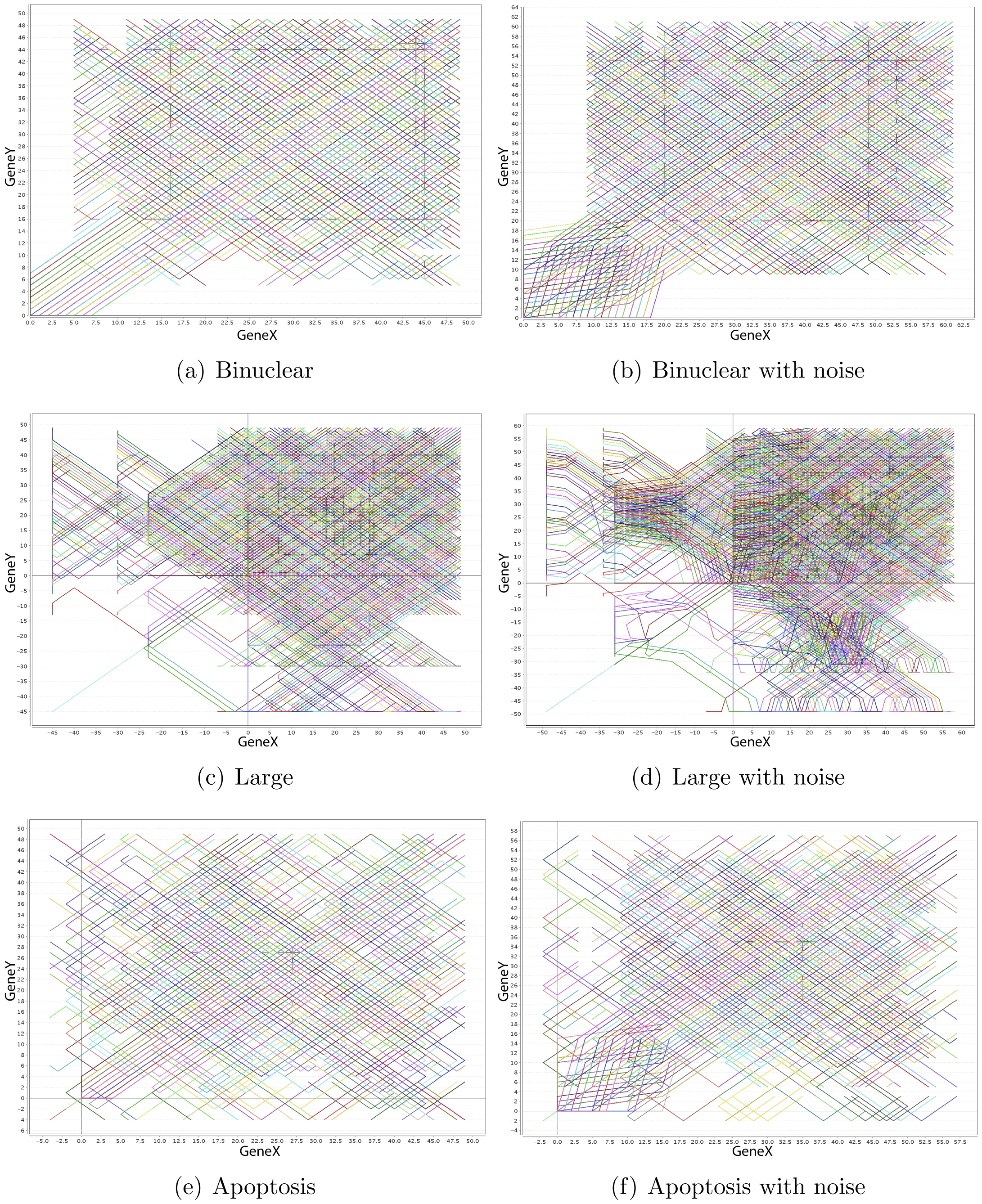


Figure 3.11: PCA of the top 10 scoring genes for the ”polylobed” phenotype. (a) The amount of variance explained by each principal component. (b) Projections of the eigenvalues for the first 9 principal components. The overall variance of the knockdown score distribution is explained by the magnitude of the projections in each principal component. The individual slices in the pie charts correspond to contributions of individual genes to the variation explained by each principal component. The genes with highest contributions are indicated in blue.

cells. At later time points the trajectories diversify directions, but we still see a more structured signal in ”binuclear” and ”apoptosis” compared to the ”large” phenotype.

The perturbations in the trajectories introduced by artificial noise make the time points with the highest impact changes promptly visible in Figures 3.12b, d and f when compared to the unperturbed plots. The most sudden changes in knockdown impact are in the beginning of the time course for the ”binuclear” and ”apoptosis” phenotypes, whereas ”large” displays stronger changes midway through the cell cycle. The changes are also more abrupt in the last case.

Even though the feature was eventually excluded from the application, the idea of visualizing trajectories of single or coupled outcomes in time offers an effective and fast way to visually identify patterns. Furthermore, the introduction



(a) Binuclear

(b) Binuclear with noise

(c) Large

(d) Large with noise

(e) Apoptosis

(f) Apoptosis with noise

Figure 3.12: **Left side**: pairwise derivatives of knockdown vectors corresponding to pairs of genes that exhibit impact changes greater than 50% at some point during the experiment. **Right side**: the same representation, but with changes greater than 80% highlighted by 5-fold perturbation. Each plot depicts the effects on one phenotype. For each phenotype all gene pair trajectories are shown, where each gene knockdown leads most prominently to the respective phenotype. A different color is randomly assigned to every trajectory. The paths represent results of 100 gene pair combinations.

of noise to perturb these trajectories in order to make sharp changes more visible is an approach than should be further explored.

3.4 Discussion

3.4.1 Summary of results

The versatility of Arena3D visualization and analysis methods has allowed extensive analysis of different aspects of phenotypic emergence. We have gained insight into the extent of *Nanog* perturbation effects from epigenetic down to protein regulation level. The analysis has suggested rewiring routes in the background regulatory network. It also identified correlated effects of this perturbation on genes in multiple regulation steps. Arena3D was also used to look into similarities among several morphologies of defective cell division, a process whose disruption has deep implications in development and disease. The impact of different knockdowns was investigated and effective targets identified. Rare and prevalent phenotypes may have consequences in disease severity or symptome instantiation. We also suggested a potential new role of gene *LSM14A* in cytokinesis, which needs to be verified experimentally. While not without caveats, the approach employed by this tool has proved there is a good potential in combining a series of visualization principles for pattern identification in holistic as well as more focused biological studies. I discuss these aspects in further detail in the following paragraphs.

3.4.2 3D versus 2D visualization

The use of 3D for visualizing anything else besides protein structures has always aroused debates in the visualization and biological communities (Tavanti and Lind, 2001; Tory et al., 2006). The main aspects of concern are occlusion and misinterpretation of size when comparing objects. These are details that the user should always be aware of when using 3D representations, but one should also note that proper design techniques can reduce and even eliminate them.

I believe that the visualization concept presented in Arena3D can be very useful in certain biological problems, if applied properly. First, 3D alleviates the problem of cluttering encountered in large 2D pathway representations, as shown previously in the SBML model visualization example. Moreover, there are cases when a 3D visualization is clearly more optimal for categorizing the data, as well as avoiding edge overlap and confusion that arises from that. I demonstrate one such example in Figure 3.13. Here I show how the interpretation in visualizing a kinase-substrate network as described in (Tan et al., 2009) improves with the use of an additional dimension. The initial figure in the paper renders the identification of several interaction partners difficult because the links often intersect. The proteins are colored according to their impact in certain diseases, but they are rather scattered than grouped into categories and this impedes interpretability. Figure 3.13b already improves the view by showing a clear separation between substrates and kinases, while Figure 3.13c creates a completely new perspective using clustering. Proteins can be clustered according to different criteria, while still separated into kinases (top, hidden layer) and substrates (grey layer). The connections are well visible, also by rotation. This view simplifies considerably the initial figure.

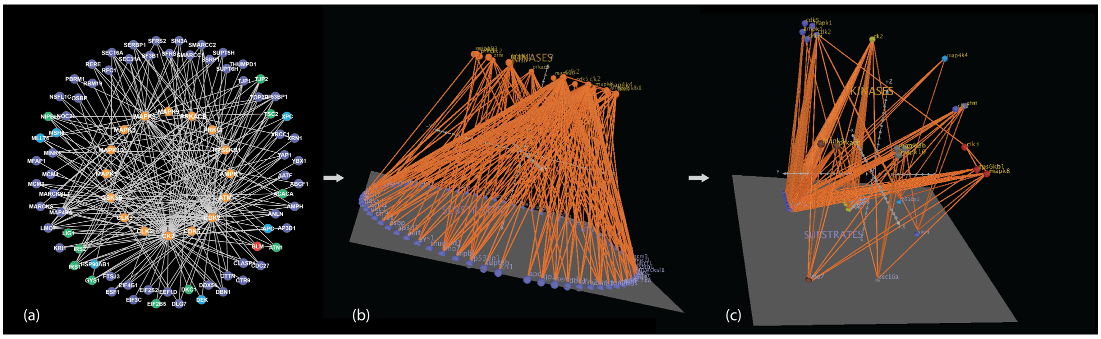


Figure 3.13: Kinase-substrate network represented in 2D as described in (Tan et al., 2009) (a), 3D without clustering (b), and 3D with clustering by kinase and substrate type (c). Part (a) was adapted after Figure 5c of the paper (Tan et al., 2009). The (b) and (c) representations were done in Arena3D. Orange nodes denote kinases; all other nodes denote substrates (for these, different colors depict associations to diseases). In (a), many of the links overlap and the connection partners are difficult to track. Also, there is no easy way to understand how proteins belonging to different categories (depicted in different colors) relate. An improvement in the clarity and interpretability of the figure can be seen from left to right. The authors might have benefited from using 3D in this case. This situation illustrates the advantages of 3D over 2D in some visualization tasks.

This brings us to the second observation: the layered approach is useful for separating categories of biological information for a better structured display, such that the connections between different classes of entities are easily visible. In the case of multiple phenotypes this provides a neat separation into categories, and following time course profiles in this manner is easier. The layered view also helps with the interpretation and may lead to discovery of new indirect connections between processes. Furthermore, while the 3D figures in the published papers might occlude some details, this is solved in real life through interactivity: the user can usually rotate and move things such that much more insight is gained by using the application to answer a specific question than by looking at a static picture.

3.4.3 Comparison to similar visualization tools

Arena3D combines several visualization concepts that confer it several points of advantage compared to other visualization tools for time course, network and gene expression data: data integration, multi-layered 3D layouts, clustering, gene tracking, detailed analysis of time course profiles. Regarding network layout, the use of several layers in 3D to both integrate and separate different levels of biological information (different phenotypes, tissues etc.) and the availability of different clustering algorithms renders a more flexible and intuitive experience compared to software like BioLayout Express(3D) (Theocharidis et al., 2009) or clusterMaker (Morris et al., 2011). While color is used to denote changes also in tools like VistaClara (Kincaid et al., 2008), SpotXplore (Westenberg et al., 2010) or GATE (MacArthur et al., 2010), Arena3D can additionally assign an overall score for the temporal profile of each gene, which in turn enables comparison of networks over time. Furthermore, Arena3D can handle both time series and non-time series heterogeneous datasets. It does not require a hierarchy and can process larger data compared to Pathline (Meyer et al., 2010a). However, heatmaps cannot be plotted and it offers fewer clustering options than clusterMaker. It can nevertheless deal with a wider variety of biological and non-biological data types. The individual gene/protein tracing is another feature missing in many of the similar tools. The key aspect is combining structured data integration, dynamic visualization approaches and interactivity for both a global as well as a focused comparison of phenotypic outcomes at multiple time points.

3.4.4 Future development and conclusions

The type of analysis showcased in the previous section is applicable to any RNA interference or microarray experiments, as well as any other types of studies with multiple time points and phenotypes. I have shown how Arena3D can enhance interpretation of medium and even high-throughput screen results by investigating whether patterns of similar gene disruption effects in time reflect related phenotypes. Once such complementary genes are identified, subpopulation profiling can provide more insight into genotype-phenotype linked differences. This is interesting to study, because patterns of cellular heterogeneity may reflect deeper divergence of the underlying regulatory networks. Also, similarity scoring techniques can build up a collection of sensitive/resistant phenotypes in relation to genes. This gives further intuition about the impact of stress, mutations or changing environmental conditions on the functioning of the cell.

The Arena3D framework can be easily extended to cover further aspects in temporal data visualization and dynamic pattern extraction. For better grouping of genes into modules of congruent time-resolved patterns, we could for instance cluster them using an approach that combines PCA, scoring schemes and self-organizing maps (SOMs) (Kohonen, 1982), similar to the one proposed in Figure 3.14. By performing PCA, the number of clusters that best represents the data can be revealed. This number can be employed then in the SOM clustering algorithm, which requires specifying an initial number of clusters. The SOM algorithm would be applied as in (Quackenbush, 2001), based on combining distances between vectors with scoring schemes for gene similarity and/or percentile reduction of entropy scores (as described in (Sangurdekar et al., 2006)). Clustering maps obtained for each phenotype would then be visualized in Arena3D and compared to discover gene-driven phenotypic differences.

The use of different geometric shapes or glyphs to represent the nodes in the networks might also be investigated in the future. Replacing the current plain spheres, these symbols could help the user immediately identify the enzymes de-

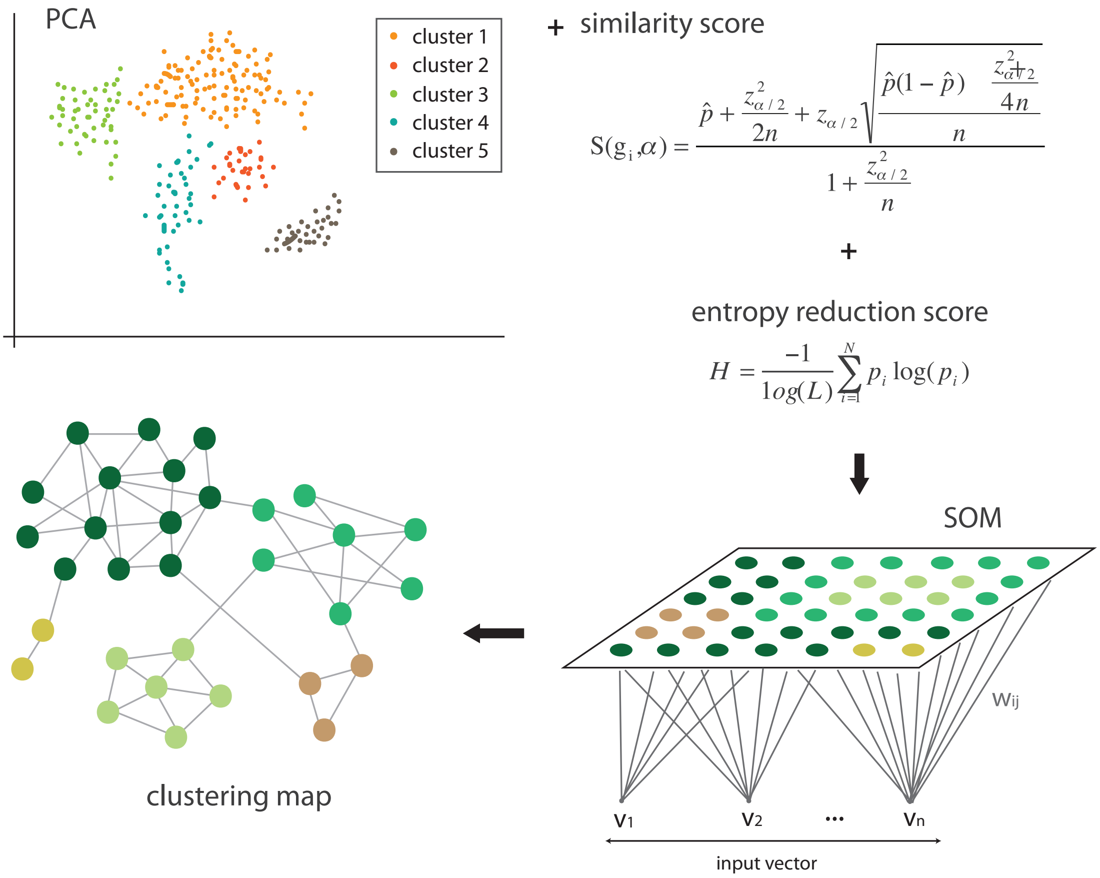

PCA
● cluster 1
● cluster 2
● cluster 3
● cluster 4
● cluster 5
+ similarity score
$S(g_i,\alpha) = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \frac{z^2_{\alpha/2}}{4n}}}{1 + \frac{z^2_{\alpha/2}}{n}}$
+
entropy reduction score
$H = \frac{-1}{1og(L)}\sum_{i=1}^{N} p_i \log(p_i)$
SOM
Wij
V1
V2
...
Vn
input vector
clustering map

Figure 3.14: Proposed method for identifying gene modules based on their time course associated values for every phenotype: combining PCA results and similarity scores into a classification using SOMs, which would result in clustering maps of genes for every phenotype. The similarity score formula is the one described in section 3.2.2.2. The entropy reduction score is taken from (Sangurdekar et al., 2006).

picted, similar to the approach used in BioCarta pathway maps (http://biocarta.com/). Thus, iconic depictions of different categories of enzymes, RNAs or transcription factors would create a better user experience by moving a step away from abstract models.

To better understand the context in which certain cellular events take place, as well as identify similarities in mechanisms of action of genes, links with external databases will be provided. I particularly wish to highlight the ArrayExpress (Parkinson et al., 2011), BioModels (Le Novère et al., 2006), Gene Ontology (GO) (Ashburner, 2000), KEGG (Ogata et al., 1999), COSMIC (Forbes et al.,

2011) and OMIM (Hamosh et al., 2005) databases as possible future integration targets. ArrayExpress constitutes an excellent resource where many time course and multiple-phenotype experiments are deposited. BioModels contains mathematical models that could be retrieved in the SBML format into Arena3D for visualization and analysis. Gene ontology and pathway information would help in annotating proteins in the loaded dataset, as well as identifying links between them and involvement in biological processes. Also, linking to diseases from time series data might allow the identification of crucial time points in disease onset and development. Finally, the incorporation into Garuda should help further dissemination and integration with other biological tools. We envisage that following these additions Arena3D will become an even more useful tool to complement the ensemble of biological methods that study hotspots of biosystem robustness.

This also stems from the structured visualization approach that Arena3D uses to represent dynamic patterns. It has become increasingly clearer in the past years that the differential regulation of processes not only among different organisms, but also within a single organism imposes a switch from a global to a time-resolved, tissue specific view when analyzing biological systems (Lopes et al., 2011). This stratification is well reflected in the principles of visualization in Arena3D, with its ability to structure and compare tissue-level expression in a temporal context. We believe this asset makes Arena3D particularly well suited to study these regulatory aspects. Looking at differential phenotypic outcomes based on tissue or organ specificity will provide insights into developmental patterns and functional complementarities of a system. This is a first step in improving prediction of regulatory specificity, eventually leading to differential diagnostics and treatment of disease. Before this, a consistent way of identifying common factors in disease outcomes is also needed, such that linking phenotypes temporally becomes an asset. In the following chapter I present a visualization tool derived from the same principles, which takes the analysis one step further in integrating and connecting phenotypic outcomes.

Chapter 4

Connecting time-resolved phenotypic landscapes with PhenoTimer

4.1 Description

In the previous chapter, I have presented different visualization strategies for comparing phenotypic outcomes in a temporal context and dissecting some details of their underlying genetics. An equally interesting topic is how these phenotypic features evolve as a result of similar mechanisms of regulation at transcriptional, translational or pathway level. The question I want to extend in this chapter is the following: what common molecular determinants do two phenotypes have and how are these two phenotypes related in a temporal context? Ultimately, answering this could have implications in tracking developmental stages and finding common strategies for disease prevention or treatment, in a temporally stratified manner.

To address this subject, experimental results must be combined with suitable visualization tools that can help make sense of the intricate relationships within the data, especially in the context of heterogeneity and multiple dimensions: space, time, tested conditions. Despite the affluence of incoming time-resolved large scale data (Furusato et al., 2008; Hitchler and Rice, 2011; Lefrancois et al., 2010), linking and visualizing such complex processes continues to be a bottleneck in systems biology.

In this chapter I introduce a novel 2D/3D visualization approach that links phenotypic outcomes through time. The software, entitled PhenoTimer, enables dynamic integration of time-course medium or high-throughput data coming from gene expression or imaging screens with networks and functionality information. The main novelty consists in visualizing relationships between phenotypes as arc projections in two- or three-dimensional space, along with dynamical highlighting of genetic regulation within networks. It is applicable to any dataset where time-course measurements associated to genes or proteins for several biological variables (phenotypes, pathways, diseases etc.) are available.

I illustrate the effectiveness of this visualization approach with different biological applications: phenotypic transitions of cell division defective-cells from the genome-wide knockdown study on cell cycle essential regulators by (Neumann et al., 2010), transcription events throughout the cell cycle linking cancer pathways and similar mechanisms of gene regulation upon acute administration of addictive drugs. The wide range of applications of this tool enables tackling important questions in the area of cellular regulation. The results have been summarized in a paper, which was under review at the moment of thesis submission.

4.2 Software design and implementation

PhenoTimer integrates a combination of 3D and 2D projections to track connections between phenotypes through time. These connections delineate genetic factors that have a shared influence for a certain phenotypic outcome. The tool was developed with the purpose of looking at the phenotypic landscape as it evolves with time and identifying patterns that might explain: (1) how two phenotypes relate to each other, (2) how the progression of a disease occurs and (3) what the common mechanisms driving disease outcomes are. Additionally, network integration helps regard these patterns in the context of systems-level regulatory circuitry.

4.2.1 Visual depictions

PhenoTimer uses a combination of graphical representations to explore time-course phenotypic data, as described in the following subsections.

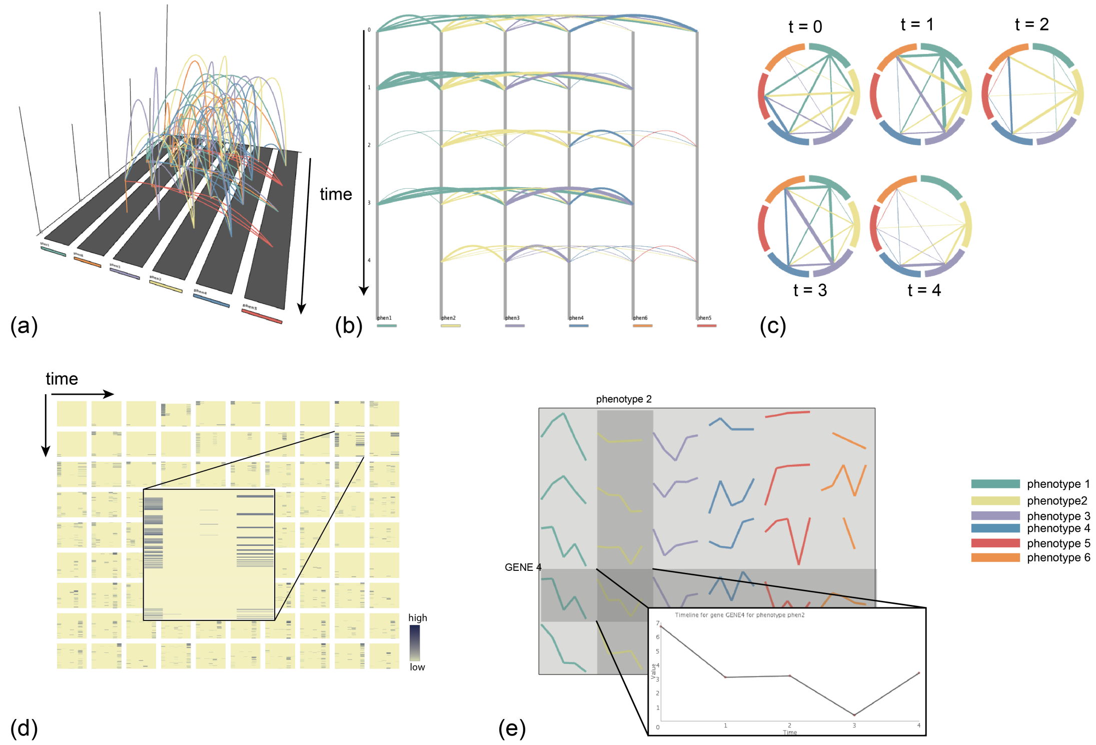


Figure 4.1: The different visualization modes available within the PhenoTimer software. (a) 3D arc view: links between phenotypes are represented as arcs connecting the respective phenotypic lanes in three dimensions. Each phenotype has a color associated, indicated in a rectangular box at the end of the respective phenotypic lane. These colors are used in the arc representations to indicate the directionality of the connection, i.e. the end point phenotype. The height of the arcs is proportional to the number of genes involved in the respective link. Connections are shown through time and bar charts of time-course values can be displayed in parallel. (b) 2D arc view: a similar view to the one described in (a), only displayed in 2D. In this case, the width of the arc instead of its height is used to denote differences in the number of involved genes. (c) Circular view: the phenotypes are arranged as segments of a circle and the connections between them are rewired at every time point. (d) Heat map view: a color gradient from yellow (lowest) to dark blue (highest) is used to depict the value associated to each gene at every phenotype. Lines are genes; columns are phenotypes. A heat map is produced for every time point, and the user can zoom into every single one by hovering the mouse over it. A click selects it for a full screen view. (e) Line plot view: the timeline of gene-associated measurements is plotted for every gene and phenotype. Hovering over a plot brings up a pop-up window with a zoomed-in version of the plot.

4.2.1.1 The arc representations

The main novelty of the tool consists in using different "arc" depictions to link phenotypes or other variables based on some common genetic mechanism. The meaning of these connections between phenotypes is defined based on the dataset and the question that the user wants to answer. For instance, a link between two disease phenotypes could mean that these two diseases share some misregulation mechanism for specific genes. A link between two pathways could underline that there are some proteins active at similar levels in the two pathways at that specific time point. The connections can also represent transitions from one phenotypic outcome to another, for instance in different stages of a disease.

There are three available arc views, as shown in Figure 4.1 : (a) 3D arc view, with arcs joining phenotypic lanes in three dimensional projections through time; (b) 2D arc view, a flattened representation of the previous view, where arcs are drawn in two dimensions joining parallel vertical phenotypic lines; and (c) circular view, with arcs joining phenotypic segments arranged in a circle, similar to representations used by Circos (Krzywinski et al., 2009) or TVNViewer (Curtis et al., 2011). Arcs have been modelled using Bezier curves (Chaudhuri and Dutta, 1986).

A series of arc attributes enriches the informational content that can be obtained by observing the arcs. The height (in 3D) or width (in 2D) of the arc is proportional to the number of genes/proteins involved in the respective connection. In a global view, this helps with comparative assessment of gene regulation impact at different time points and for different pairs of phenotypes.

Color is used to indicate directionality of transitions between phenotypes, in the cases where this is valid and known. The color of an arc will match the color of the phenotype towards which the transition occurs (i.e. the end phenotype). A monochrome option is available in the case when the directionality of the connection is irrelevant for the biological problem. Different color schemes are available for the user, including color-blind safe. The schemes have been chosen to abide by rules of good color combinations for visual balance and include the following: "Standard", "Paired", "Pastel", "Strong", "Color blind", "Single color", all accessible from the "Change color scheme" submenu of the "View" menu. These schemes have been taken from http://colorbrewer2.org/. Additionally, the user can customize colors for every phenotype.

The arcs in 3D mode are interactive: by clicking on an arc, a pop-up is displayed with the names of genes/proteins involved in the respective connection, along with functionality information, if available (if the user has loaded GO files), as shown in later in Figure 4.9 of the *Results* section.

In the circular view, the plot is generated dynamically and connections between phenotypes are rewired for each time point. This last option offers a global view of all circular plots for all time points, as well as a zoomed-in view of a single plot occupying the full canvas. Clicking on a plot in the global circular view zooms into the respective plot, and they can also be visualized sequentially for every time point.

4.2.1.2 The heat map representation

Heat maps are used to summarize the landscape of gene-associated values for every phenotype. They are generated separately for every time point (see Figure 4.1d). In the heat map, rows correspond to genes/proteins, columns correspond to phenotypes and the color maps to the value for the respective time point on a scale from yellow (lowest) to dark blue (highest). The user can change the color gradient. The rows and columns are grouped using hierarchical clustering to reflect similarities among genes and/or phenotypes.

There are two display options: the user can visualize all heat maps for all time points, or a specific one for the time point of interest, as well as go through all of them one by one, with the display being updated at every time point. In the global view, hovering over a plot with the mouse pops up a window with a zoomed-in version of the heat map. Clicking on it will display it on the full canvas. There is also a button available in the GUI to switch from global to single heat map view.

In the single heat map view, hovering over a specific tile will highlight the corresponding gene and phenotype.

4.2.1.3 The line plot representation

Line plots of time series profiles for all genes and all phenotypes can be drawn to visualize global trends (Figure 4.1e). All plots are merged on the canvas in the same order as the tiles in the heat map, with rows corresponding to genes and columns to phenotypes. Hovering over a specific plot with the mouse opens a pop-up window with a zoomed in version of that plot, similar to the heat map behavior.

To switch between view modes, the user must access the ”Mode” submenu from the ”View” menu.

4.2.1.4 Bar charts

A bar chart can be loaded by the user and will be displayed on the time axis in the 3D arc view in parallel to the phenotypic links, as shown in Figure 4.1. The bar chart is user-defined and typically contains parametric values associated to each time point. The number of values loaded must be equal to the number of time points in the dataset and they must be listed one per line. These values can signify anything, depending on the biological problem (for instance, the total number of expressed genes at a particular time point). Thus, it is the user's responsibility to decide whether the data loaded has feasible meaning in the given biological context. The bar chart can only be displayed in the 3D arc mode.

4.2.1.5 Networks

Different networks can be visualized dynamically along with the 3D and 2D plots: GO, pathways, PPI, metabolic and other types of networks of the user's choice, as well as networks where connections between genes signify participation in the same phenotypes. Networks can be generated and displayed in parallel with the graphical views mentioned previously in three ways: (a) by loading a network (GO files have a different format compared to other network files and thus must be loaded using the ”Gene Ontology” option); (b) by retrieving a network from the STRING database (Szklarczyk et al., 2011; von Mering et al., 2003), as described below; or (c) by having the application automatically generate one at each time point.

Network generation

Currently, GO networks are not retrieved automatically by PhenoTimer, but must be loaded by the user. Any program can be used for this purpose. The necessary files to be loaded are: (1) an enrichment file that specifies the biological processes, molecular functions or cellular components in which genes in the dataset of interest are enriched; and (2) an interaction file which details the structure of the GO tree for the specific terms, i.e. how the GO terms are related (an interaction meaning a hierarchical relationship). The latter is optional, but if not loaded the GO terms will appear disconnected. To load the GO files, the user must select the corresponding options in the ”Load gene ontology” submenu of the ”File” menu. After loading, the option to display the network must be selected from the ”View” menu. Similarly, a PPI or other type of network file loaded by the user would have to specify the interaction list of the network nodes. Other networks are loaded using the ”Load network” option. More details on the format of the files loaded by the user can be found in section 4.2.6.1.

Another option featured by the application is the automatic generation of networks from the data. At every time point, a network of genes is calculated and displayed, where nodes represent genes and they are linked if the respective pair of genes is involved in the same phenotype. The thickness of the links is proportional to the number of phenotypes shared by two genes. In this context, ”shared” refers to any similar regulation mechanism of the two genes that leads to the same phenotype. For display feasibility and performance reasons, only networks with less than 500 nodes will be generated and displayed.

It is important to emphasize here the dynamic nature of these networks: the networks change at every time point to reflect the genetic regulation underlying the phenotypic links at the particular time point. In the case of GO networks, different GO terms will be highlighted in red at each time point to indicate the functions of the genes that are involved in connections at the respective time point. For the network retrieved from STRING, genes involved in connections at a specific time point are highlighted in red. The dynamically generated network is updated at every time point.

Network layout

PhenoTimer uses a force directed layout (Fruchterman-Reingold) for network display. This helps optimize the use of space and minimize overlap in the network. For GO networks, this layout was kept instead of the hierarchical one usually used for such representations because of display reasons (big trees are difficult to constrain in a relatively small portion of the canvas) and because the focus in this visualization is on the individual terms highlighted rather than on the way they relate to each other. The nodes in the network can also be moved around to correct any display suboptimality.

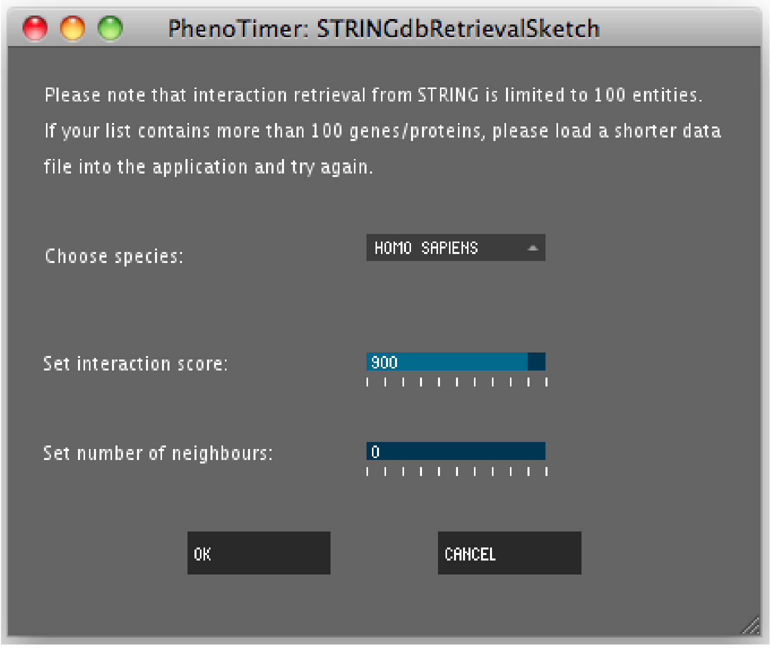
4.2.2 Data integration

PhenoTimer integrates data from different databases as described below.

4.2.2.1 Reconstructing networks from STRING

PhenoTimer is able to retrieve connections from the STRING database (Szklarczyk et al., 2011; von Mering et al., 2003) for a given list of proteins. To access the option, the user must select ”Get network from STRING” from the ”Databases” menu. The interactions are retrieved on the fly, so this feature can only be used if an internet connection is available. Several filters can be set before making the query, including: species, number of neighbors and interaction score, as described in the paper (Szklarczyk et al., 2011). The species selection menu provides the additional option of retrieving networks of orthologous genes from other organisms.

Figure 4.2 shows the query window that enables the user to get interactions from the database. The retrieved connections will be displayed in the form of a network where nodes represent the proteins in the loaded dataset and the links their PPIs from STRING. To show the network, the user must select the ”Show loaded network” option from the ”Network” submenu of the ”View” menu.



PhenoTimer: STRINGdbRetrievalSketch

Please note that interaction retrieval from STRING is limited to 100 entities. If your list contains more than 100 genes/proteins, please load a shorter data file into the application and try again.

Choose species: HOMO SAPIENS

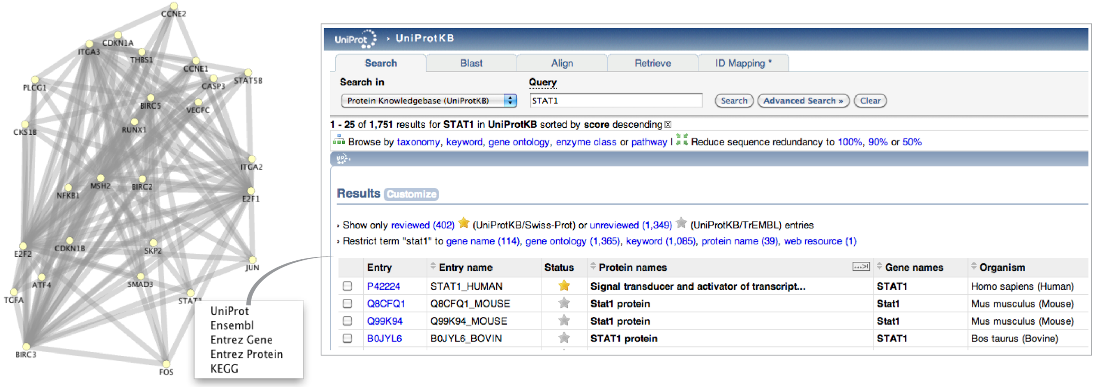
Set interaction score: 900

Set number of neighbours: 0

Figure 4.2: The window opened by the application for a query to the STRING database. The user can set different filters (species, interaction score and number of neighbors). Clicking OK will retrieve the interactions for the loaded dataset of proteins from the STRING database.

4.2.2.2 Linking out to other databases

Besides STRING, users can also obtain information from other databases for selected genes/proteins of interest by right-clicking the respective node in the network and choosing a database. This action will open up a window in the default browser of the user's operating system with the query result for the specific gene/protein. The following databases are available: UniProt (The UniProt Consortium, 2012), Ensembl (Flicek et al., 2013), Entrez Gene (Maglott et al., 2007), Entrez Protein (Coordinators, N. C. B. I. Resource, 2012) and KEGG (Ogata et al., 1999). For instance, clicking on protein *STAT1* in a network retrieved from STRING, loaded by the user or generated automatically from the data and choosing the option "UniProt" will open the page with the UniProt result(s) after searching for protein *STAT1* in this database, as in Figure 4.3. Thus, the user can obtain additional insight for the particular protein by further investigating the available information in the different databases directly from the application.



Search in: Protein Knowledgebase (UniProtKB)

Query: STAT1

1 - 25 of 1,751 results for STAT1 in UniProtKB sorted by score descending ☒

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%

Results

› Show only reviewed (402) (UniProtKB/Swiss-Prot) or unreviewed (1,349) (UniProtKB/TrEMBL) entries

› Restrict term "stat1" to gene name (114), gene ontology (1,365), keyword (1,085), protein name (39), web resource (1)

	Entry	Entry name	Status	Protein names	Gene names	Organism
□	P42224	STAT1_HUMAN	★	Signal transducer and activator of transcript...	STAT1	Homo sapiens (Human)
□	Q8CFQ1	Q8CFQ1_MOUSE	☆	Stat1 protein	Stat1	Mus musculus (Mouse)
□	Q99K94	Q99K94_MOUSE	☆	Stat1 protein	Stat1	Mus musculus (Mouse)
□	B0JYL6	B0JYL6_BOVIN	☆	STAT1 protein	STAT1	Bos taurus (Bovine)

Figure 4.3: Linking out to different databases. The user can right-click on a gene/protein name in the displayed network to query for it in the following databases: UniProt, Ensembl, Entrez Gene, Entrez Protein, KEGG. Upon selection, a browser window will open with the query results in the respective database.

4.2.2.3 Additional controls

Several other controls are available within the PhenoTimer GUI, to allow for flexibility and optimization in the visualization and analysis process, as described below. Their accessibility from within the application is shown in a later subsection in Figure 4.5.

Temporal tracking A time slider is available for visualizing the flow of phenotypic connections through time. The view is updated at each time point. The user can either visualize a single time point or trace all events up to the current time point. Pressing the key ”t” switches between these two options.

Thresholds Upper and lower thresholds can be set for the gene-associated values for every phenotype separately using range sliders. The visualization will be updated accordingly.

Time offset By default, the application visualizes connections between pairs of phenotypes at a certain time point. However, connections can also be visualized for a specific time interval, i.e. from time point t to time point $t + x$, where t is the current time point and $x \in \{0, T\}$, with T being the final time point. A slider entitled ”Time offset” is available for modifying the interval x.

Arc prominence The slider with this name allows the user to modify the height (in 3D) or width (in 2D) of the arcs, for optimal visualization.

Transparency The arc transparency can be modified to optimize visualization, especially in the case of arc overlap. A transparency slider is available for this purpose.

Single phenotypes The user can switch between viewing all connections between phenotypes and only those connections belonging to a particular phenotype of choice, as shown later in Figure 4.10. This is done by accessing the submenu "Phenotype" from the "View" menu.

Gene query Genes of interest can be queried and only links where those genes are involved will be displayed. The option can be accessed as shown in Figure 4.4. There is no limit to the number of genes that can be queried at the same time. This means that the user can ask questions like: (1) on which phenotypes do genes A, B and C have a shared influence? or (2) in which pathways are genes D and E jointly activated?

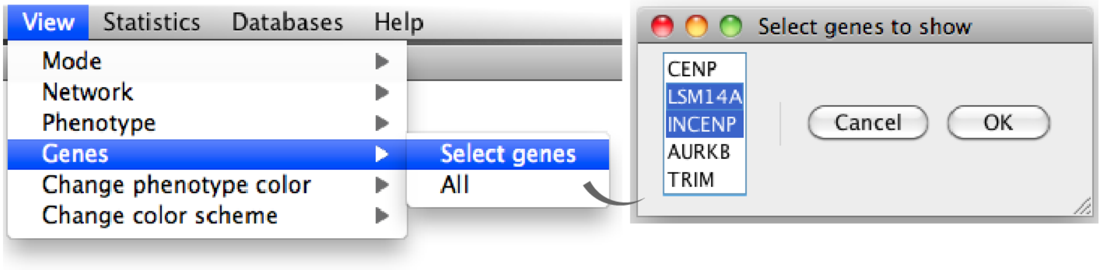


Figure 4.4: Querying for specific genes in the dataset.

4.2.2.4 Interactivity

Besides interactions with the arcs in the 3D mode and the zoom-in capabilities upon mouse hover or click events in the circular, heat map and line plot modes, the tool supports other types of interactivity. In the 3D arc view, the plot can be moved (by dragging the mouse while pressing key "m"), rotated (by dragging the mouse while pressing key ”r”), zoomed in and out (using the scroll button of the mouse). The network nodes can be dragged using the mouse to optimize display and avoid overlap. Additionally, pressing on a particular node highlights only the connections of that node to others in the network in red, as long as the button is pressed. Upon release, all connections are displayed again.

4.2.3 Statistical methods

4.2.3.1 Phenotypic ordering

For the 3D and 2D arc modes, phenotypic lanes are ordered so as to minimize cluttering and overlapping of connections between them. To achieve this, I have used an agglomerative hierarchical clustering algorithm (Hastie et al., 2009) that successively rearranges the phenotypes until the number of links between two adjacent phenotypes is maximized.

4.2.3.2 Heat map clustering

The rows, representing genes/proteins, and columns, representing phenotypes, of the heat map are clustered using the same algorithm as in the previous subsection, agglomerative hierarchical clustering. This allows to group similar genes and phenotypes. The clustering calculations can be performed using single, complete or average linkage and either Euclidean or Manhattan distance (Hastie et al., 2009). The default is complete linkage with Euclidean distance. The clustering may be recalculated at any point by pressing the ”Recompute clustering” button in the interface after having selected a method of choice.

4.2.3.3 Other considerations

It is important to note that for the loaded input data no normalization is performed. This option is not available because of the heterogeneity of the possible input data, which deems finding a suitable normalization method in each particular case impractical. If needed, the user should perform the normalization beforehand.

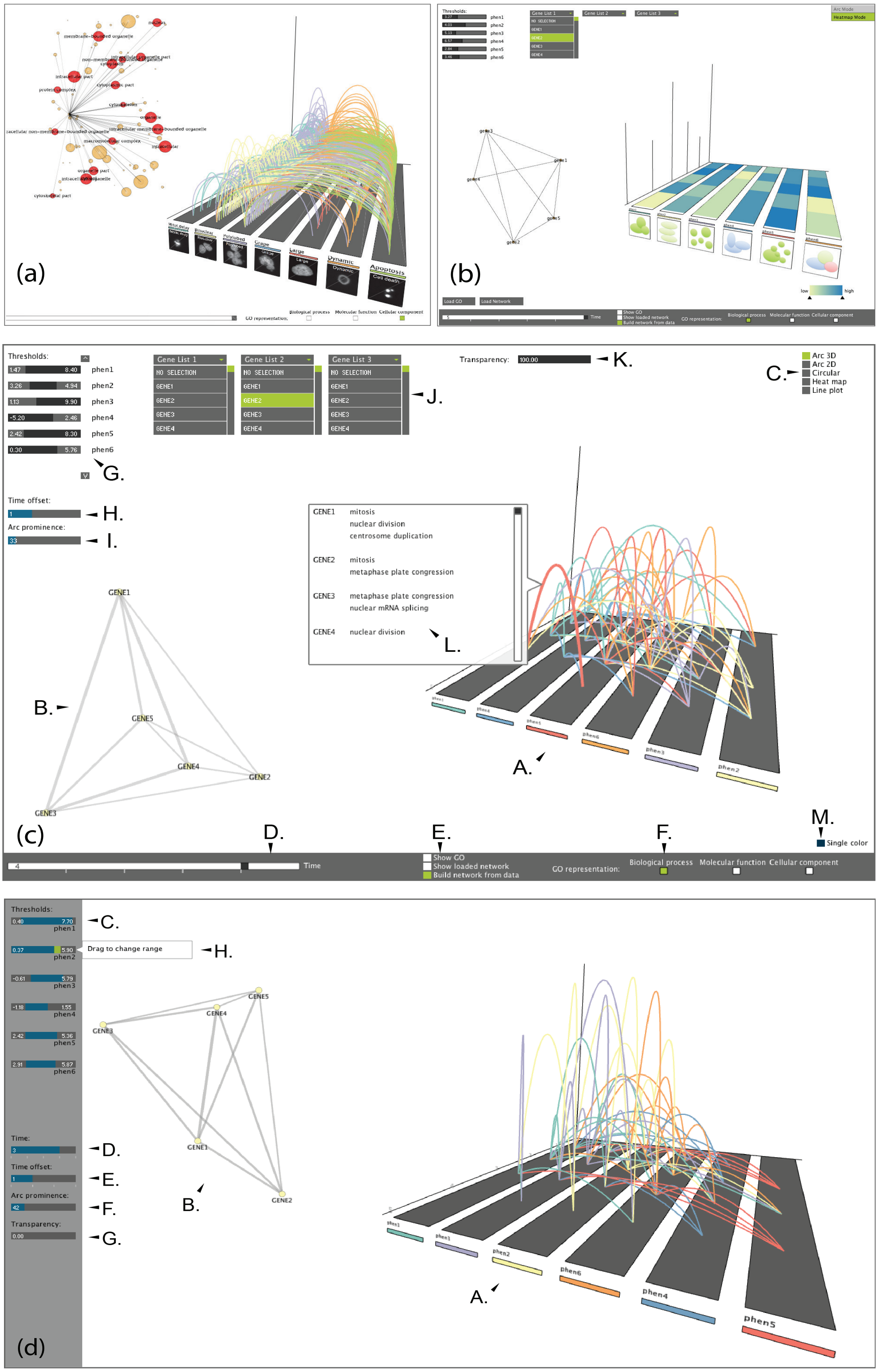
4.2.4 Stages of implementation and graphical user interface

The tool has gone through different stages of implementation, as summarized in Figure 4.5), and we believe it has considerably improved with time. Initially, PhenoTimer featured only 3D implementations. I conceived the 3D arc view as a starting point for pattern discovery in the dataset of phenotypic transitions between cell populations as described in (Neumann et al., 2010). I further enriched this view with GO network information, the networks being displayed adjacently in 2D (Figure 4.5a). At a later point, I added a 3D heat map for comparison of cell cycle transcription events in different organisms. The two views were combined into a fully functional software (Figure 4.5b). However, at the subsequent stage of implementation, the 3D heat map view was discarded for reasons of occlusion of perspective induced by three-dimensionality. This drawback is further discussed in section 4.5 of this chapter.

It was decided that a complementary approach of different 3D and 2D views would offer the best compromise for exploratory data analysis and visualization. Consequently, the tool was extended to comprise, besides the 3D arc view, the following other modes: 2D arcs, circular, heat maps and line plots, all in two dimensions, and new GUI controls were added (Figure 4.5c). The final stage witnessed a redesign of the graphical user interface (GUI) to comply with fundamental principles for effective information visualization (Bertin, 1984; Tufte, 2001), as shown in Figure 4.5d. The GUI shown in this last figure is the one currently available in the software and is much simplified compared to the previous version: the different sliders and buttons are grouped logically on the left side, to allow for more space for the graphical visualization on the right side. Many of the commands have been transferred into the menu of the application (not shown).

4.2.5 Workflow

PhenoTimer is suitable for visualizing medium or high-throughput datasets coming from gene expression screens (microarray, RNA-Seq etc.) or imaging experiments with multiple phenotypic outcomes and time points, provided some filtering is performed beforehand and the data is formatted into a specific space-


(a)
Biological process Molecular function Cellular component
GO representation:
(b)
Thresholds:
phen1
phen2
phen3
phen4
phen5
phen6
Gene List 1
Gene List 2
Gene List 3
NO SELECTION
GENE1
GENE2
GENE3
GENE4
Arc Mode
Heatmap Mode
low
high
Load GO
Load Network
Time
Show GO
Show loaded network
Build network from data
GO representation:
Biological process Molecular function Cellular component
(c)
Thresholds:
1.47 8.40 phen1
3.26 4.94 phen2
1.13 9.90 phen3
-5.20 2.46 phen4
2.42 8.30 phen5
0.30 5.76 phen6
G.
Time offset:
H.
Arc prominence:
I.
Gene List 1
Gene List 2
Gene List 3
NO SELECTION
GENE1
GENE2
GENE3
GENE4
J.
Transparency: 100.00
K.
C.
Arc 3D
Arc 2D
Circular
Heat map
Line plot
GENE1 mitosis
nuclear division
centrosome duplication
GENE2 mitosis
metaphase plate congression
GENE3 metaphase plate congression
nuclear mRNA splicing
GENE4 nuclear division
L.
B.
GENE1
GENE5
GENE4
GENE2
GENE3
A.
phen1
phen4
phen5
phen6
phen3
phen2
D.
E.
F.
M.
Single color
Time
Show GO
Show loaded network
Build network from data
GO representation:
Biological process Molecular function Cellular component
(d)
Thresholds:
0.40 7.70 phen1
C.
0.37 5.90 phen2
Drag to change range
H.
-0.61 5.79 phen3
-1.18 1.55 phen4
2.42 5.36 phen5
2.91 5.87 phen6
Time:
D.
Time offset:
E.
Arc prominence:
F.
Transparency:
0.00
G.
GENE5
GENE4
GENE3
GENE1
GENE2
B.
A.
phen1
phen3
phen2
phen6
phen4
phen5

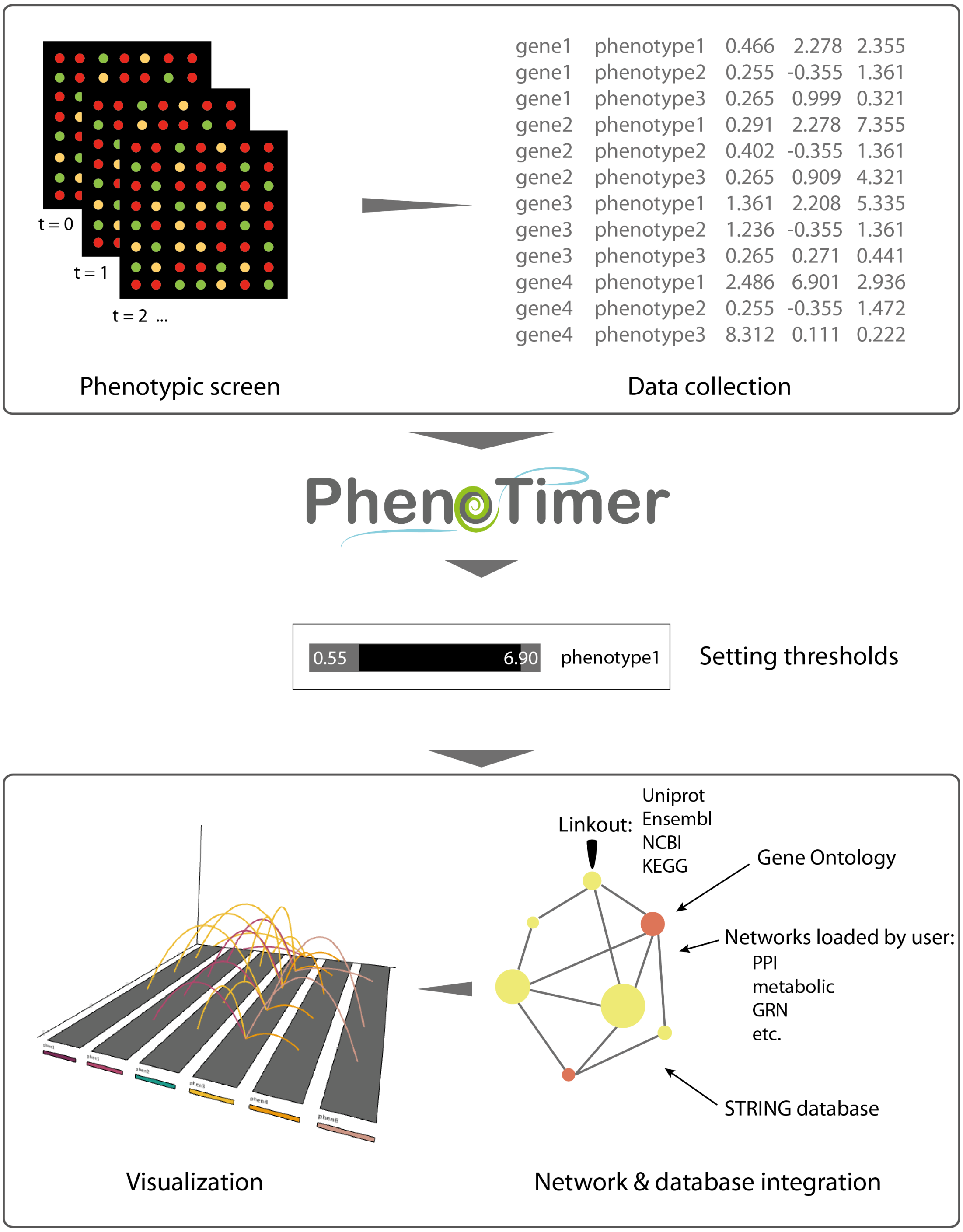
delimited text file, as described in subsection 4.2.6.1. Upon loading the data, the application produces the visualization instantly. However, filtering is necessary in order to observe patterns of connecting phenotypes. This is done by setting upper and lower thresholds for the values that each gene can have for a particular phenotype. For instance, one might wish to filter for only up- or downregulated genes common to phenotypic outcomes. One should keep in mind that the user makes the decision on the thresholds and thus they should either be calculated separately or reasoned biologically. Integrating dynamic network visualizations in parallel to different view modes and linking to databases as previously described helps to link the observed patterns to biological functions and identify interesting behavior or explain regulation of processes. The workflow is illustrated in Figure 4.6.

4.2.6 File formats

4.2.6.1 Input files

The multiple time point and phenotype files that can be loaded into PhenoTimer have a predefined space-delimited format. The first column must specify the genes/proteins, the second column the phenotypes, and the following columns contain time-course values for the specific gene and phenotype, all separated by white space. There is no restriction on how many time points can be loaded. For

Figure 4.5 *(preceding page)*: Stages of implementation of the PhenoTimer software: (a) 3D arc view accompanied by GO network. The GUI allows only time tracking and switching between GO term representations. An image can be loaded and visualized for each phenotypic lane, a feature that was later removed. (b) A 3D heat map view was added to the 3D arc view, another feature that was later removed. More functionality is available, like switching between views, setting thresholds, displaying networks. (c) The application now supports a combination of 3D and 2D arc views, accompanied by heat maps and line plots, and the more elaborate GUI gives greater flexibility for analysis: A. the plots of the five view modes are depicted here; B. network graphics space; C. controls to switch between view modes; D. time slider; E. switch between different network representations; F. select GO representation; G. phenotypic threshold sliders; H. time offset slider; I. arc prominence control; J. gene selection drop-down lists - up to three genes can be selected; K. arc transparency control; L. pop-up with gene information for a selected arc. (d) The same views as in (c), but with a GUI redesign: A. and B. like before; C. phenotypic threshold sliders; D. time slider; E. time offset slider; F. arc prominence slider; G. transparency control; H. indication pop-up. The rest of the functionality was moved to the application menu. For more details on the functionality, see subsection 4.2.1.



gene1	phenotype1	0.466	2.278	2.355
gene1	phenotype2	0.255	-0.355	1.361
gene1	phenotype3	0.265	0.999	0.321
gene2	phenotype1	0.291	2.278	7.355
gene2	phenotype2	0.402	-0.355	1.361
gene2	phenotype3	0.265	0.909	4.321
gene3	phenotype1	1.361	2.208	5.335
gene3	phenotype2	1.236	-0.355	1.361
gene3	phenotype3	0.265	0.271	0.441
gene4	phenotype1	2.486	6.901	2.936
gene4	phenotype2	0.255	-0.355	1.472
gene4	phenotype3	8.312	0.111	0.222

Figure 4.6: PhenoTimer workflow. The experimental data coming from medium to high-throughput microarray, imaging screens or similar measurements must first be structured into a specific input file format, similar to the one shown in the top panel. Loading this file into PhenoTimer will produce a visualization where all phenotypes are connected. The user must filter the connections according to some biological reasoning by setting thresholds to phenotypes. This will result in a clearer, patterned visualization of the data (bottom panel, left). One can afterwards enrich the informational content by integrating different types of networks, as well as linking out to databases (bottom panel, right).

each gene all phenotypes should be specified in the same order before moving on to the next gene. Comments can be added to the file using the "#" symbol. For an example, see Table 2 in Appendix C.

For GO networks, two files need to be loaded: an enrichment file and an interaction file. The enrichment file has separate columns specifying GO identifiers, GO term names, p-values for the respective GO terms and the enriched genes for each GO term, separated by "|" (see Table 3A of Appendix C). The interaction file lists pairs of interacting GO terms (depicted by their identifiers), one pair per line, as in Table 3B of Appendix C. These connections represent hierarchical relationships in the GO tree. Both files are tab-delimited. These files are not obtained by the software, but must be loaded by the user, who can employ any tool he considers suitable for this purpose.

Other networks, like PPI, metabolic, pathways etc., are loaded using a single file with format similar to the GO term interaction file: a list of protein interaction partners, one pair per line.

More examples are available at http://phenotimer.org/.

4.2.6.2 Export files

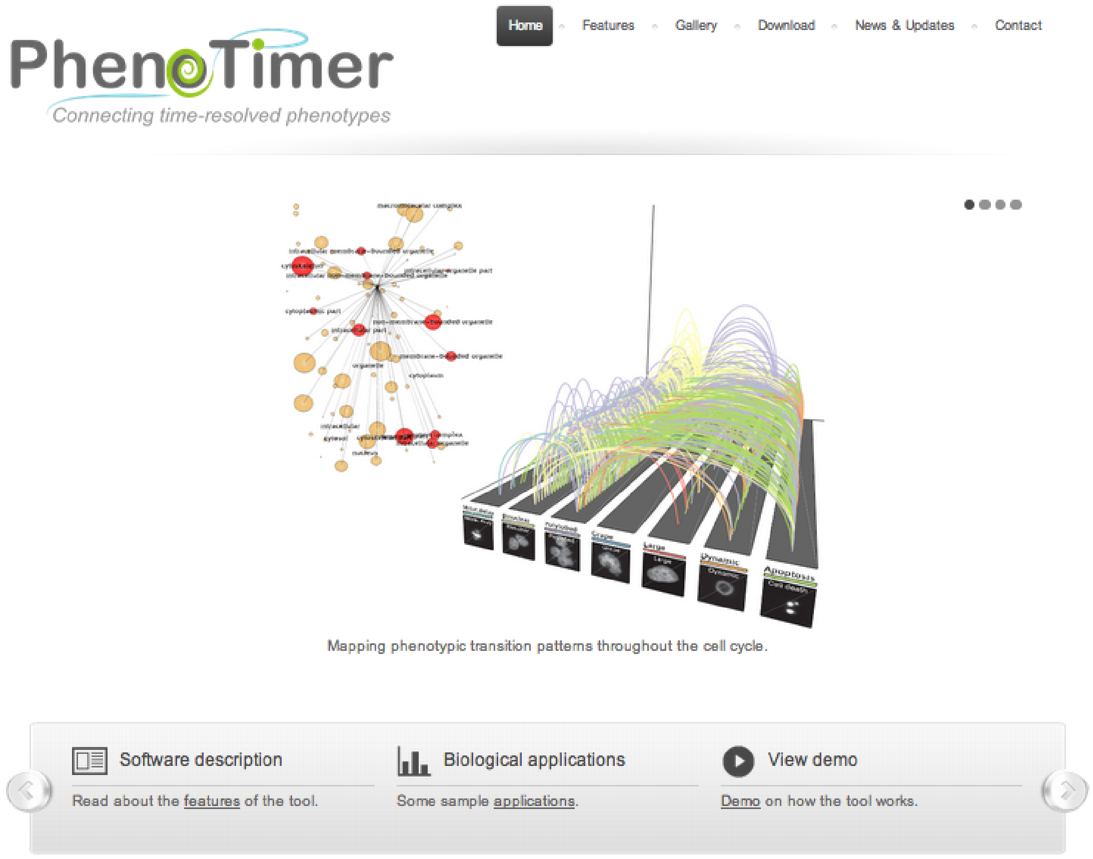
PhenoTimer can export the plots and networks as image files. This option is available from the "File" menu ("Save image as..."). The image will be saved with the extension given by the user in a specified location (as PNG, JPEG or TIFF file).

4.2.7 Technical specifications and availability

PhenoTimer was developed in Processing 1.5.1 (http://processing.org/), a Java-based environment with OpenGL integration. The stand-alone application and its source code are freely available for academic use under the GNU GPL v3.0 license at the website: http://phenotimer.org/. The website was also designed by me. For a screenshot of the homepage, see Figure 4.7.

PhenoTimer runs on Mac OSX, Windows and some Linux environments (Ubuntu 9.04, limited testing). The users need to install the Java Runtime Environment (http://www.java.com/) to run the tool. Macintosh users should also install the

JOGL libraries (http://opengl.j3d.org/).



Mapping phenotypic transition patterns throughout the cell cycle.

Software description

Read about the features of the tool.

Biological applications

Some sample applications.

View demo

Demo on how the tool works.

Figure 4.7: The PhenoTimer website homepage, located at http://phenotimer.org/.

4.2.8 Limitations

One limitation of the software is inconsistent performance among various Linux platforms. However, this stems from the lack of adequate support for the Processing environment for Linux rather than from any weakness of the software itself. Some bugs may arise that are platform and environment version-dependent. Besides this, PhenoTimer also exhibits performance limitations when loading datasets larger than the following dimensions: a few thousand genes x 50 phenotypes x 100 time points. The reasons for this limitation are related to both memory usage and physical visualization feasibility. The main memory performance limitation is the number of phenotypes. It is recommended not to exceed these limits for optimal functioning of the software.

4.3 Analysis methods

I used PhenoTimer for the visualization and analysis of three datasets with different biological setups, as follows: (1) the dataset of genome-wide knockdown effects on cell division processes from (Neumann et al., 2010), (2) a dataset of transcription events throughout the cell cycle and their conservation compiled from (Gauthier et al., 2008) and (Kasprzyk, 2011), and (3) a dataset of timed drug addiction effects in mouse, extracted from (Piechota et al., 2010). All datasets are available for loading into the application at http://phenotimer.org/samplefiles.html. The following subsections detail specific analysis methods used for each dataset.

4.3.1 Progression dynamics of mitotic defects

4.3.1.1 Data preparation

The first application used the time-course phenotypic data underlying the analysis performed in (Neumann et al., 2010). This data has already been described in previous chapters. I applied thresholds to the phenotypic scores assigned to every gene knockdown event according to the values mentioned in the paper, as follows: 0.04 for ”mitotic delay”, 0.092 for ”binuclear”, 0.11 for ”polylobed”, 0.03 for ”grape”, 0.0676 for ”large”, 0.06197 for ”dynamic” and 0.072 for ”apoptosis”. If the phenotypic event at the time point scored lower than the corresponding threshold, it was marked with ”-1”. After loading the data using PhenoTimer, I then filtered only for those events that were greater than -1 and obtained the phenotypic patterns presented in the *Results* section.

4.3.1.2 Networks

Two sources were used for network reconstruction in the cases where the network was not directly provided by PhenoTimer. First, the GO term enrichments and tree in Figure 4.9 were retrieved using BiNGO (Maere et al., 2005) within

Cytoscape (Shannon et al., 2003). Second, I used GeneMania (Montojo et al., 2010; Mostafavi et al., 2008) for the validation of links in the network of potentially synchronized genes. According to this database, 62.4% were found to have some biological motivation. The following distribution characterized the types of interactions extracted from the literature: co-expression 64.24%, physical interactions 14.68%, genetic interactions 11.16%, co-localization 5.46%, predicted 4.37%, shared protein domains 0.09%.

4.3.1.3 Evaluation of synchronized gene activities

The network of hypothesized synchronous genes was determined by direct comparison of the vectors containing the time course gene knockdown scores for the prevalent phenotypes. If the genes showed identical sequence of phenotypic events (including transitions at exactly the same time point), then they were marked as "synchronous" and linked in the network. The principle is further illustrated in Figure 4.8.

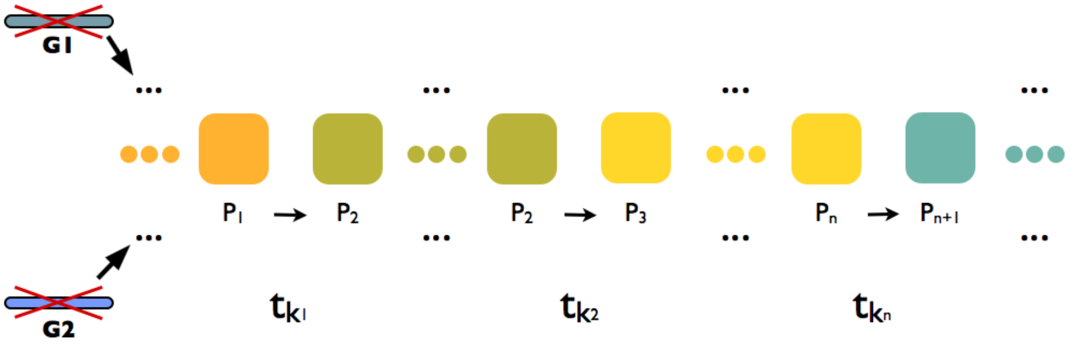


Figure 4.8: Two genes G1 and G2 are hypothesized to be "synchronous" if their knockdowns produce identical phenotypic succession events. The knockdown of genes G1 and G2 will each cause a series of phenotypic observations in the cell population. For them to be synchronous, all phenotypes P_1 ... P_{n+1} should occur in the same order and the transitions $P_1 \rightarrow P_2$ ... $P_n \rightarrow P_{n+1}$ should take place at the same time points t_{k_1} ... t_{k_n}. Different colors denote different phenotypes in the cell population.

4.3.1.4 Clustering of phenotypic profiles

To classify the genes in the network of hypothesized synchronous activities, I performed a PCA using the *princomp* method in R, after which I used k-means clustering (Pavlopoulos et al., 2011) to divide the first two principal components into groups. After several iterations with different values for k, I decided the best distinguishable groups were obtained for the value of $k = 4$.

4.3.2 Conservation of transcriptional events throughout the cell cycle

4.3.2.1 Data preparation

A series of data processing steps were necessary in the analysis of transcription conservation throughout the cell cycle. First, I compiled a collection of 600 human genes that have periodic peaks of transcription throughout the cell cycle from Cyclebase (Gauthier et al., 2010, 2008). These genes have been experimentally shown to have the highest expression at a fixed time point in the cell cycle. Second, I mapped the transcription peaks of these genes against the landscape of orthologs in other 51 species collected from BioMart (Kasprzyk, 2011). Next, I used iTOL (Letunic and Bork, 2007) to reorder these species according to the tree of life and attached a heat map of high transcription events to the obtained diagram. Cell cycle phases and phenotypes corresponding to the genes with transcription peaks at the same time point were afterwards mapped.

4.3.3 Transcriptional regulation linking cancer pathways

4.3.3.1 Data preparation

For every gene that exhibited periodic transcription peaks throughout the cell cycle according to (Gauthier et al., 2008), I searched for disease pathways where it was enriched using bioCompendium at http://biocompendium.org/. A threshold of 0.05 was employed for the p-values of the enrichments. The background considered was the whole human proteome. The list of enriched cancer pathways and the corresponding genes was then loaded into PhenoTimer.

	Quantile				
Drug treatment	**0%**	**25%**	**50%**	**75%**	**100%**
Control	6.376778	7.182875	7.746917	8.936444	12.368333
Saline	6.353444	7.358917	7.830667	9.148375	12.523333
Ethanol	6.487556	7.410542	8.135417	9.358958	12.667000
Nicotine	6.449667	7.326972	7.905417	9.147083	12.490333
Cocaine	6.286667	7.300000	7.972083	9.347958	12.610667
Heroin	6.392889	7.413083	8.165917	9.371813	12.531667
Morphine	6.504444	7.386278	8.114833	9.446313	12.745667
Methamphetamine	6.335889	7.466917	8.108417	9.417646	12.595000

Table 4.1: The calculated quantiles of the normalized and log2-transformed expression values, as recorded for every drug treatment.

4.3.4 Linking drug abuse phenotypes

4.3.4.1 Data preparation

The dataset of time course gene expression changes following induction of six different drugs (ethanol, nicotine, cocaine, heroin, morphine and methamphetamine), along with the controls, was extracted from NCBI's Gene Expression Omnibus database (Edgar et al., 2002) (GEO series accession number GSE15774).

To filter the drug effect measurements, quantiles have been calculated for the values of gene expression upon induction of the six drugs using R, as indicated in Table 4.1. The lower (25%) and upper (75%) quantile values have been used as thresholds in the visualization and analysis of drug effects. Thus, out of the 42 genes termed drug-responsive in the paper (Piechota et al., 2010), i.e. already identified as differentially expressed, I selected for further evaluation only those genes with expression values in the lower and upper quantile. I termed these genes ”relatively lowly” and ”relatively highly” expressed within the group, respectively.

4.4 Results

4.4.1 Discovering patterns in cell cycle regulation

As pointed out in previous chapters, cell cycle regulation has been intensely studied for its implications in organismal development and disease (Malumbres and Barbacid, 2009). However, the time dependence of this process still holds many unanswered questions. I used PhenoTimer to analyze different regulatory aspects and show how visualization can enhance the discovery of new informative patterns in the timewise modulation of events.

4.4.1.1 Progression dynamics of mitotic defects

As a first application, I illustrate the timewise tracing of morphological outcomes of cell division defects, as described by (Neumann et al., 2010). As mentioned in previous chapters, the knockdown of genes essential to the cell cycle resulted in cell division defective morphologies classified into seven main categories. It was observed that the cells don't adopt a single phenotype, but transition from one phenotype to the other before becoming arrested into a particular morphology or dying. Hence, there is a succession of phenotypes within the cell populations that can be traced through time. Using a combination of visualization and analysis strategies in PhenoTimer, I am able to show distinct patterns of morphological transitions as a result of disrupting the cell cycle and dig deeper into the potential causes and consequences.

I used PhenoTimer to visualize the transitions among phenotypes within the cell populations globally for all knockdown events in a time interval spanning approximately two cell cycles (see Figure 4.9). I only represented transitions to the most prominent phenotypes at every time point, i.e. maximally scored for the respective gene knockdown. Thus, the transitions visualized were the result of filtering according to thresholds as detailed in the 4.3.1.1 subsection. The three arc modes allowed for a global overview of knockdown effects, as well as a more detailed analysis of events at individual time points.

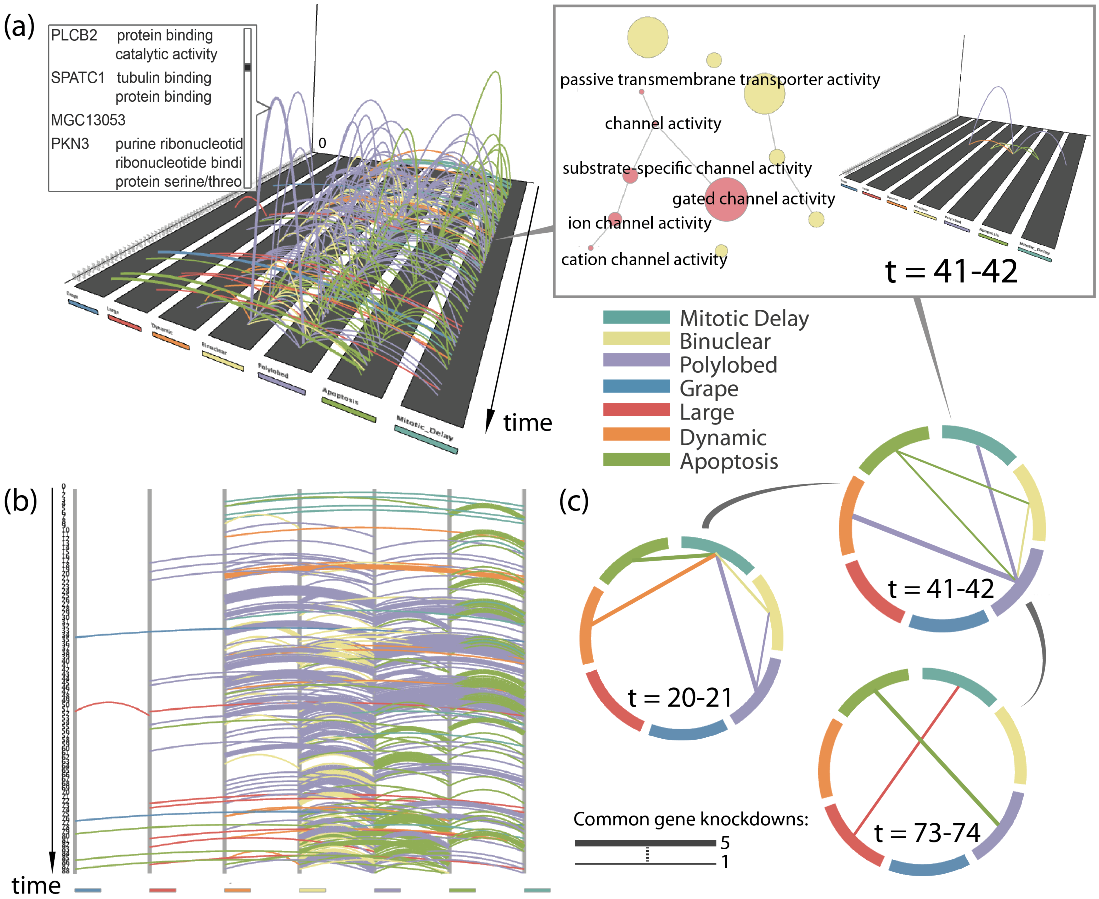


Figure 4.9: Phenotypic transition patterns observed in the imaged cell populations upon essential gene knockdown. (a) The 3D arc representation displays transitions at consecutive time points among the seven defined phenotypes, each labeled with a different color, as indicated in the legend. The color of the arc indicates the phenotype towards which the transition occurs. A number of 96 time points, equaling approximately two cell cycles, are traced. The height of the arcs is proportional to the number of gene knockdown events for which the transition occurs at the respective time point. Selecting an arc reveals a pop-up with gene information for the respective link. (b) The 2D arc view reproduces the previous plot in 2D. The width of the arcs in this case has the same meaning as their height for the previous plot. (c) Transitions at three time points are visualized using the circular view. The box in the upper right corner shows the transition events at a single time point (41-42) in 3D, along with the network of GO terms related to the dataset. The GO terms highlighted in light red correspond to the genes involved in the transitions at the given time point.

Global transition traces delineate prevalent and rare phenotypes

We can easily observe patterns: frequent transitions to ”apoptosis” (caused by severe cellular damage) and ”binuclear”-”polylobed” coupling (cytokinesis defects that are likely to succeed each other), as well as rarer transitions among ”mitotic delay”, ”grape” and ”large” phenotypes. The latter are likely slow and final transitions, as opposed to the former, which are more prevalent in the cell population. Thus, these global patterns serve well in distinguishing prominent (”binuclear”, ”polylobed”, ”apoptosis”) and rare (”grape”, ”large”) phenotypes. This is even more apparent when analyzing single phenotypic plots using the option provided in PhenoTimer (see Figure 4.10).

The timeline of gene functionality helps explain phenotypes

Coupling timing of gene events and their functional profiles, visualized dynamically in networks, reveals the genetic background that explains the phenotypes. Along with the morphological transitions throughout the cell cycle, I have visualized the GO network dynamically for each time point (shown in the box in Figure 4.9 with highlighted terms for time point 41-42). Following the changes in the network with time, one discovers a succession of molecular functions that can help explain the observed cellular transitions. While the dynamic network term highlighting cannot be shown here, Figure 10 in Appendix C summarizes the enriched biological processes through time. Many cell cycle related processes are enriched in the timeline, with periodic spikes for genes with roles in cell division, complex assembly and metabolic processes. The genes with relevance to spindle assembly peak early on, while the ones related to ubiquitination appear at later stages. The reconstruction of process activity reflects quite accurately the chronology of events one would expect to observe during the cell cycle. This suggests there is a link between the timing of the cell division event that is disrupted through gene knockdown and the onset of the resulting phenotype.

Potential new function for gene MGC12053

Using the integrated transition tracking strategy of PhenoTimer, one might infer new functions for unknown genes. Such an example is gene *MGC13053* (Figure 4.9a), which is involved in the same type of transition as *PLCB2*, *SPATC1* and

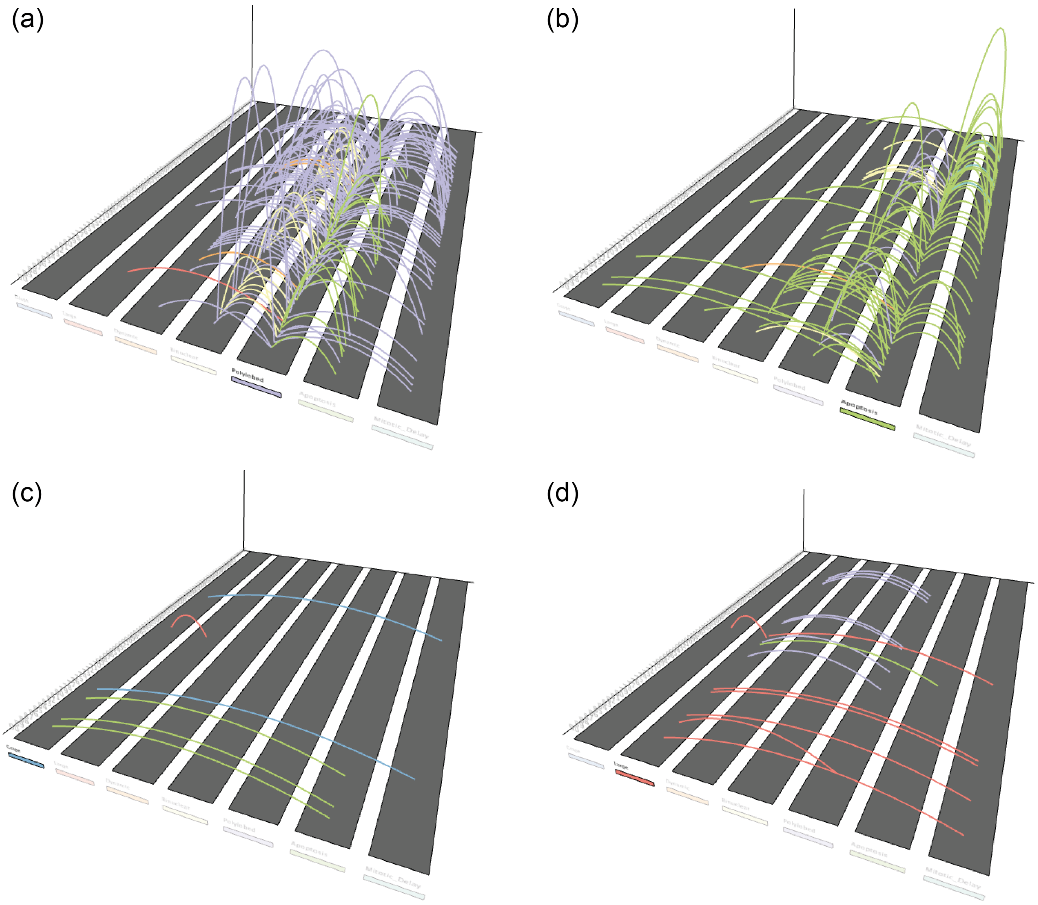


Figure 4.10: Single phenotype transition plots. The figure shows only those transitions that begin or end in a specific phenotype: (a) ”polylobed”, (b) ”apoptosis”, (c) ”grape”, and (d) ”large”. One can easily distinguish the more prevalent (a,b) and rarer phenotypes (c,d), even when considering only the transitions ending in the respective phenotype (depicted in purple for ”polylobed”, green for ”apoptosis”, blue for ”grape” and red for ”large”).

.

PKN3. These are genes associated to ribonucleotide binding processes, namely tubulin binding, centrosomal activity and phosphorylation events (according to GeneCards (Rebhan et al., 1997)). Thus, a valid hypothesis to test is whether *MGC13053* affects microtubule dynamics. While this would need experimental validation, the cytokinesis arrest phenotype exhibited (”polylobed”) suggests that it might be involved in defective spindle poll assembly or chromosome segregation.

Identical timing of phenotypic transitions suggests potentially synchronous gene activities

Analyzing the global overview of phenotypic transition profiles for all gene knockdowns led to the observation that some transitions are recapitulated for several knockdown events. This gave rise to the question of whether there are genes whose suppression causes similar phenotypic changes and whether this could be functionally motivated. Further investigation allowed me to certify that a substantial number of genes did indeed have not only similar, but identical phenotypic transition patterns. They were identified as described in subsection 4.3.1.3. Since their malfunction affected the cell populations in similar ways, I reasoned that the products of these genes should be involved in closely related activities, at least temporally if not also spatially. Thus, I could build a network of hypothesized synchronous gene processes, where connected genes have identical succession of phenotypic outcomes, as shown in Figure 4.11. These genes might be transcribed simultaneously or have products acting in the same pathways.

The network contains 482 genes that form clusters of interconnected nodes. Bigger clusters correspond to genes whose suppression causes ”binuclear” and ”polylobed” phenotypes, which are also the more prevalent morphologies overall. This suggests that many of these clusters could refer to proteins that participate in the same complex. Such large biological machineries would be ubiquitous in essential processes like the cell division and this could also explain why they act together. If they are part of the same complex, then its disruption through one knockdown or the other will affect the same cellular subprocesses. Nevertheless, this is not the only way in which identical phenotypes could arise: we cannot exclude co-expression, genetic interactions or pathway sharing. Therefore, these hypotheses needed further validation, as explained below. Interestingly, no synchronous events were observed for genes with first phenotype ”mitotic delay”, ”grape” or ”large”, suggesting that smaller pathways may be affected when these morphologies occur.

62.4% of the hypothesized gene ”interactions” were found to be linked biologically either through co-expression, physical interactions, co-localization or shared protein domains using GeneMania (Montojo et al., 2010; Mostafavi et al.,

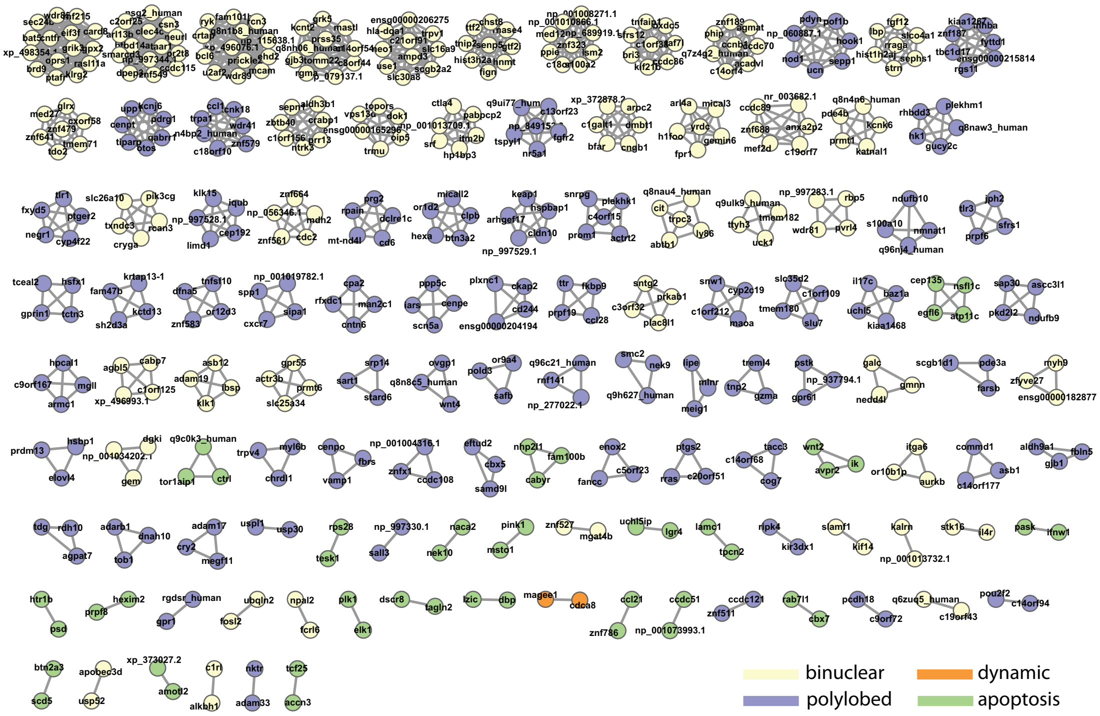


Figure 4.11: Networks of potentially synchronous gene activities. The nodes represent genes from the tested knockdown dataset and they are connected if their suppression produces identical sequences of phenotypic transition events. The colors of the nodes correspond to the first phenotype obtained in the cell populations upon knockdown of the respective gene. The networks were visualized using Cytoscape.

2008), as detailed in subsection 4.3.1.2. Among these, co-expression and physical interaction events were the most common and these are also the most likely ways to explain potential synchronous roles in the same processes. Functionality enrichment revealed frequent involvement of these genes in mRNA splicing events, with their protein products likely constituting parts of the spliceosome. A list of enriched molecular functions for the network of potentially synchronous genes is available in Table 4.2.

A closer look at four of the largest gene network modules, as depicted in Figure 4.12, shows that many of the hypothesized connections lack validation. However, the common effects upon knockdown, especially for well-defined phenotypes like ”binuclear”, imply the possibility of biologically-motivated links. Therefore, the novel interactions should be tested experimentally.

GO annotation	Q-value	Genes in networks	Genes in genome
catalytic step 2 spliceosome	$1.820605E-3$	12	80
spliceosomal complex	$4.184084E-3$	13	109
nuclear mRNA splicing, via spliceosome	$7.688298E-2$	15	196
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	$7.688298E-2$	15	196
RNA splicing, via transesterification reactions	$8.786896E-2$	15	202
mRNA processing	$8.967626E-2$	17	256
spliceosome assembly	$1.275452E-1$	6	32
nuclear body	$1.291282E-1$	11	123
ribonucleoprotein complex assembly	$1.491444E-1$	9	85
ribonucleoprotein complex subunit organization	$1.927164E-1$	9	89

Table 4.2: The molecular function enrichment of the genes that have hypothesized synchronous behavior, as obtained from GeneMania.

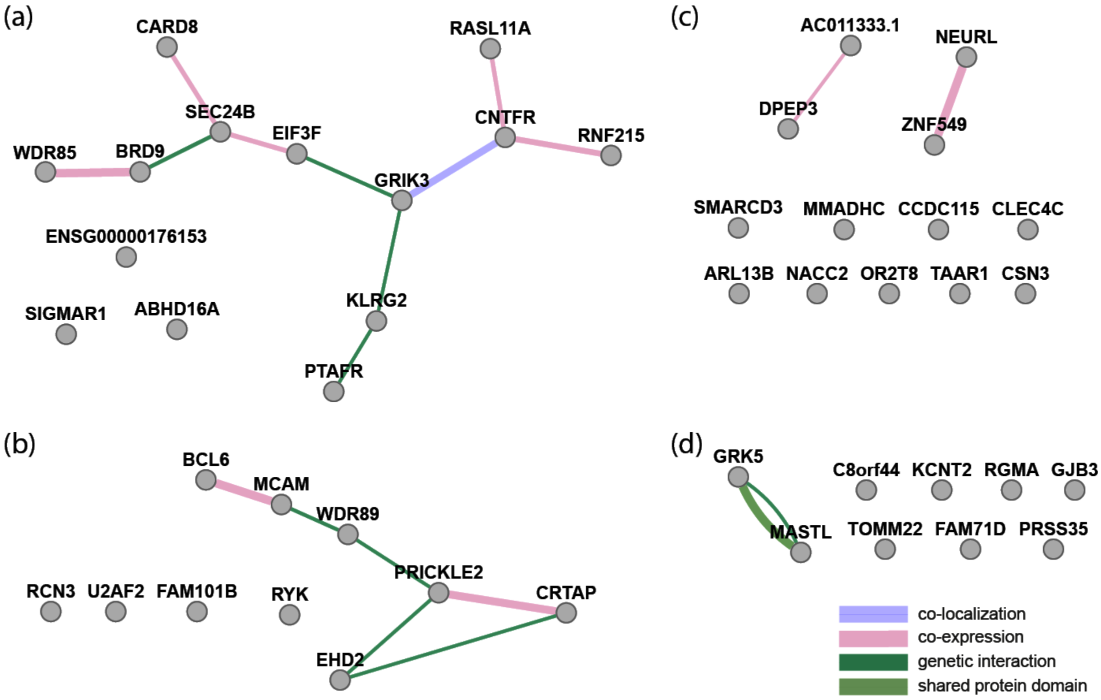


Figure 4.12: Four of the largest modules of potential synchronous gene activities. The connections between genes, retrieved from GeneMania, are colored differently to represent proof of biological relationships from the literature. The knockdown of these genes causes a shift to the ”binuclear” morphology in the cell population after: (a) 16.5 hours; (b) 15 hours; (c) 15.5 hours; (d) 26 hours.

Grouping of genes by time-course profiles recapitulates phenotypes

In order to discover whether this network of potential synchronous events could be further subdivided into classes, I used a clustering algorithm to group the genes according to their phenotypic succession profiles, as described in subsection 4.3.1.4. The resulting four groups are shown in Figure 4.13. This classification resembles the one according to the first phenotype in the population. The respective genes are probably involved in critical points of later stages of cell division, which require good coordination because of their complexity. Thus, the novel interactions identified may be of significant interest. Since no further substructure could be identified, they should be investigated individually.

4.4.1.2 Conservation of transcriptional events throughout the cell cycle

After having looked at consequences of perturbing the cell cycle, I asked to what extent the observed patterns might be explained by differences in conservation of regulatory events. To obtain an answer to this, I looked at how regulation of the cell cycle occurs across evolutionary scales and how these regulatory programs differ. For this purpose, I combined the transcription profiles of 600 essential genes that periodically fire throughout the progression of the cell cycle (Gauthier et al., 2010, 2008) with orthology information obtained from BioMart (Kasprzyk, 2011) to trace conservation of transcription in the cell cycle. Finally, I mapped phenotypic outcomes from suppressing these genes, as analyzed in the previous subsection, to the conserved expression peaks. This allowed me to assess the extent of phenotypic variability explained by time-resolved expression variation.

Figure 4.14 shows the conservation patterns of transcription peak events throughout the cell cycle. The evolutionary perspective and the mapping of cell cycle phases allow us to identify the most novel (less conserved) regulatory events. These are found in the G1 and S phases. In contrast to this, the end of G1 and beginning of S phase are remarkably conserved. Therefore, it appears that the G1 and S phases contain both the least and the most conserved periodic transcription events. This visualization approach enables mapping of transcription conservation throughout the cell cycle at time-point resolution. This fine-grained

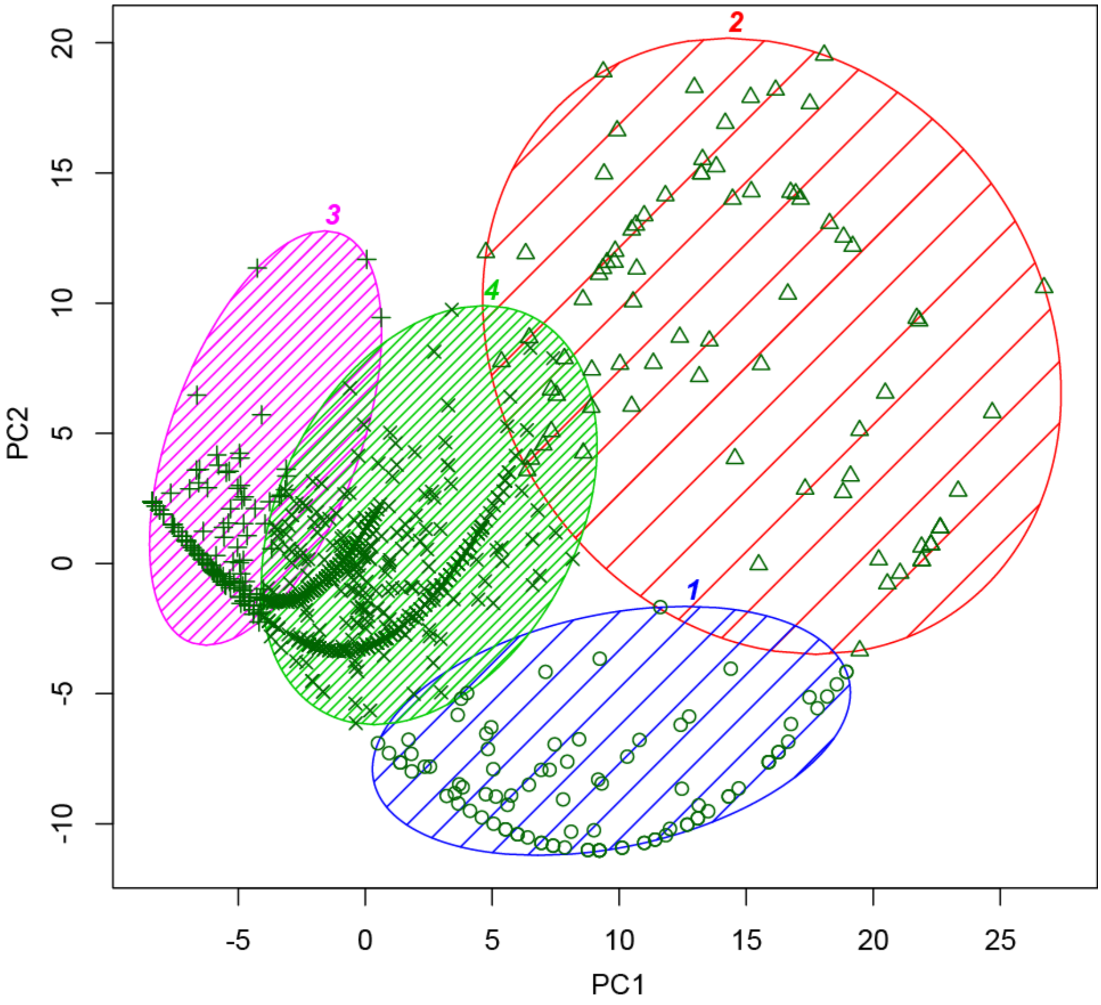


Figure 4.13: K-means clustering of synchronized genes according to their phenotypic succession profiles. The figure shows the grouping according to the first two principal components, which explain 71.27% of the variability. The four categories are distinguished by color and are numbered from 1 to 4.

view of cell cycle regulation offers the possibility to identify time points of interest for further investigation.

Combining this dataset with the one described in the previous subsection, I looked at how conservation of transcription may relate to effects of perturbing the respective genes (Figure 4.14b). The intersection of the two datasets was low: only 66 common genes out of 1067 in the first dataset and 600 in the second dataset. This shows how different techniques capture different aspects of a process and are best used complementarily. The discrepancies may also point out false

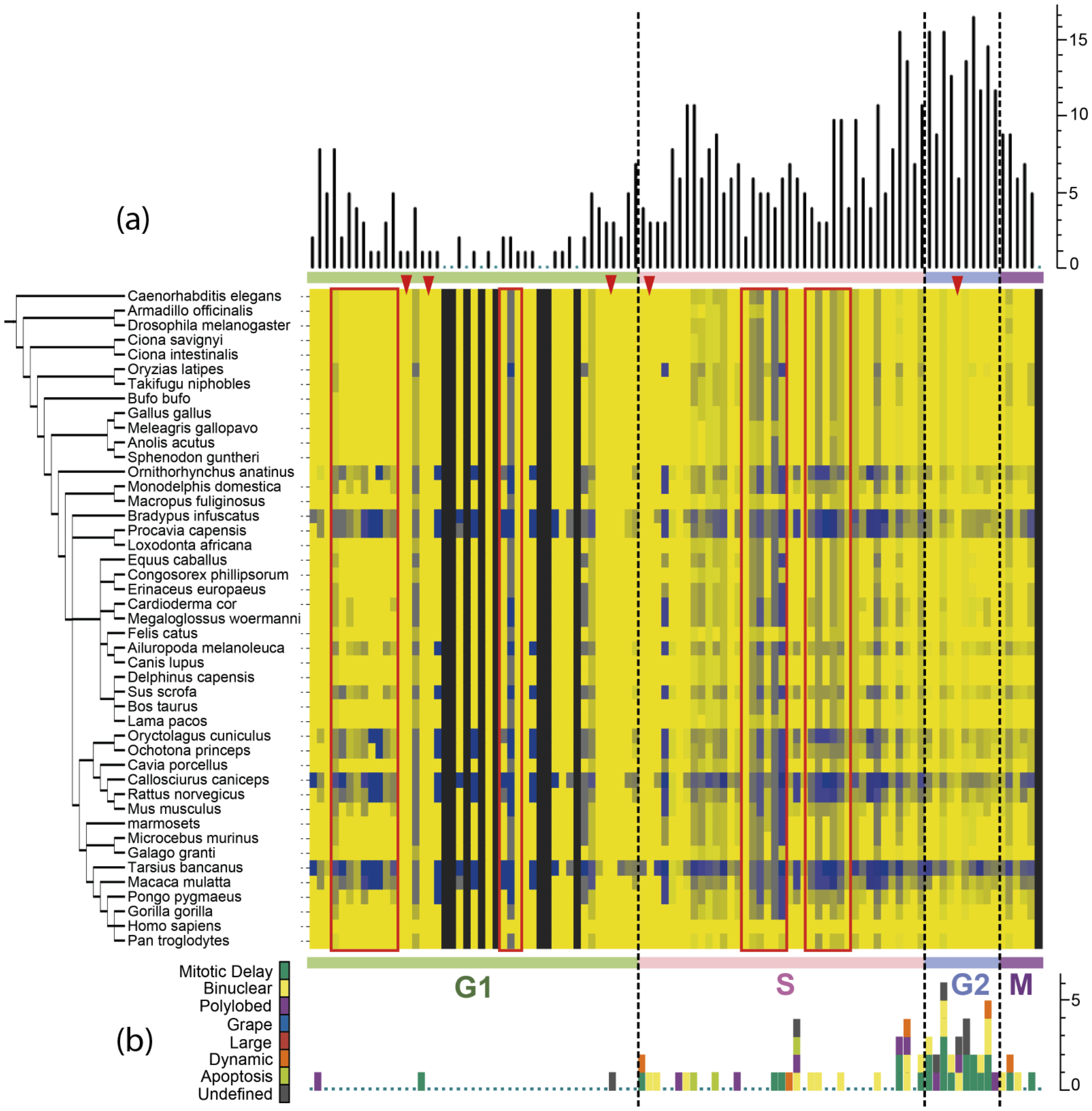


Figure 4.14: Mapping the human orthologs landscape to cell cycle events. The figure shows a heat map view of transcription peak points throughout the cell cycle in human and the degree of conservation of the corresponding genes in other species. A color gradient from blue to yellow is used to denote increase in percentage of conserved genes that have a transcription peak at a particular time point. The species are arranged according to the tree of life (the farthest species from human have been eliminated). The black vertical portions in the heat map correspond to time points where there were no peaks of transcription recorded for periodic genes. The cell cycle is timed from 0 (beginning of G1 phase) to 100% (end of M phase). The least conserved subphases are highlighted in red blocks and the best conserved ones are indicated by red triangles. Peak transcription count plots are displayed along the heat map: (a) the number of genes with a periodic peak of transcription at a certain time point; (b) phenotypic profiling after knockdown for genes that show a periodic peak of transcription throughout the cell cycle. In (b), the colored bars denote the phenotypes that result from silencing periodic genes. One bar unit corresponds to one gene. ”Apoptosis” appears only in the S phase for the available data and occurs for genes that are highly expressed in the more novel time points. The G2 and M phases are dominated by cytokinesis defects (”mitotic delay”, ”binuclear”, ”polylobed”).

positives and false negatives in the data, but this is beyond the scope of my analysis. I believe it is worth to further study the overlap between the two datasets, as it may facilitate the discovery of interesting regulatory aspects of a subset of cell cycle-essential genes.

The effects of perturbing genes that peak in the S-phase include some cytokinesis defects but also apoptosis. The latter occurs upon silencing of genes *KIFC2*, *USP1* and *CDCA5*, all essential for the cell cycle, and, notably, less conserved. Knockdowns of genes with transcription peak in the G2 phase seem to mostly cause cytokinesis defects (”mitotic delay”, ”binuclear”, ”polylobed”), indicating that the cell’s survival after division may be connected to the transcription firing rate of individual genes. Transcription events seem to be more conserved for the latter group of genes, which suggests the existence of some novel regulators of the human cell cycle with key roles in development.

4.4.1.3 Transcriptional regulation linking cancer pathways

Continuing the analysis of highly transcribed genes throughout the cell cycle, I investigated the impact of these differently conserved events on disease outcomes. From this dataset, I selected only those genes that were enriched in cancer pathways, as obtained from http://biocompendium.org/.

I used PhenoTimer to visualize high transcription events common to pathways affected in the following diseases: bladder, prostate, pancreatic, colorectal, small cell and non-small cell lung cancer and chronic myeloid leukemia. Figure 4.15 shows the transcription peaks of periodic genes through the phases of the cell cycle, along with the disease pathways where these genes are enriched. Two pathways are connected if at least one gene firing at the respective time point impacts both pathways.

From Figure 4.15b it is evident that high transcription activity of some genes, especially in the beginning of S phase and middle of G2 phase, is common to almost all types of cancers. In contrast, some other transcriptional events seem to be more specific for particular cancers. Such is the case of *VEGFC*, a gene highly expressed in M-phase and enriched in only two cancer types (bladder and pancreatic cancer). *VEGFC* [ENSG00000150630] is an important growth factor,

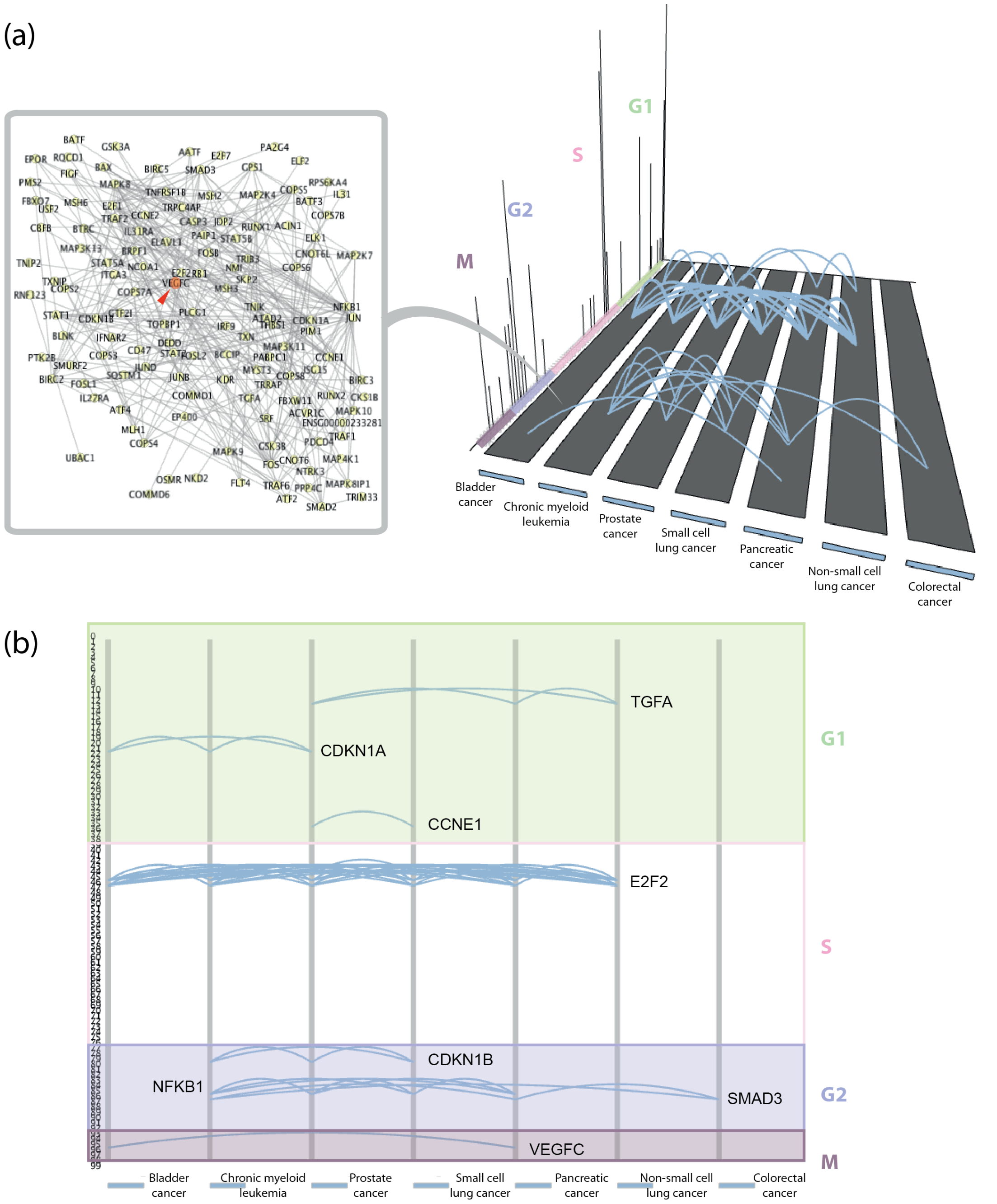


Figure 4.15: Common high transcription events in cancer pathways. (a) The 3D arc view in PhenoTimer has been used to visualize cancer pathways that involve genes with peaks in expression at periodic times throughout the cell cycle. The lanes correspond to different cancer types and the bar chart depicts the number of genes that have a peak in transcription at the respective time point in the cell cycle. The phases of the cell cycle are marked and time intervals colored correspondingly: G1, S, G2, M. Two cancer pathways are connected at a specific time point if at least one gene is commonly affected in these pathways and has the highest expression at that time point in the cell cycle. The connection between bladder and pancreatic cancer in the M phase reveals gene *VEGFG* to be involved in both pathways, as highlighted in the network retrieved from STRING. (b) The 2D arc view enables an easier inspection of all connections between cancer pathways, with the annotated genes for every connection.

with roles in angiogenesis and endothelial cell growth. Figure 4.15a also places this gene in the context of a PPI network derived from STRING, where it appears highly connected.

I further investigated how this gene relates to the other periodic genes enriched in at least one cancer pathway (not necessarily appearing in Figure 4.15). Figure 4.16 shows evidence from GeneMania that *VEGFC* interacts mainly through genetic interactions and co-expression with its network partners. Two of its interactors, *E2F2* and *NFKB1*, are also commonly enriched in several cancer pathways and they both have a peak of expression after the G1 phase. Hence, malfunctioning of either *E2F2* or *NFKB1* might play a role as tumor-triggering factor through the disruption of the interaction with *VEGFC*.

This analysis suggests that the regulation of different types of cancer seems to involve similar mechanisms for DNA replication, but cell division errors leading to disease may be cancer type-specific. Visualizing the modulation of transcription in different cancer pathways has thus enabled us to make some interesting, albeit naive (from lack of deeper investigation) preliminary observations. This manner of mapping disease links would, in a more complex context, enhance the interpretation of general and specific mechanisms of misregulation in related diseases.

4.4.2 Linking drug abuse phenotypes

Prolonged drug intake impacts the brain's reward system by generating addiction, and this has severe physical and social consequences. Many of the drugs of abuse have similar mechanisms of generating dependence, so studying the genes or pathways that are jointly altered by such drugs is an important step in elucidating their downstream effects. Furthermore, finding links between drug abuse outcomes might also suggest side effects of combining different drugs.

I looked at the impact of six addictive drugs on the mouse transcriptome, particularly at the gene expression changes in the striatum, as described in (Piechota et al., 2010). In this paper the authors measured alterations in transcription at intervals of 1, 2, 4 and 8 hours after acute administration of the following drugs: ethanol, nicotine, cocaine, heroin, morphine and methamphetamine. I used Phe-

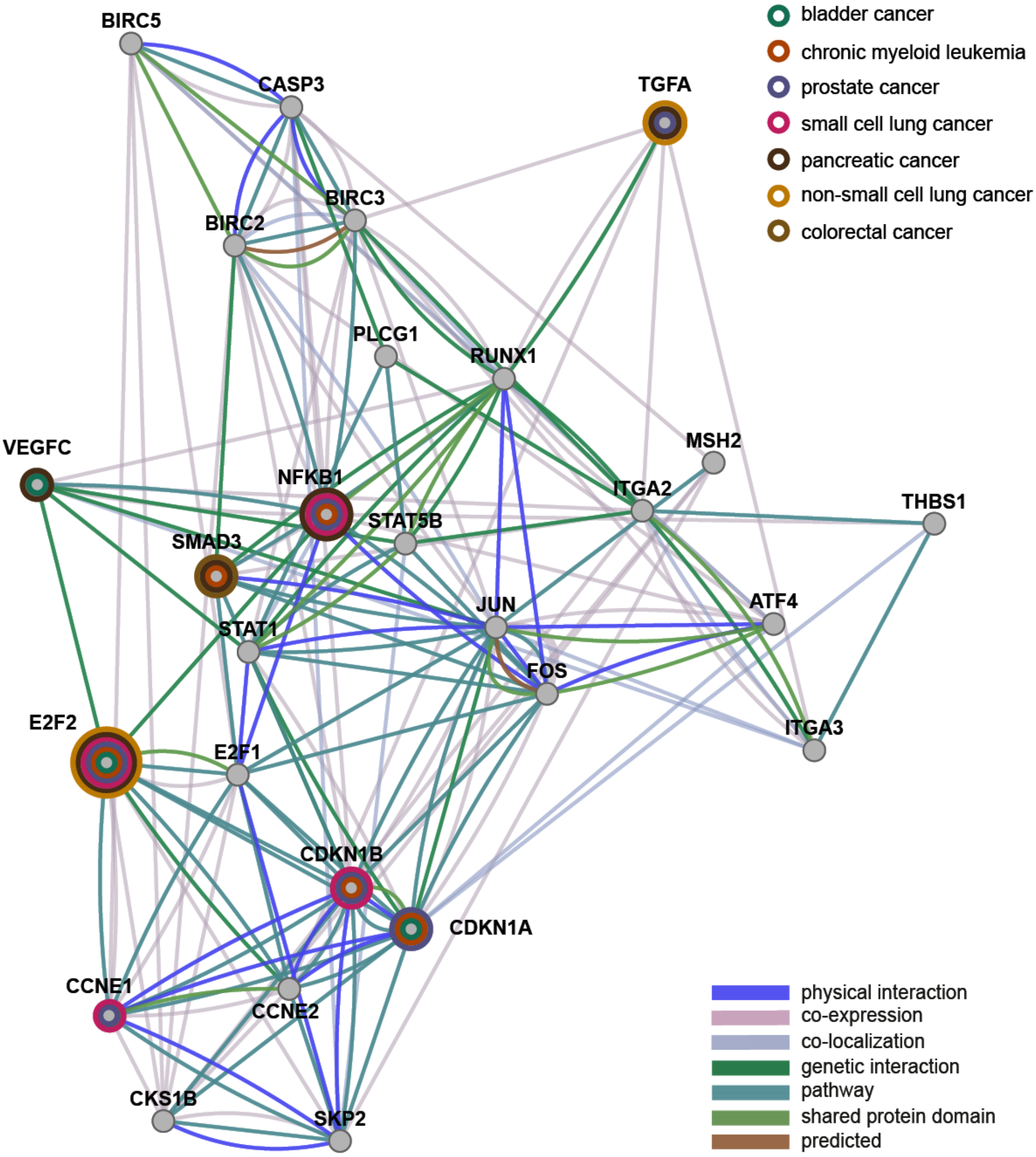


Figure 4.16: The network of periodically transcribed genes in the cell cycle that are also enriched in at least one cancer pathway. Nodes represent the genes and the links between them highlight evidence about their interactions, as obtained from GeneMania. If a gene is affected in more than one cancer pathway, the corresponding node is colored to indicate all the respective cancer types; otherwise it is depicted in grey. Most genes that are involved in the same types of cancer display genetic or physical interactions.

noTimer to visualize and analyze common outcomes of drug intake, as they result from the data. To do this, I selected drug-responsive genes that are relatively highly or lowly expressed (see subsection 4.3.4.1 for methods) and looked at their common modulation by pairs of drugs.

Drugs of abuse affect the transcription of many genes in a similar manner

Figure 4.17 shows connections between drugs that underlie a common genetic mechanism through time. If the expression of a particular gene is in the lower, respectively upper quantile upon administration of both drugs A and B at a particular time point, a link will be drawn to connect the two drugs. Networks of genes similarly modulated by the same drug(s) are shown in parallel to the evolutionary inspection of drug relationships. Thus, Figure 4.17 captures the evolution of connections (1) between drugs that impact the same genes, and (2) between genes that are affected by the same drugs. One can observe rather uniform overall drug impact on lowly and highly expressed genes in the dataset at all time points (i.e. very similar drug effects), but a more dynamic regulation landscape at the level of individual genes, as displayed by the changing networks.

At every time point, there is a core network of genes (yellow nodes) that stay constantly lowly or highly expressed throughout the treatment. Many of the lower expressed genes are related to proliferation, whereas the higher expressed genes modulate transcription and signalling processes through phosphorylation (see Tables 4 and 5 in Appendix C). In both cases there are also genes that have roles in stress response. These functions indicate that drug injection will trigger reduction in cell proliferation, enhancement of transcription regulation and will modify reactions in signalling pathways as a response to stress. These core genes are likely genes whose transcription changed quickly as a response to treatment, and thus more susceptible.

The variable genes (orange nodes - lower quantiles, green nodes - upper quantiles), on the other hand, were slower in responding to the drugs, either because their products are downstream factors of directly affected proteins or because they are more robust to perturbations in general. Some of them are major regulators of transcription, cell division or hypoxia response (see Tables 6 and 7 in

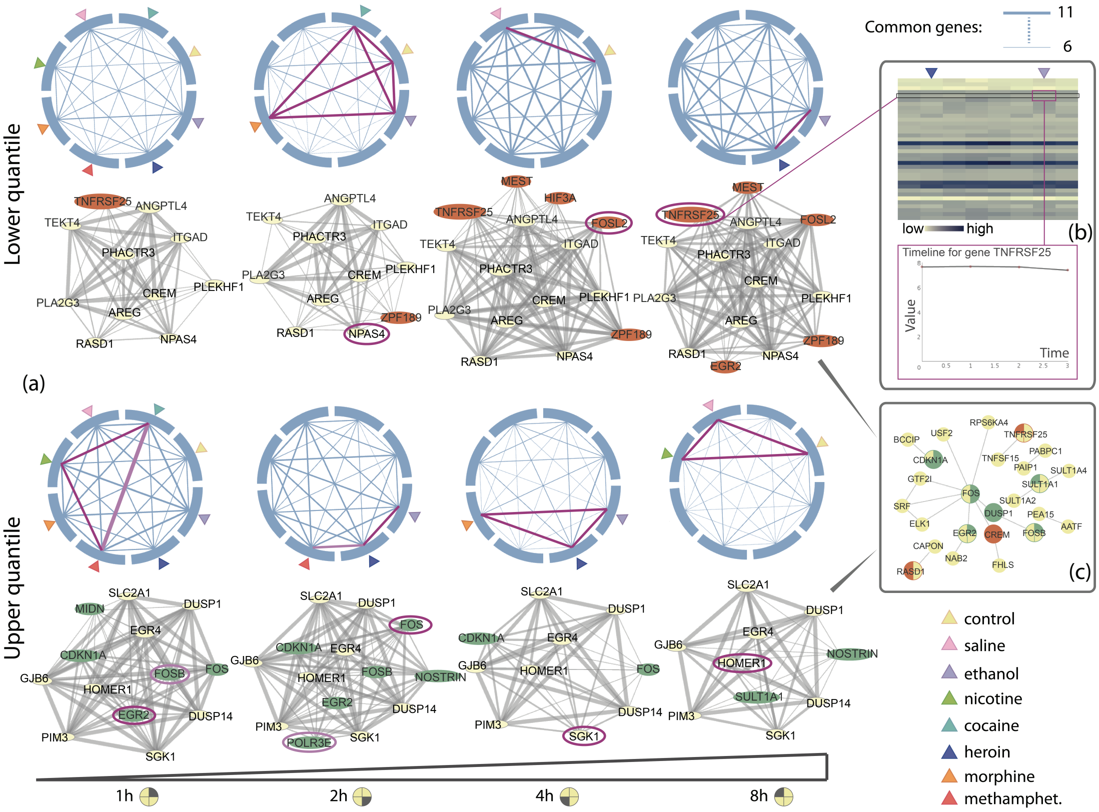


Figure 4.17: Similarities in drug effects on transcriptional regulation for up to 8 hours after injection. (a) The PhenoTimer circular view is used to depict drugs that regulate gene expression in similar ways. Arc segments correspond to drugs and colored triangle glyphs indicate the drug type. Controls are included. The order of the drugs in the circular view is the same for all time points. Two drugs are connected if at least one gene has expression in the lower quantile range (upper plots) - termed ”lowly expressed”, or upper quantile range (bottom plots) - termed ”highly expressed”, after treatment with either drug. The thickness of the links indicates the number of shared genes (maximum 11). Each circular plot depicts connections at a particular time point, from 1 to 8 hours. Below these plots, the networks of genes affected by the same drug(s) are shown, automatically generated from the data. The yellow nodes form the core network and correspond to genes whose expression stays constantly low/high throughout the entire time course. Genes whose transcription varies and thus appear and disappear from the network at various intervals are depicted in orange (lower quantile) and green (upper quantile). Magenta and pink links between drugs and correspondingly circled elements in the network refer to situations when the drug pair shares a maximum number of commonly regulated genes and at least one of these genes (indicated in the network) is unique for the connection (i.e. it is not involved in other drug pair links). (b) The heat map of transcription levels for all genes is shown after 8 hours of drug treatment for every drug. Rows are genes and columns are drugs, heroin and ethanol highlighted in blue and purple, respectively. The expression levels of gene *Tnfrsf25* are emphasized in a rectangle, with a zoomed-in line plot of the expression changes for ethanol treatment. (c) The network of human homologs for the drug-responsive genes has been retrieved from STRING. The yellow nodes are core elements from the other networks and the variable elements are highlighted in corresponding color according to their presence/absence in the network at different time points. All plots have been obtained using PhenoTimer and then combined and annotated to emphasize different aspects of the analysis.

Appendix C).

Core network genes are more conserved in human

The human homolog network for the full set of drug-responsive genes, retrieved from STRING, is displayed in Figure 4.17c. It contains many of the core genes and some of the variable highly expressed genes, but not many of the variable lowly expressed genes. The regulation of the latter might be mouse-specific. Thus, the core network is probably more consistently affected by drugs in both human and mouse, being more conserved.

Ethanol and heroin effects are more similar at later time points

The dynamics of the gene and drug connections underlie differences in mechanisms of action. Among the lowly expressed genes, *Tnfrsf25* is uniquely downregulated by heroin and ethanol after 8 hours of injection. It seems that this member of the tumor necrosis factor receptor superfamily [ENSG00000215788] is slightly more strongly affected by these two drugs. This subtle difference would barely be visible otherwise, as the heat map and line plot show in Figure 4.17b. In this case, the arc view helped capture this effect.

Heroin and ethanol have other similarities in modulating transcription, as shown in the case of genes *Fos* and *Sgk1*, both uniquely influenced by the two drugs after 2 and 4 hours, respectively. *Fos* is an important regulator of cell proliferation and differentiation and has been associated with apoptosis (Schwartz et al., 2000). *Sgk1*, a serine/threonine-protein kinase, is involved in DNA damage response (You et al., 2004). From these observations, we might infer an increased crosstalk between heroin- and ethanol-regulated pathways that relate to neuronal death and tumor development.

Stimulants and depressants share more regulatory effects between groups rather than within one category

When analyzing stimulants (cocaine, methamphetamine) and depressants (heroin, morphine) separately, one finds more synergies between drugs belonging to different classes rather than to the same class. The pairs of drugs cocaine-heroin and morphine-methamphetamine, respectively, share more common regulatory events, as seen in Figure 4.18. One would normally expect drugs from the same class to have more similar effects, but the results of the experiment seem to indicate the contrary. These are subtle differences that might not have a biological impact on the brain, but they should nevertheless be further investigated.

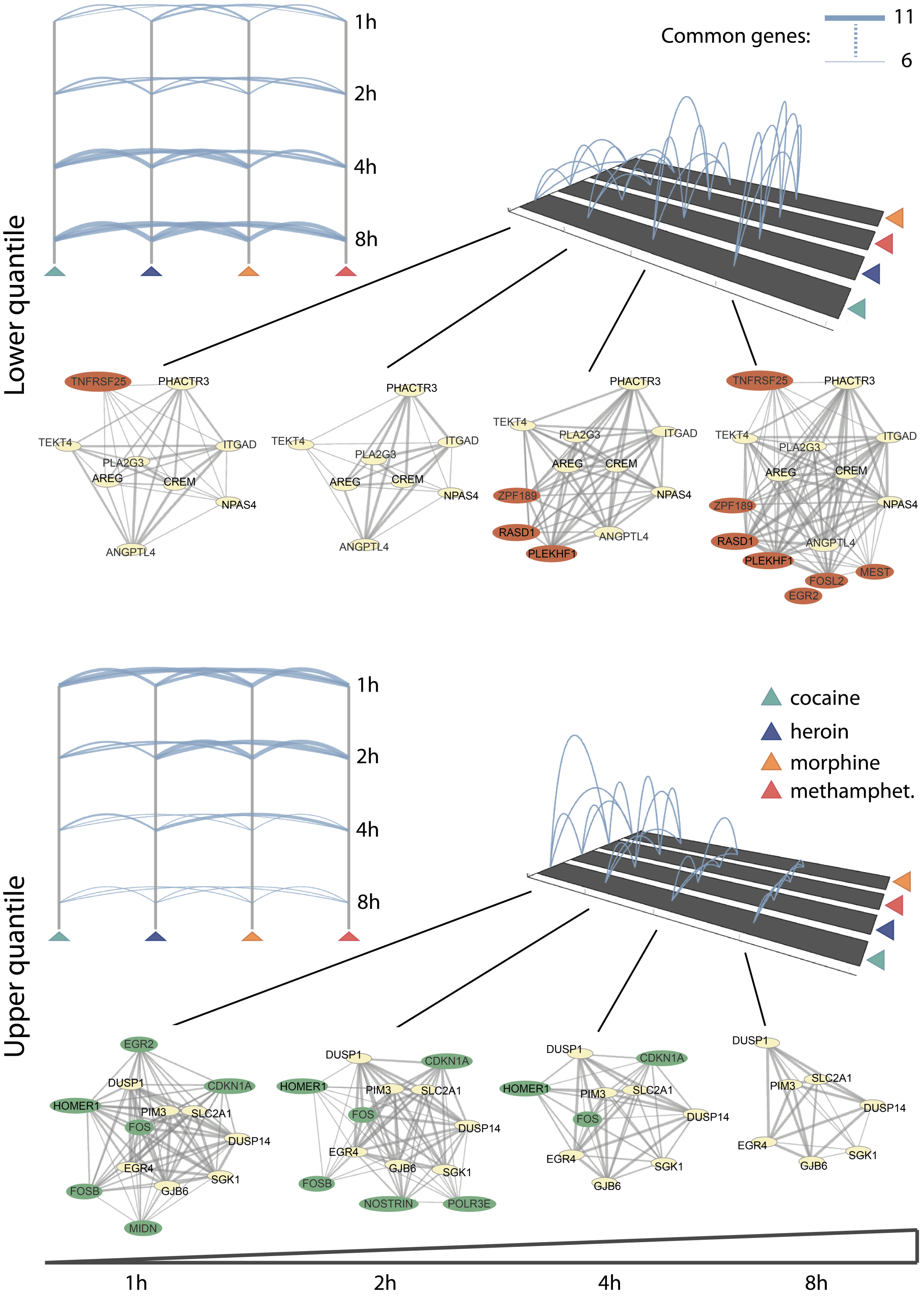
Gene subgroups reveal dynamic rewiring of drug links

The paper (Piechota et al., 2010) classifies the 42 genes responsive to drug treatment into four subclasses based on expression patterns. These categories underlie different biological processes as follows: A - behavioral sensitization and reward learning, B1 - reward learning and drug dependence, B2 - drug dependence, B3 - anti-neurotoxic response. Visualizing similarities in drug action mechanisms for these subclasses reveals a considerably more dynamic landscape, with group B1 showing the highest variation, as depicted in Figure 4.19. The underlying gene networks, extracted from GeneMania, range from mostly connected in A to almost no interactions in B3. The nodes highlighted in red correspond to the genes that are similarly downregulated by pairs of drugs. Their partners in the network are most often not affected in the same way, suggesting some compensatory mechanisms in the pathways. The details of the exact mechanisms are, nevertheless, too complex to be inferred from this visualization and should be further studied. Figure 4.19 proves that the visualization of a dataset is highly adaptable and can change depending on the way of organizing the data and on the question asked, to discover different patterns.

4.5 Discussion

4.5.1 Summary of results

In this chapter, I have presented a novel visualization approach to identify relationships between phenotypes depicting system perturbation effects or disease outcomes, at different time scales. I have introduced PhenoTimer as a tool that combines 2D and 3D representations of links between processes and integrates dynamic networks, in order to track regulatory events and infer new connections


Lower quantile
1h
2h
4h
8h
Common genes:
11
6
TNFRSF25
PHACTR3
TEKT4
ITGAD
PLA2G3
AREG
CREM
NPAS4
ANGPTL4
ZPF189
RASD1
PLEKHF1
FOSL2
MEST
EGR2
Upper quantile
cocaine
heroin
morphine
methamphet.
DUSP1
CDKN1A
HOMER1
PIM3
SLC2A1
FOS
DUSP14
EGR4
GJB6
SGK1
FOSB
MIDN
NOSTRIN
POLR3E
1h
2h
4h
8h

at molecular or organismal level. On the one hand, I have shown its effectiveness in exploring dynamic patterns in the regulation of the cell cycle and across evolutionary scales. On the other hand, I have applied its functionality to discover links between diseases or elucidate consequences of drug treatment.

First, I looked at the dynamics of cell populations when their cell cycle is disrupted by single gene knockdown. I was able to track the morphologies that these populations undertake and follow the transitions between different phenotypes. Since the experimental setup did not allow for perfect synchronization of the cells with respect to their cell cycles, I visualized patterns at the level of cell populations and not individual cells. These patterns are nevertheless relevant in the context of cell proliferation, tissue and organ development, especially since disruption of mitosis has severe implications in a wide range of diseases. Nonlethal cytokinesis defects (”binuclear”, ”polylobed”) were found to perpetuate in the population in a coupled fashion. I could distinguish prevalent and rare phenotypes, probably underlying differences in the centrality of pathway segments that were disrupted. A closer visual analysis suggested a new function for gene *MGC12053*, and simultaneous phenotypic successions were hypothesized to occur in a background of gene interactions. Thus, potentially novel synchronous genetic events (either by activity, physical interaction or participation in the same pathway) were proposed.

Further dissection of transcriptional events in different phases of the cell cycle

Figure 4.18 *(preceding page)*: Similarities in transcription regulation within two distinct drug categories: stimulants (cocaine, methamphethamine) and depressants (heroin, morphine). The upper plot shows connections between drugs that commonly regulate lowly expressed genes (with transcription levels in the lower quantile) and the lower plot displays connections for common highly expressed genes (with transcription levels in the upper quantile). The 2D (left) and 3D (right) arc views represent connections between drugs with similar mechanisms. Different lanes correspond to different drugs (indicated by the colored triangle glyphs). At every time point, the width (2D) or height (3D) of the links is proportional to the number of genes that two drugs modulate in a similar manner, maximum 11. The networks of genes that are down/upregulated by the same drug(s) were automatically generated and are displayed for every time point correspondingly. Yellow nodes form the core (gene expression stays in the same range for all time points) and orange (lower quantile) or green (upper quantile) nodes represent genes with variable expression. The thickness of the edges is proportional to the number of drugs that similarly regulate the two genes. The plots have been obtained using PhenoTimer and then merged and annotated to emphasize different aspects.

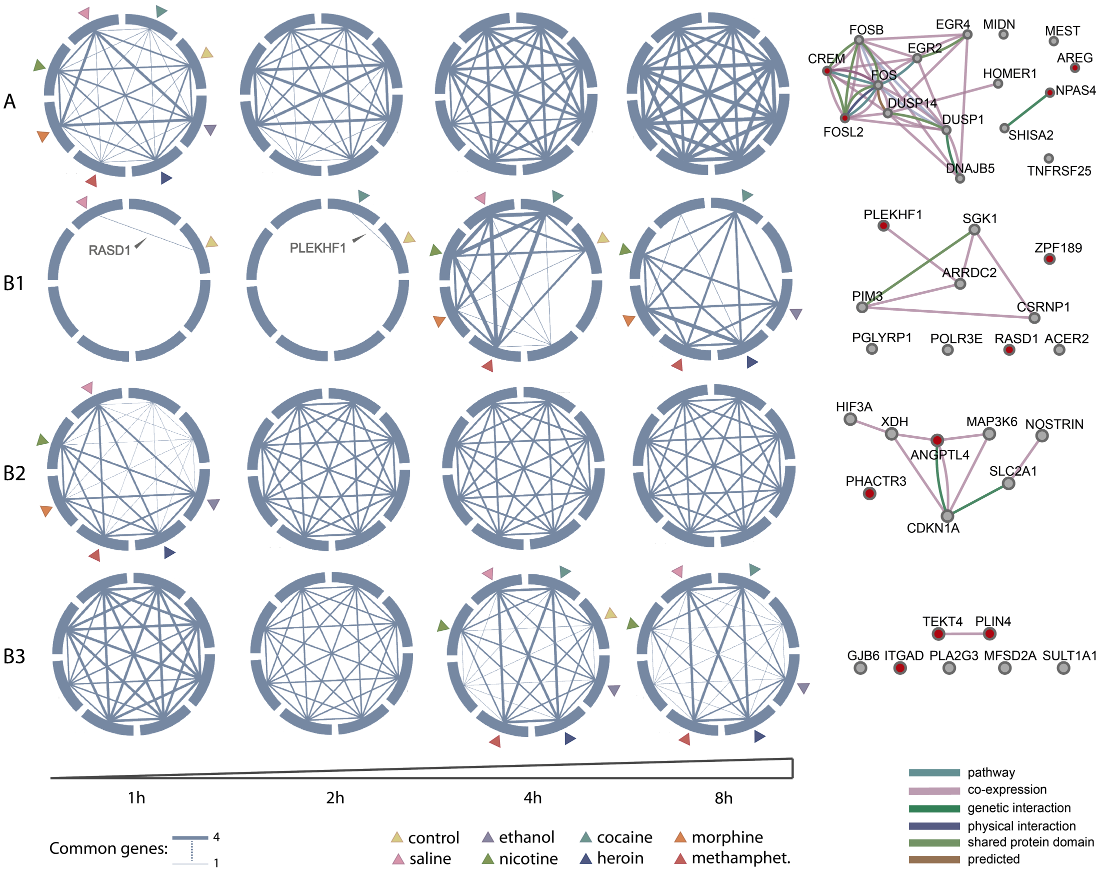


Figure 4.19: Drug mechanism similarities for different gene subclasses. At every time point from 1 until 8 hours after injection two drugs are connected if after treatment the transcription of the same gene dropped to levels in the lower quantile, as calculated in the subsection 4.3.4.1. Only effects on lowly expressed genes are shown. The more genes commonly regulated by two drugs, the thicker the connection between them is. The connections are plotted separately for genes belonging to different classes: A (behavioral sensitization, reward learning), B1 (reward learning, drug dependence), B2 (drug dependence) and B3 (anti-neurotoxic response). Each arc segment represents a drug, triangle glyphs indicate the drug type; the same order is kept in all plots. On the right hand side, networks depict the interactions between genes in each category, as obtained from GeneMania. Interactions are colored differently depending on the type, according to the legend. Genes highlighted in red are involved in drug connections for the respective class. The circular plots have been obtained using PhenoTimer and then combined and annotated to emphasize different aspects of the analysis.

and their conservation allowed the identification of the most novel subphases from an evolutionary perspective. They occurred in the beginning and middle of the G1 phase and middle of the S phase. Most cytokinesis phenotypes appeared for genes with transcription peaks in the G2 phase. This evolutionary perspective of transcription throughout the cell cycle was followed up by the discovery of such events that are common to different cancer pathways. A potential link between pancreatic and bladder cancer was proposed in the mechanism of disease onset: the disruption of genetic interactions between *VEGFC*, *E2F2* and *NFKB1*. While this analysis simplified many of the biological aspects, it did serve in showing how PhenoTimer can be used in a larger setting for the discovery of new connections between pathways, processes or diseases.

Finally, visualizing drug addiction effects at a molecular level using PhenoTimer helped identify similarities in the drugs' mechanisms of action that are often hard to dissect. Explorative and interactive visualization comes to use in this case to obtain further details about the genes involved. Unexpectedly, drugs within the same class exhibited less similarity in transcriptional modulation compared to others in different classes, at least at the level of relatively highly and lowly expressed genes. While more thorough investigation is needed to allow for causal inferences, this approach has helped to emphasize further details in the complexity of addiction mechanisms. Most importantly, phenotypic links between two drugs may suggest synergistic effects or interference between them. While the uniformity of transcriptional changes for different drugs in this dataset did not allow for such discoveries, this strategy could prove much more effective for other datasets. Linking drug effects would provide hypotheses of combinatorial drug effects for experimental testing. Of course, this has potential relevance in drug trials and therapy design.

4.5.2 Combining 3D and 2D visualization

PhenoTimer combines five visualization modes into an integrated platform. The reason for this design choice stems from the desire to extend the limits of visual exploration in 2D with additional 3D features, but without deviating from the rules of proper information visualization. After several stages of design, it was decided that an approach that combines 2D and 3D visualization is optimal as a general framework for analyzing heterogeneous data that aggregates measurements for multiple time points and phenotypes. Depending on the size and content of the dataset, as well as on the biological question, the user may choose to use one visual depiction or another, or combine several of them to get the most out of the available data.

While the heat map and line plots constitute complementary features that offer extra detail, the arc modes are more useful in detecting patterns and identifying interesting connections. In this respect, some of the arc views work better than others in depicting such patterns depending on the data. When many links exist between phenotypes or other variables, the 3D arc view might alleviate visibility problems caused by extensive overlap in the 2D view. Also, the height of the arcs in 3D is visually easier to compare than their width in 2D when differences are not so large, such that the underlying gene numbers are indicated more effectively. Nevertheless, 3D representations are not without weaknesses: occlusion and perspective distortion can misguide interpretation (Tory et al., 2006). Interactive features like zooming, panning or rotating partially overcome this, but more often the linear 2D and circular layouts would prove more effective. The 2D arcs provide a better overview of time traces, while the circular representation is more advantageous for single time point analysis.

4.5.3 Comparison to similar visualization tools

PhenoTimer uses a novel display of links between phenotypes through time in the form of arcs, an idea that, to our knowledge, is not present in this form in any of the software available for biological data analysis so far. A similar feature is described by (Tominski and Schumann, 2008), but they use links only in 3D and solely for connecting genes. PhenoTimer offers more flexibility with the additional 2D views and, more importantly, looks at connections between phenotypic outcomes rather than at the genetic background. Classical methods like clustering or simple heat map plotting would in most cases still work better for connecting genes, but, when it comes to the observed phenotypes, this way of visualizing relationships allows for an extended exploration and opens a larger repertoire of questions.

The idea of visualizing dynamics in networks in parallel with the other plots is similar to strategies employed by tools like GATE (MacArthur et al., 2010), VistaClara (Kincaid et al., 2008) or SpotXplore (Westenberg et al., 2010), which use color gradients to map expression changes within the network. It also resembles some aspects of the visualization in TVNViewer (Curtis et al., 2011), where links between nodes are rewired at every time point (this is also valid for the circular arc view, which draws both from Circos (Krzywinski et al., 2009) and TVNViewer (Curtis et al., 2011)). However, the network visualization employed in PhenoTimer is fundamentally different in several aspects: (1) it is dynamically linked to the arc plots, heat maps and line plots and updated at every time point; (2) it highlights genes and functions involved in certain processes at the respective time point; (3) there are several ways to generate a network, and different types of networks can be queried; (4) it links to databases for more information; and, importantly, (5) the network can be generated automatically by the application throughout the time course to reveal genes that determine similar phenotypes. PhenoTimer is, however, considerably more limited in the size of the networks it can visualize, as well as the size of the initial *gene* x *phenotype* matrix loaded.

Like some of the tools, e.g. STEM (Ernst and Bar-Joseph, 2006), PhenoTimer also uses linear plots and clustering to visualize time course profiles. Adding to this there is the heat map view, with rows and columns clustered dynamically at every time point. This dynamic clustering, along with visualization of different plots in a global as well as a single time point view, is an advantage of PhenoTimer.

Ultimately, the combination of different arc visualization modes in 2D and 3D, networks and other graphics, along with data integration methods grants extended versatility in data inspection. All this together constitutes an innovative way of visualizing phenotypic relationships in a temporal context.

4.5.4 Future development and directions

In terms of the software, future plans include porting the application development to the newest release (2.0) of the Processing platform. This new version includes some bug fixes that would solve inconsistencies in resizing windows and provide a smoother control and better cross-platform compatibility.

To enhance data analysis within PhenoTimer, we plan to implement a series of graphical and statistical methods that should ease the interpretation of large datasets. More layout algorithms for networks would enable the users to discover hidden structures or groupings in the data. This could be combined with clustering (K-means, hierarchical etc.) and PCA to extract similarities within gene expression time series. Retrieving information from other databases like GO (Ashburner, 2000), KEGG (Ogata et al., 1999), Reactome (D'Eustachio, 2011) etc., and linking out to more resources would add to the data integration aspect of the tool. These enhancements should expand the potential for analysis after the visualization step.

The innovative combination of features comprised in PhenoTimer, along with the generalized framework, makes it suitable to a whole range of biological applications. Beyond the already shown examples, PhenoTimer could also be used to visualize spatiotemporal regulatory programs encoded within chromosomes or to map correspondences between different chromosomes (including crossover or other rearrangements). Dynamics of populations could be studied by visualizing variations within a population along a time frame (for instance, bacterial species variation in the human gut and how it changes with diet). Moreover, timing common steps in disease progression could suggest new strategies for drug repurposing and optimal start of treatment. We thus envision a good potential for this software to help with visual patterning of phenotypic processes at different temporal and systemic scales.

Chapter 5

Conclusions

One of the ongoing challenges in biology is understanding how phenotypes emerge from genetic and environmental factors (Bochner, 2003). The aim is twofold: (1) to understand how changes in the genetic structure or environmental components can affect phenotypes, so that we can predict diseases, and (2) given a particular disease phenotype, to be able to reconstruct the regulatory mechanisms that failed and thus come up with ways to repair or attenuate the damage. My thesis has focused on improving the understanding of such aspects in the context of cell cycle regulation and of dynamical systems and has employed a series of analysis and visualization methods for elucidating gene-phenotype relationships.

I have investigated the dynamics of motif-mediated interactions within the cell cycle and their consequences in disease. Biological systems are tuned to a fine degree of detail; therefore, it is not surprising that short peptides can have such a big influence. The enrichment analysis indeed suggests that they are relevant not only in the regulation context, but also for phenotypic outcomes in human. Knowledge about disease-associated mutations of these elements allows us to make the connection from the relatively abstract phenotypes defined by (Neumann et al., 2010) to the concrete outcomes of various disorders. All this is achieved by dissecting the network of interactions underlying cell cycle regulation.

Understanding the temporal determinants of phenotypic instantiation and succession at a basic biological level required the development of new tools for dynamically linking such factors. From a global perspective, I wanted to be able to understand how phenotypes differed and how they would succeed each other in the cell populations. Visualization is a great aid in discovering global patterns and I developed tools to address these questions. Arena3D and PhenoTimer approach the challenge of linking phenotypes from different perspectives. While Arena3D focuses on the connections within the regulatory network underlying different phenotypes, PhenoTimer identifies connections between these phenotypes through time based on genetic determinants. In Arena3D, phenotypes are compared as separate biological outcomes and their evolution is followed through time. In PhenoTimer, a higher level of data integration allows simultaneous inspection of how the relationships between phenotypes (either transitions or common factors) develop with time and how gene perturbations result in similar outcomes. Global patterns, as well as events at single time points can be easily detailed. Ultimately, the transitions observed between these phenotypes are not very different from the stages of disease progression. Hence, the visualization tools developed would be just as well suited for biological questions with application to disease.

Overall, this study extends the knowledge of genotype-phenotype relationships in the context of the cell cycle. On the one hand, the approaches used have enabled a more detailed understanding of morphological defects that appear in cases of cell division failure. Beyond tracking phenotypic transitions in the cell populations, I was also able to look at the conservation of these events and at potential interactions that might underline the key points of cell cycle robustness. On the other hand, I could explore different visual strategies to identify optimal representations of connective patterns with biological relevance in large-scale datasets.

In the future, we envision that placing the observations of linear motif and post-translational synergies in the context of spatial and temporal dynamics will lead to a better understanding of the regulatory circuits behind essential biological processes. Since it is becoming increasingly clear that the modular proteome organization shapes complex phenotypic patterns (Gstaiger and Aebersold, 2013), such analysis will become crucial in linking genes and phenotypes at a more detailed level.

On the visualization side, there is no doubt that strategies for biological data representation will necessitate continuous improvements to adapt to the new types of analysis coming especially from genomics and personalized medicine approaches (Holmes et al., 2009). Visualization remains important for biological data inspection, even more so in the future. The visualization tools developed within this project enable exploratory data analysis for a wide variety of datasets, from small to large scale, and should be of interest to researchers working in many biological and bioinformatics fields. They provide systematic frameworks to study dynamic phenotypes coming from medium to high-throughput experiments. It seems that there is a considerable interest of the scientific community in such visualization strategies, as shown by the access statistics of articles published about these approaches in the Publications section. This indicates their potential to be useful in several research contexts.

With the help of tools like the ones described, gene-phenotype relationships can be investigated in further detail to understand the mechanism of diseases and the reasons behind the various responses to drugs in human and in other organisms. Ultimately, combining such analysis and visualization strategies will be of the essence to bridge the genotype-phenotype gap.

Appendix A

This appendix contains the supplementary information for Chapter 2.

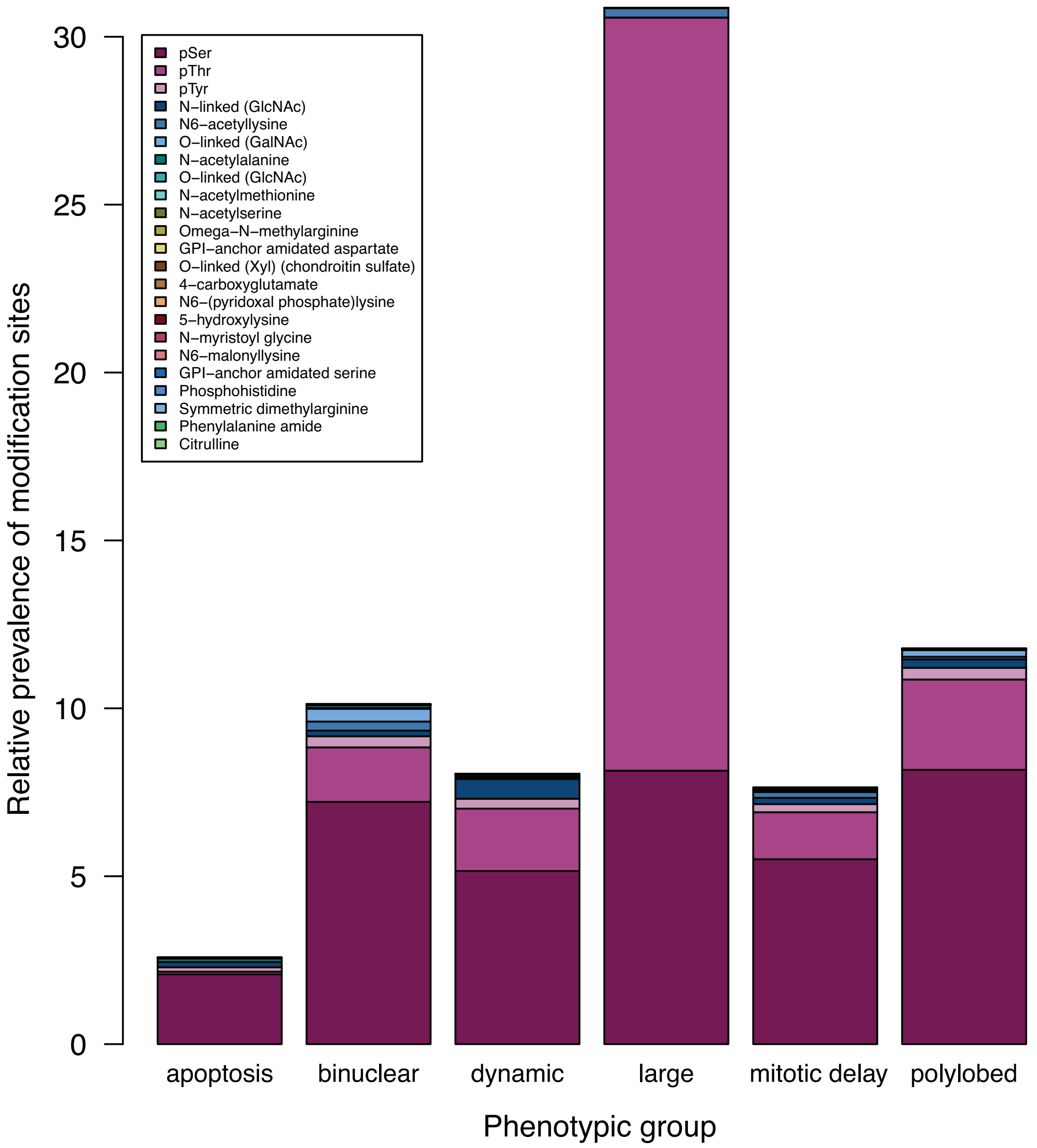


Figure 1: Relative PTM site count per phenotypic group. Different modification sites are denoted by different colors. Abbreviations: pSer - phosphoserine; pThr - phosphothreonine; pTyr - phosphotyrosine.

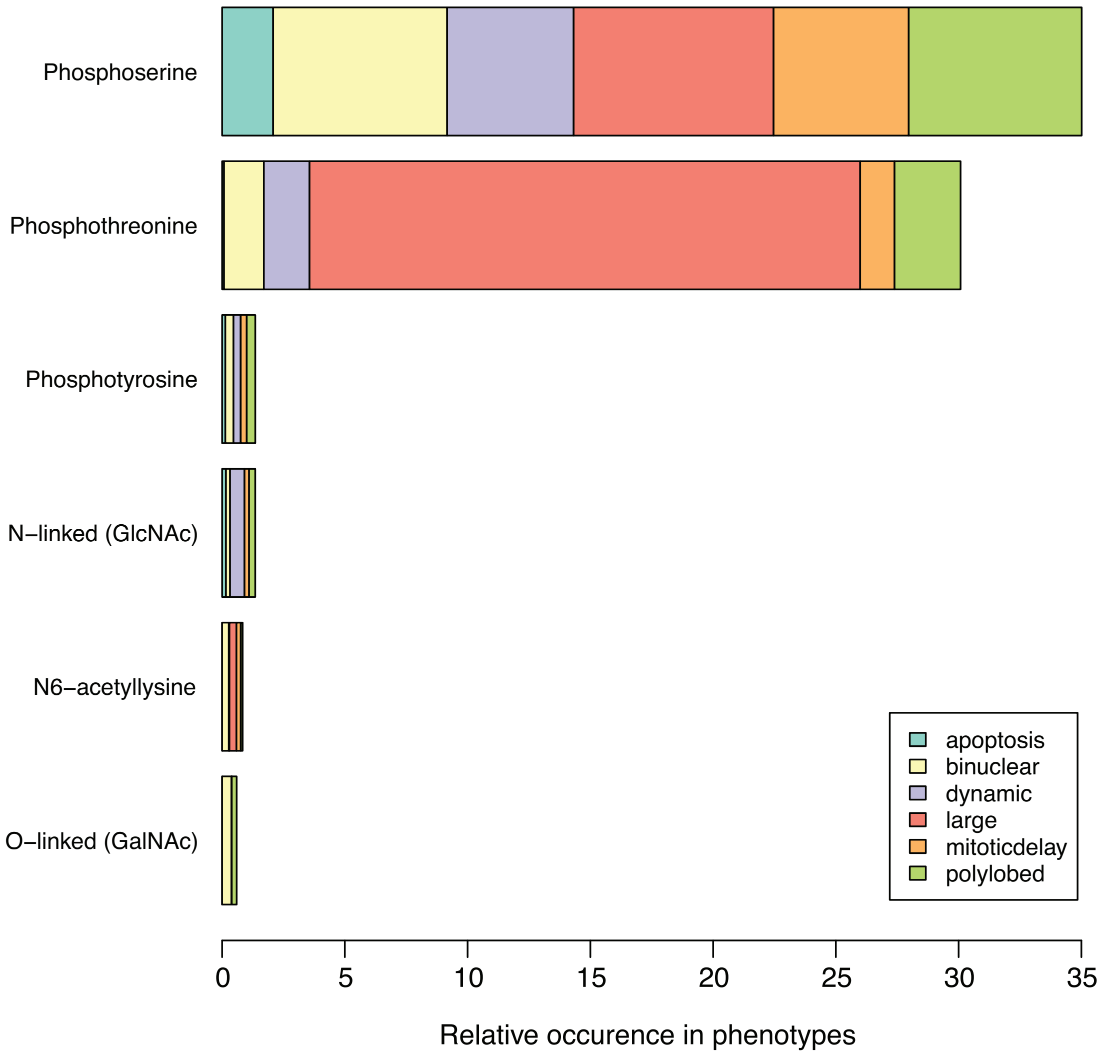


Figure 2: Distribution of top PTM classes, by phenotype.

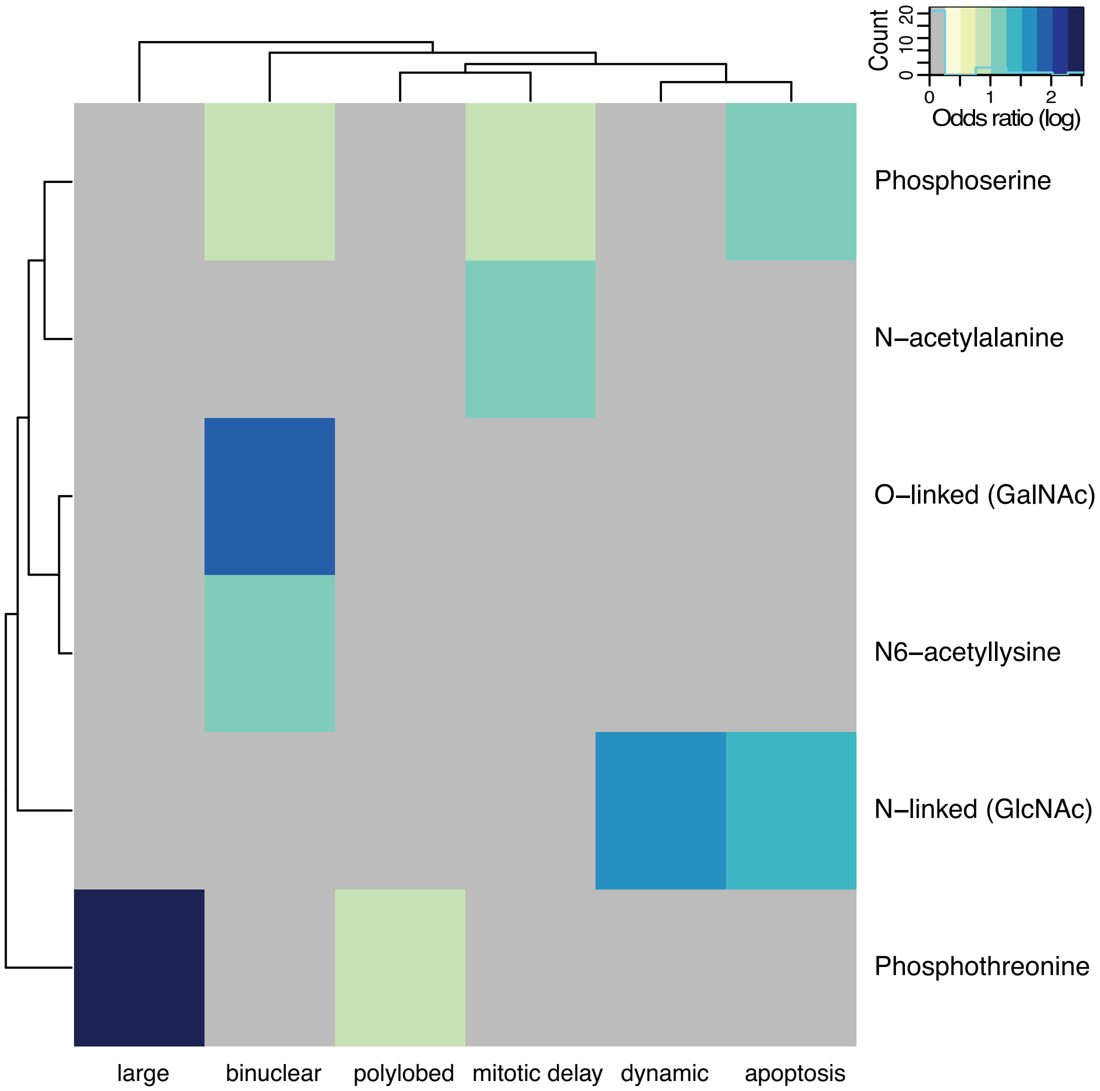


Figure 3: Major classes of PTMs enriched around SLiMs in different phenotypic groups. The heat map shows the natural logarithm of the odds ratio, mapped on a color gradient from yellow to blue. Grey indicates no enrichment.

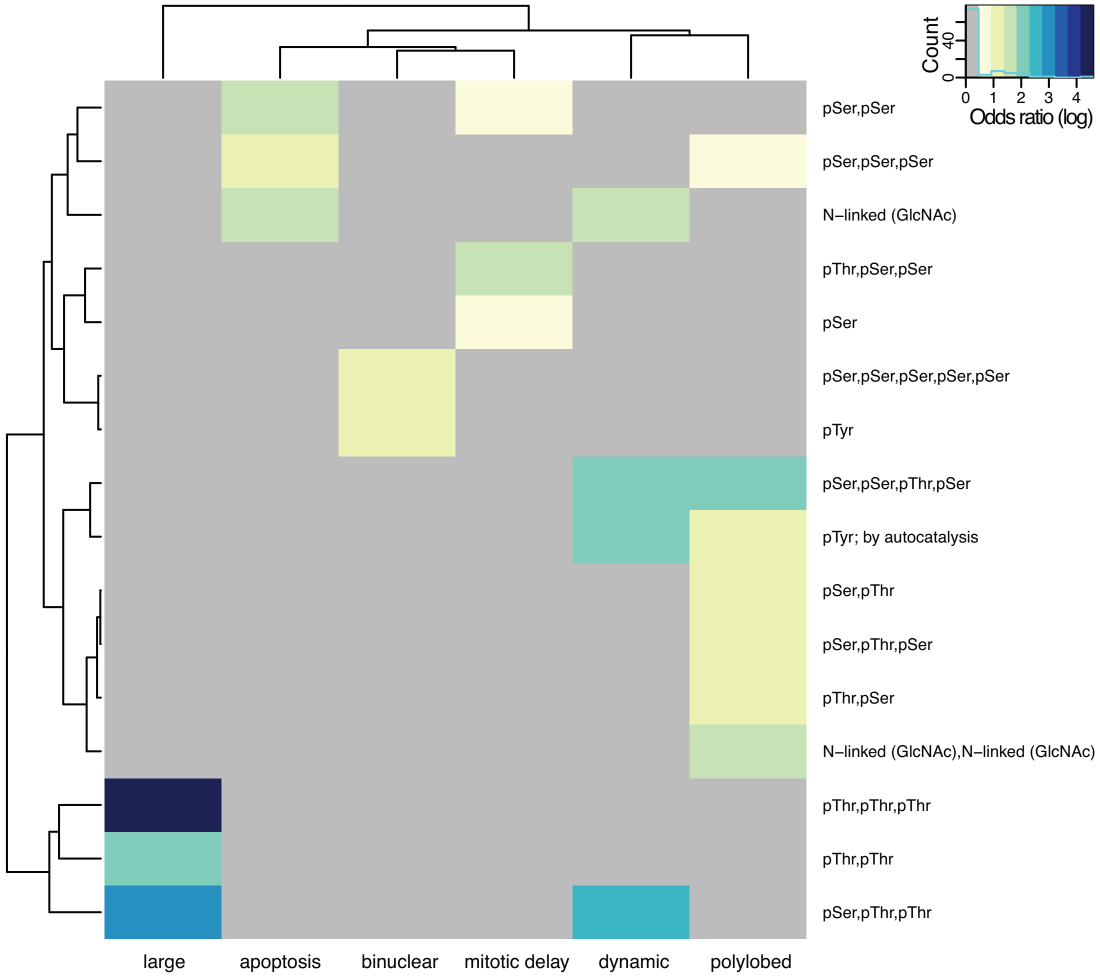


Figure 4: PTM patterns enriched around SLiMs in different phenotypic groups. The heat map shows the natural logarithm of the odds ratio, mapped on a color gradient from yellow to blue. Grey indicates no enrichment. Abbreviations: pSer - phosphoserine; pThr - phosphothreonine; pTyr - phosphotyrosine.

Appendix A

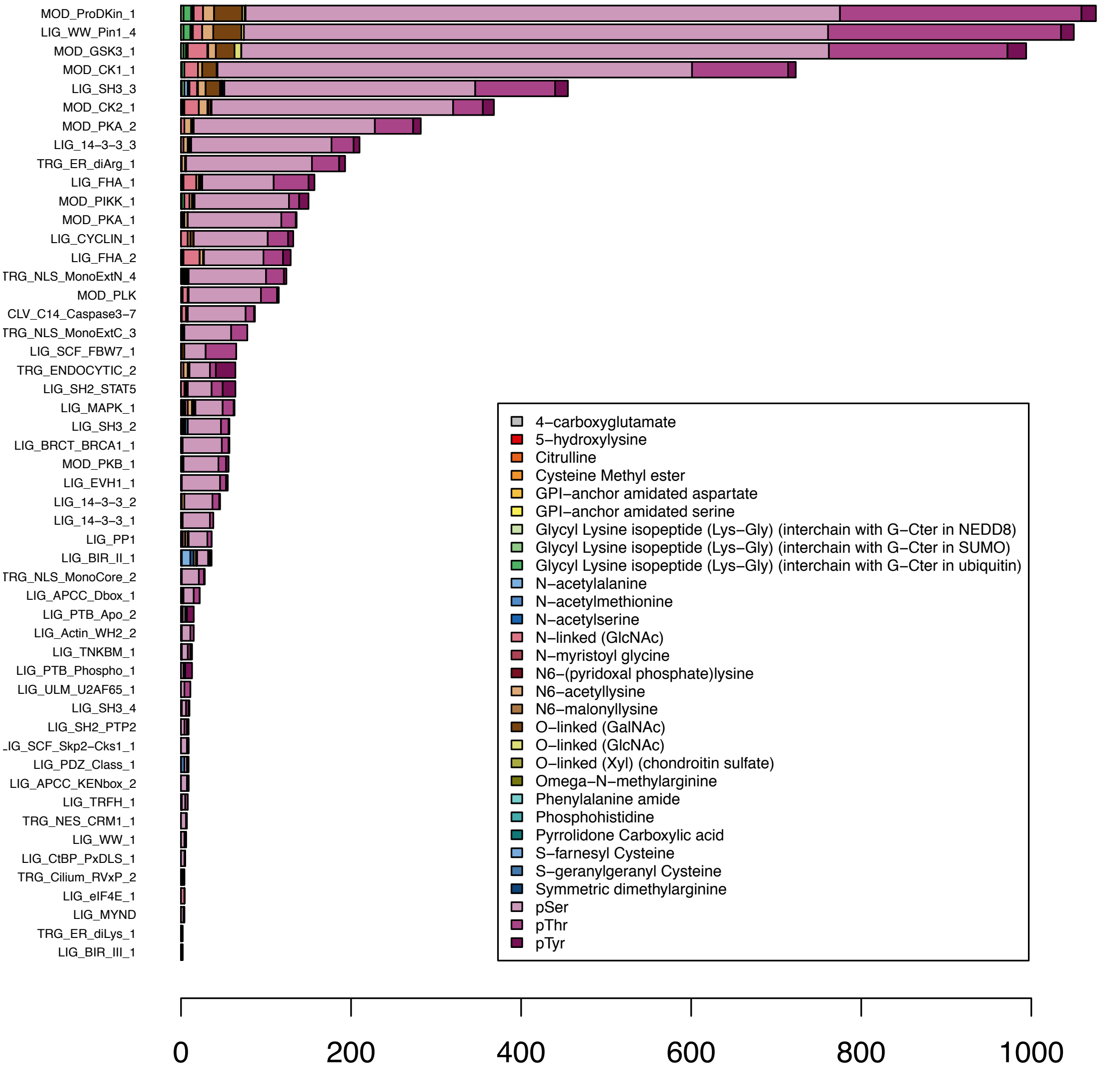


Figure 5: PTM site count around different SLiMs. Different modification sites are denoted by different colors. Abbreviations: pSer - phosphoserine; pThr - phosphothreonine; pTyr - phosphotyrosine.

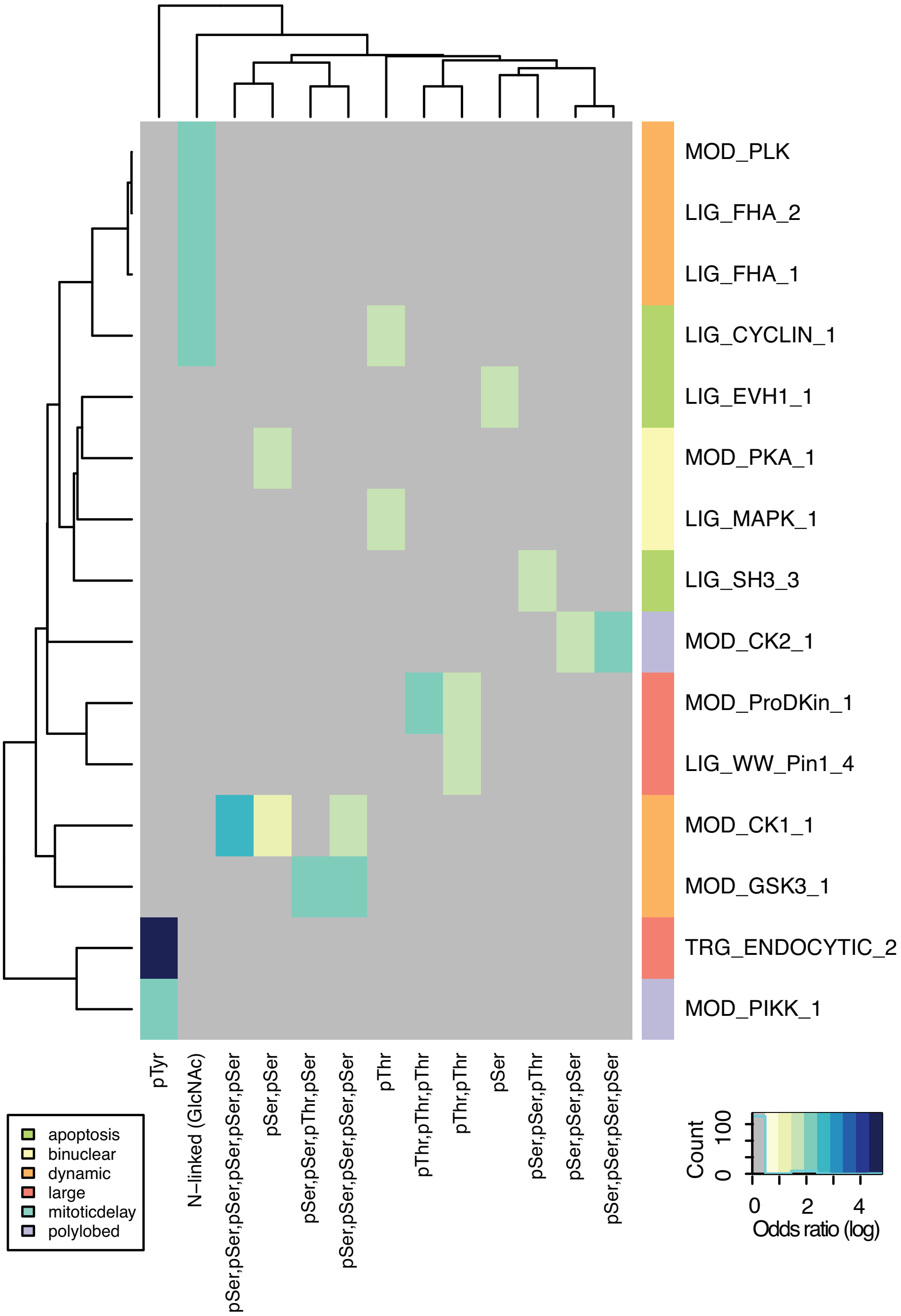


Figure 6: PTM patterns enriched around different SLiMs. The heat map shows the natural logarithm of the odds ratio, mapped on a color gradient from yellow to blue. Grey indicates no enrichment. Colored rectangles in front of the SLiM names indicate the phenotypes where these SLiMs are enriched. Abbreviations: pSer - phosphoserine; pThr - phosphothreonine; pTyr - phosphotyrosine.

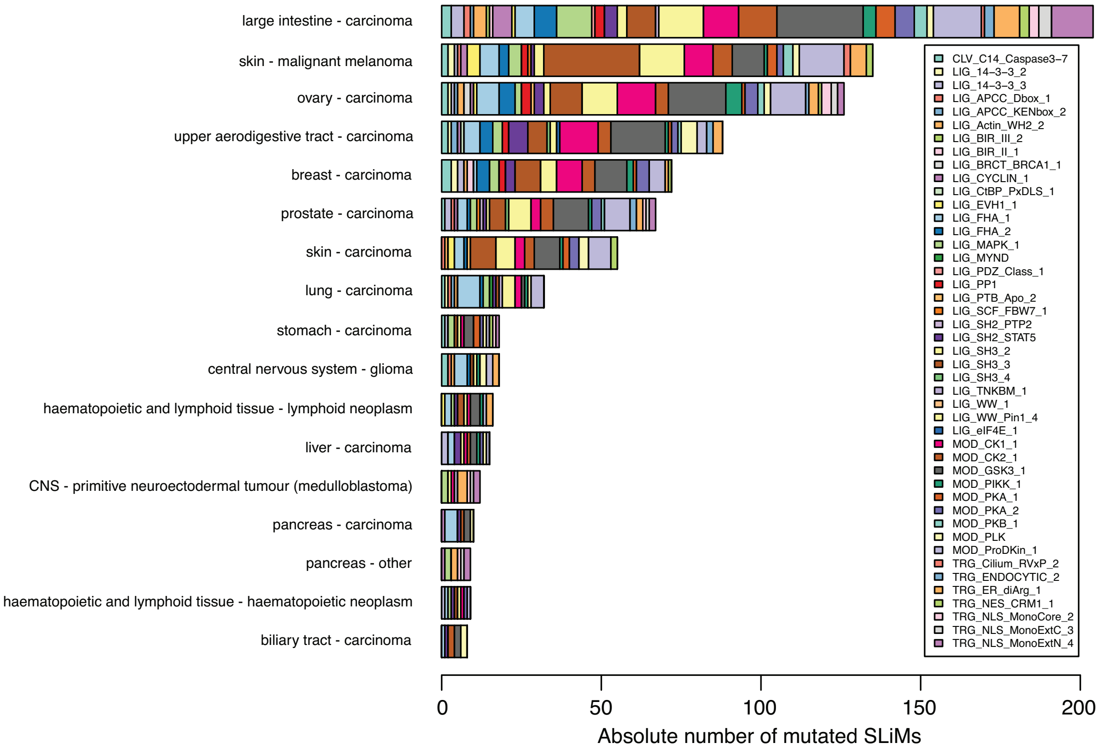


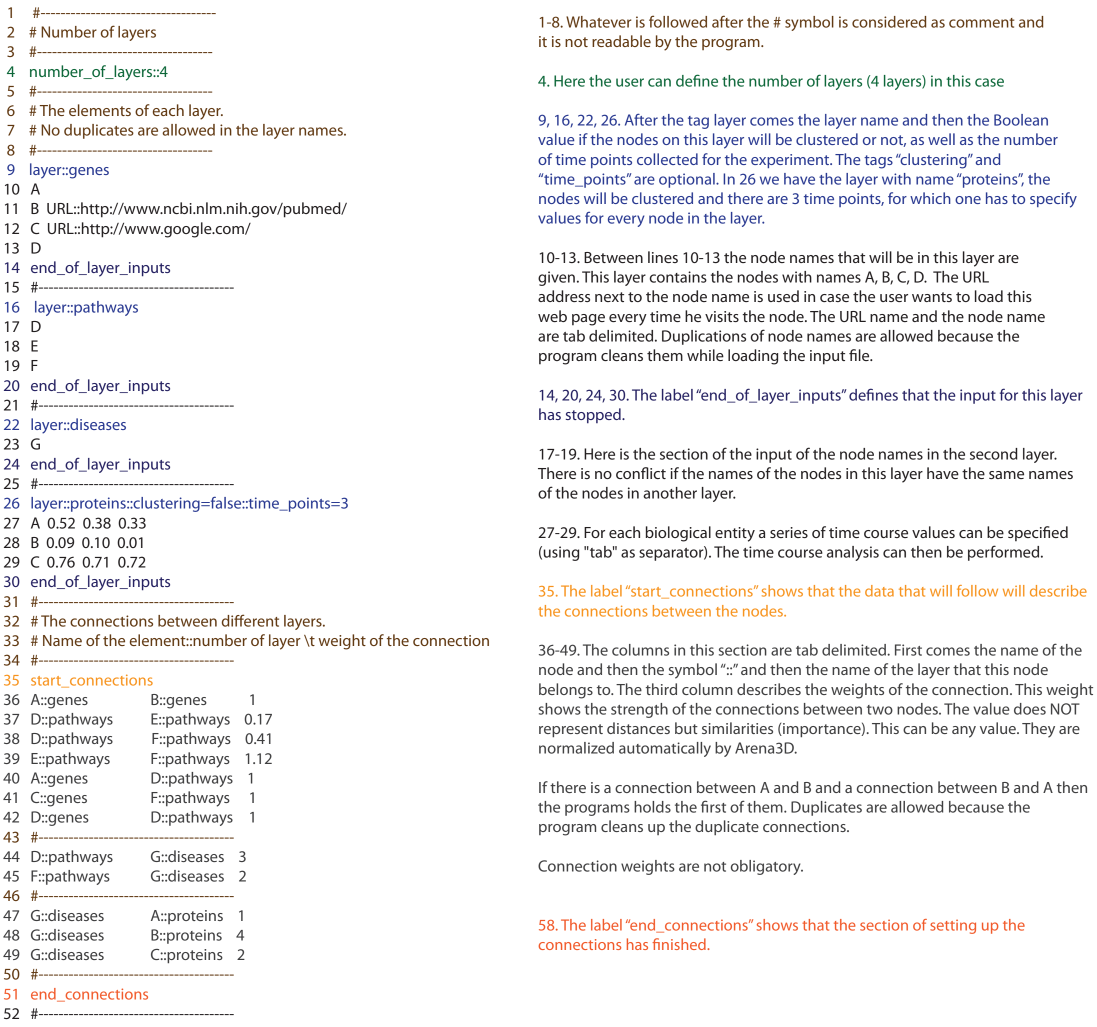
Figure 7: Diseases with the highest number of SLiM mutation evidence, according to the COSMIC database. For each type of cancer, the bars show the number of cases when a specific SLiM site is mutated, from the analyzed dataset. CNS: central nervous system.

Appendix B

This appendix contains the supplementary information for Chapter 3.

Gene name	Description
Arid3b	AT rich interactive domain 3B (Bright like)
Cdc2a	Cyclin-dependent kinase 1
Elys	AT hook containing transcription factor 1
Errs	estrogen-related receptor beta
Ewsr1	RNA-binding protein EWS
Gata2b	GATA-binding protein 2b
Hdac2	histone deacetylase 2
Ilf2	interleukin enhancer binding factor 2
Mybbp1a	MYB binding protein (P160) 1a
Nac1	nucleus accumbens associated 1, BEN and BTB (POZ) domain containing
Nanog	homeobox protein NANOG
Oct4	POU domain, class 5, transcription factor 1
Prmt1	protein arginine methyltransferase 1
Rex1	REX1, RNA exonuclease 1 homolog
Rif1	RAP1 interacting factor homolog (yeast)
Rnf2	ring finger protein 2
Sall1	sal-like 1 (Drosophila)
Sall4	sal-like 4 (Drosophila)
Smarcad1	SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1
Smarcc1	SWI/SNF related, matrix associated actin dependent regulator of chromatin, subfamily c, member 1
Trim28	tripartite motif containing 28
Wdr18	WD repeat domain 18
Yy1	YY1 transcription factor
Zfp198	zinc finger protein 198
Zfp219	zinc finger protein 219
Zfp281	zinc finger protein 281

Table 1: List of genes involved in the ESC core network, as described in Lu et al. (2009). Information about the genes has been extracted from the GeneCards database Rebhan et al. (1997).



```
1   #-----------------------------------
2   # Number of layers
3   #-----------------------------------
4   number_of_layers::4
5   #-----------------------------------
6   # The elements of each layer.
7   # No duplicates are allowed in the layer names.
8   #-----------------------------------
9   layer::genes
10  A
11  B  URL::http://www.ncbi.nlm.nih.gov/pubmed/
12  C  URL::http://www.google.com/
13  D
14  end_of_layer_inputs
15  #------------------------------------------
16   layer::pathways
17  D
18  E
19  F
20  end_of_layer_inputs
21  #------------------------------------------
22  layer::diseases
23  G
24  end_of_layer_inputs
25  #------------------------------------------
26  layer::proteins::clustering=false::time_points=3
27  A  0.52  0.38  0.33
28  B  0.09  0.10  0.01
29  C  0.76  0.71  0.72
30  end_of_layer_inputs
31  #------------------------------------------
32  # The connections between different layers.
33  # Name of the element::number of layer \t weight of the connection
34  #------------------------------------------
35  start_connections
36  A::genes          B::genes        1
37  D::pathways       E::pathways   0.17
38  D::pathways       F::pathways   0.41
39  E::pathways       F::pathways   1.12
40  A::genes          D::pathways    1
41  C::genes          F::pathways    1
42  D::genes          D::pathways    1
43  #------------------------------------------
44  D::pathways       G::diseases   3
45  F::pathways       G::diseases   2
46  #------------------------------------------
47  G::diseases       A::proteins   1
48  G::diseases       B::proteins   4
49  G::diseases       C::proteins   2
50  #------------------------------------------
51  end_connections
52  #------------------------------------------
```

1-8. Whatever is followed after the # symbol is considered as comment and it is not readable by the program.

4. Here the user can define the number of layers (4 layers) in this case

9, 16, 22, 26. After the tag layer comes the layer name and then the Boolean value if the nodes on this layer will be clustered or not, as well as the number of time points collected for the experiment. The tags "clustering" and "time_points" are optional. In 26 we have the layer with name "proteins", the nodes will be clustered and there are 3 time points, for which one has to specify values for every node in the layer.

10-13. Between lines 10-13 the node names that will be in this layer are given. This layer contains the nodes with names A, B, C, D. The URL address next to the node name is used in case the user wants to load this web page every time he visits the node. The URL name and the node name are tab delimited. Duplications of node names are allowed because the program cleans them while loading the input file.

14, 20, 24, 30. The label "end_of_layer_inputs" defines that the input for this layer has stopped.

17-19. Here is the section of the input of the node names in the second layer. There is no conflict if the names of the nodes in this layer have the same names of the nodes in another layer.

27-29. For each biological entity a series of time course values can be specified (using "tab" as separator). The time course analysis can then be performed.

35. The label "start_connections" shows that the data that will follow will describe the connections between the nodes.

36-49. The columns in this section are tab delimited. First comes the name of the node and then the symbol "::" and then the name of the layer that this node belongs to. The third column describes the weights of the connection. This weight shows the strength of the connections between two nodes. The value does NOT represent distances but similarities (importance). This can be any value. They are normalized automatically by Arena3D.

If there is a connection between A and B and a connection between B and A then the programs holds the first of them. Duplicates are allowed because the program cleans up the duplicate connections.

Connection weights are not obligatory.

58. The label "end_connections" shows that the section of setting up the connections has finished.

Figure 8: The format of files readable by Arena3D.

Appendix B

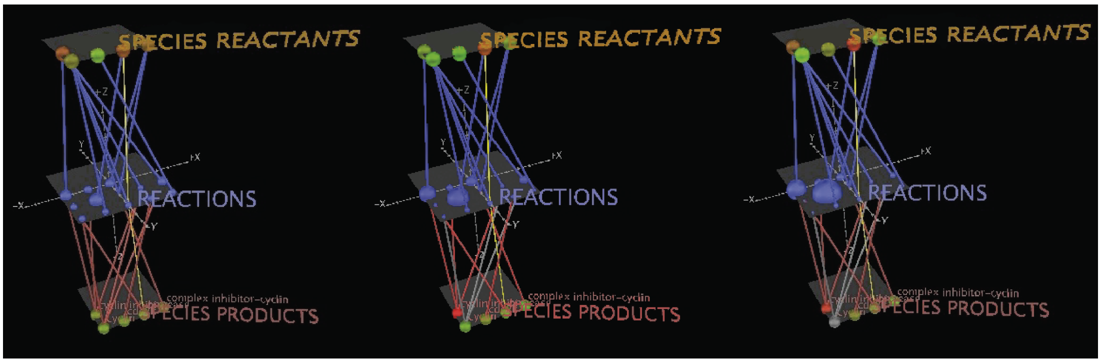


Figure 9: Visualization of how time lapse SBML models would be simulated within Arena3D. The figure depicts the state of the network at three different time points, from left to right. The reaction flow is followed through increases and decreases in size of the corresponding reaction nodes: increase in node volume means the reaction is relatively more active at the respective time point. The layers of species and products contain the reaction participants represented as nodes, and the colors indicate their concentrations, on a scale from green to red (lowest to highest). At every time point, one is able to follow the fluxes through the reactions and the corresponding species that are being converted. The figure uses random data for prototype purposes only and does not simulate an actual biological model.

Appendix C

This appendix contains the supplementary information for Chapter 4.

Appendix C

#Gene name	#Phenotype name					#Time points
gene1	phenotype1	0.13	0.22	1.30	1.55	0.98
gene1	phenotype2	1.56	0.74	1.80	0.51	-0.84
gene1	phenotype3	0.59	0.12	0.06	-0.50	-1.28
gene1	phenotype4	-0.11	-0.89	0.04	0.17	-0.90
gene2	phenotype1	1.43	0.92	1.22	-1.13	0.75
gene2	phenotype2	-1.03	-0.47	-1.26	0.84	-0.81
gene2	phenotype3	0.99	0.49	1.27	1.51	1.38
gene2	phenotype4	0.64	-0.10	1.57	-1.55	-0.01
gene3	phenotype1	1.58	0.97	1.02	1.12	0.99
gene3	phenotype2	2.10	1.86	1.45	1.07	0.91
gene3	phenotype3	-0.72	0.04	-0.93	-0.56	-0.62
gene3	phenotype4	0.07	0.11	0.24	0.23	0.22

Table 2: Input file format for the main dataset of phenotypic values measured for multiple time points. The columns are separated by white space.

A. Enrichment				***B. Interactions***	
#GO id	**#GO name**	**#p-value**	**#Genes enriched**	**#Partner 1**	**#Partner 2**
32653	mitosis	0.0012	gene1|gene2|gene4	32653	54588
54588	nuclear division	0.0034	gene2|gene5	32653	54588
13857	metaphase plate congression	0.0049	gene1|gene3	32653	54588

Table 3: Gene ontology enrichment (A) and network interactions (B) file format. The columns are separated by tabs.

Gene	Description
Angptl4	induced under hypoxic conditions in endothelial cells; target of peroxisome proliferation activators
Areg	autocrine growth factor as well as a mitogen for a broad range of target cells
Crem	transcriptional regulator that binds the cAMP response element
Itgad	receptor for ICAM3 and VCAM1; role in atherosclerosis
Npas4	transcriptional activator in the presence of ARNT
Phactr3	nuclear scaffolding in proliferating cells
Pla2g3	catalyzes the calcium-dependent hydrolysis of the 2-acyl groups in 3-sn-phosphoglycerides
Plekhf1	may induce apoptosis through the lysosomal-mitochondrial pathway
Rasd1	small GTPase
Tekt4	structural component of ciliary and flagellar microtubules

Table 4: Relatively lowly expressed genes belonging to the core network similarly regulated by drugs. The descriptions were taken from UniProt The UniProt Consortium (2012).

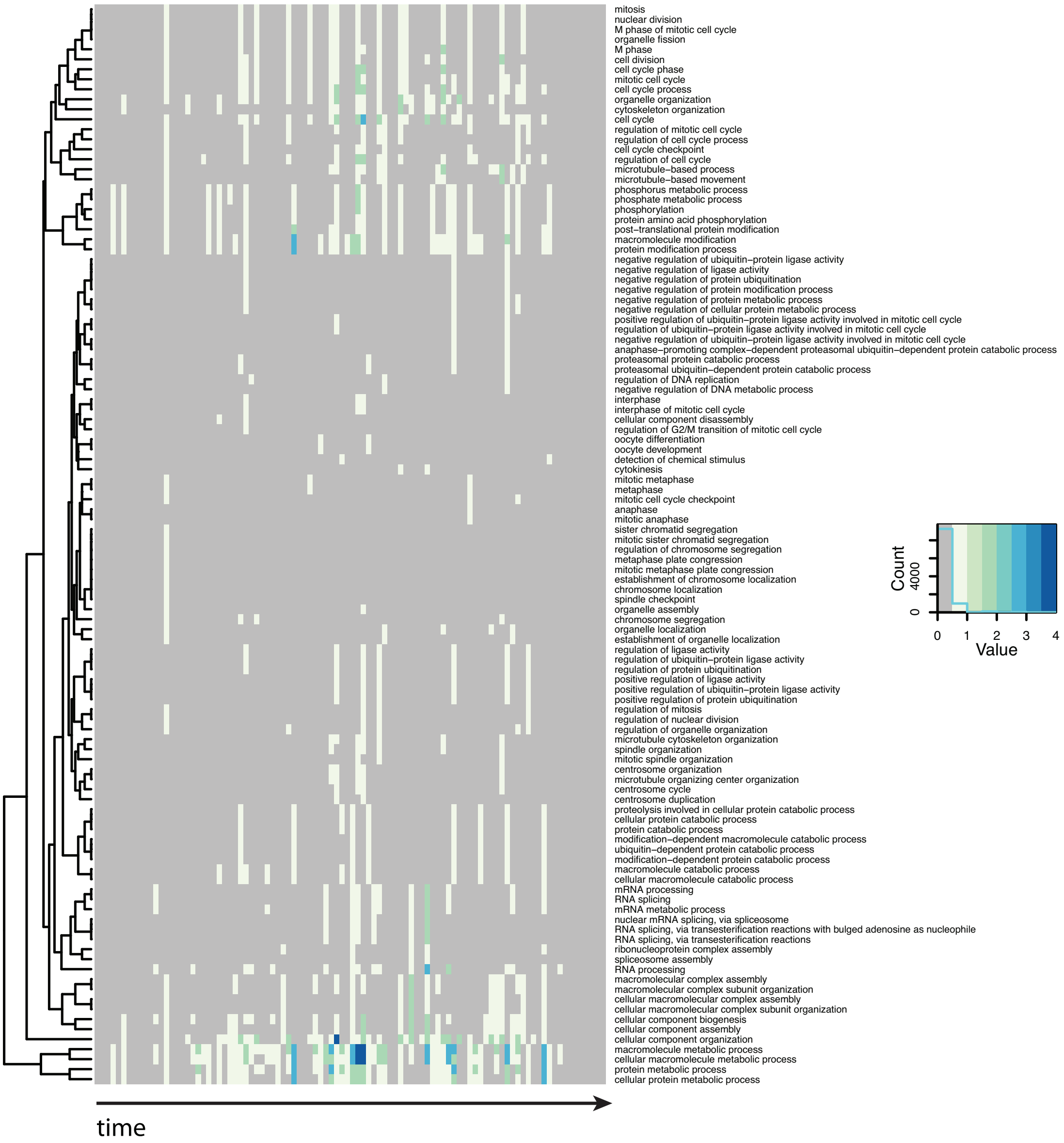


Figure 10: Timeline of enriched biological processes for the entire dataset of genes essential to cell division. The colors map on a gradient to the number of genes that are involved in phenotypic transitions at a particular time point and are enriched for the respective process. The time axis spans approximately two cell cycles, from 0 to 48 hours in half an hour time intervals.

Appendix C

Gene	Description
Dusp1	dual specificity phosphatase that dephosphorylates MAP kinase MAPK1/ERK2
Dusp14	involved in the inactivation of MAP kinases
Egr4	transcriptional regulator
Gjb6	gap junction protein
Homer1	postsynaptic density scaffolding protein
Pim3	proto-oncogene with serine/threonine kinase activity
Sgk1	serine/threonine-protein kinase with important role in cellular stress response
Slc2a1	facilitative glucose transporter

Table 5: Relatively highly expressed genes belonging to the core network similarly regulated by drugs. The descriptions were taken from UniProt The UniProt Consortium (2012).

Gene	Description
Egr2	sequence-specific DNA-binding transcription factor
Fosl2	controls osteoclast survival and size
Hif3a	involved in adaptive response to hypoxia
Mest	mesoderm specific transcript
Tnfrsf25	mediates activation of NF-kappa-B and induces apoptosis
Zfp189	may be involved in transcriptional regulation

Table 6: Relatively lowly expressed variable genes in the network similarly regulated by drugs. The descriptions were taken from UniProt The UniProt Consortium (2012).

Gene	Description
Cdkn1a	binds to and inhibits cyclin-dependent kinase activity, blocking cell cycle progression
Fos	role in signal transduction, cell proliferation and differentiation
Fosb	interacts with Jun proteins enhancing their DNA binding activity
Egr2	sequence-specific DNA-binding transcription factor
Midn	may be involved in regulation of genes related to neurogenesis in the nucleolus
Nostrin	endothelial nitric oxide synthase traffic inducer
Polr3e	DNA-directed RNA polymerase III 80 kDa polypeptide
Sult1a1	thermostable phenol sulfotransferase

Table 7: Relatively highly expressed variable genes in the network similarly regulated by drugs. The descriptions were taken from UniProt The UniProt Consortium (2012).

Publications

Secrier, M., Schneider, R. Visualizing time-related data in biology, a review. *Briefings in Bioinformatics.* 2013. (in press)

Secrier, M., Pavlopoulos, G. A., Aerts, J., Schneider, R. Arena3D: visualizing time-driven phenotypic differences in biological systems. *BMC Bioinformatics.* 2012 Mar 22;13:45. (highly accessed)

Access statistics:

· Last 30 days: 126 accesses

· Last 365 days: 2911 accesses

· All time: 2911 accesses

Pavlopoulos, G.A., **Secrier, M.**, Moschopoulos, C.N., Soldatos, T.G., Kossida, S., Aerts, J., Schneider, R., Bagos, P.G. Using graph theory to analyze biological networks. *BioData Mining.* 2011 Apr 28;4:10. (highly accessed)

Access statistics:

· Last 30 days: 721 accesses

· Last 365 days: 7696 accesses

· All time: 13836 accesses

· 2nd most accessed article in the journal

Secrier, M., Schneider, R. PhenoTimer: mapping time-resolved phenotypic landscapes. *submitted*

References

Adami, C., Ofria, C., and Collier, T. C. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences*, 97(9):4463–4468. 2

Agresti, A. and Coull, B. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, 52:119–126. 59

Asai, Y., Abe, T., Okita, M., Okuyama, T., Yoshioka, N., Yokoyama, S., Nagaku, M., Hagihara, K.-I., and Kitano, H. (2012). Multilevel modeling of physiological systems and simulation platform: PhysioDesigner, Flint and Flint K3 service. *2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet*, 0:215–219. 67

Ashburner, M. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25:25–29. 88, 137

Atallah, J. and Larsen, E. (2009). Chapter 3 Genotype - phenotype mapping: developmental biology confronts the toolkit paradox. In *International Review Of Cell and Molecular Biology*, volume 278, pages 119 – 148. Academic Press. 2

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48. 50

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(1):289–300. 58

Berson, E. L. (1996). Retinitis pigmentosa: unfolding its mystery. *Proceedings of the National Academy of Sciences*, 93(10):4526–4528. 29

Bertin, J. (1984). *Semiology of graphics: diagrams, networks, maps.* University of Wisconsin Press. 103

Bicknell, K. A. and Brooks, G. (2008). Reprogramming the cell cycle machinery to treat cardiovascular disease. *Current Opinion in Pharmacology*, 8(2):193–201. 6

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 19. 67

Bochner, B. B. (2003). New technologies to assess genotype-phenotype relationships. *Nature Reviews Genetics*, (4):309–314. 139

Bornstein, B. J., Keating, S. M., Jouraku, A., and Hucka, M. (2008). LibSBML: an API library for SBML. *Bioinformatics*, 24(6):880–881. 64

Bulavin, D. V., Higashimoto, Y., Demidenko, Z. N., Meek, S., Graves, P., Phillips, C., Zhao, H., Moody, S. A., Appella, E., Piwnica-Worms, H., and Fornace, A. J.

(2003). Dual phosphorylation controls Cdc25 phosphatases and mitotic entry. *Nature Cell Biology*, 5(6):545–551. 6

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822. 3

Case, D. A., Cheatham, T. E., r., Darden, T., Gohlke, H., Luo, R., Merz, K. M., J., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–88. 16

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(suppl 1):D532–D539. 24

Chatr-aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., ODonnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K., and Tyers, M. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Research*, 41(D1):D816–D823. 24

Chaudhuri, B. and Dutta, S. (1986). Interactive curve drawing by segmented Bezier approximation with a control parameter. *Pattern Recognition Letters*, 4(3):171 – 176. 94

Chen, W. W., Niepel, M., and Sorger, P. K. (2010). Classic and contemporary approaches to modeling biochemical reactions. *Genes & Development*, 24(17):1861–1875. 13

Chica, C., Labarga, A., Gould, C., Lopez, R., and Gibson, T. (2008). A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, 9(1):229. 22, 23

Ciliberti, S., Martin, O. C., and Wagner, A. (2007). Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Computational Biology*, 3(2):e15. 4

Clapperton, J. A., Manke, I. A., Lowery, D. M., Ho, T., Haire, L. F., Yaffe, M. B., and Smerdon, S. J. (2004). Structure and mechanism of BRCA1 BRCT domain recognition of phosphorylated BACH1 with implications for cancer. *Nature Structural & Molecular Biology*, 11(6):512–518. 29

Coordinators, N. C. B. I. Resource (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. 99

Crippen, G. M. and Havel, T. F. (1988). Distance geometry and molecular conformation. *John Wiley and Sons, New York*. 56

Curtis, R. E., Yuen, A., Song, L., Goyal, A., and Xing, E. P. (2011). TVNViewer: an interactive visualization tool for exploring networks that change over time or space. *Bioinformatics*, 27(13):1880–1. 94, 136

Davey, N. E., Cowan, J. L., Shields, D. C., Gibson, T. J., Coldwell, M. J., and Edwards, R. J. (2012a). SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Research*, 40(21):10628–10641. 22, 24

Davey, N. E., Haslam, N. J., Shields, D. C., and Edwards, R. J. (2010).

SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic Acids Research*, 38:534–539. 22

Davey, N. E., Haslam, N. J., Shields, D. C., and Edwards, R. J. (2011). SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Research*, 39(suppl 2):W56–W60. 22

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T. J. (2012b). Attributes of short linear motifs. *Molecular BioSystems*, 8(1):268–281. 10, 21

Davis, S. B., Bevan, E., and Kudikov, A. (2010). Just in time: defining historical chronographics. In *Proceedings of the 2010 international conference on Electronic Visualisation and the Arts*, EVA'10, pages 355–362, Swinton, UK. British Computer Society. 12

de Haan, J. (2006). How emergence arises. *Ecological Complexity*, 3(4):293 – 301. Complexity and Ecological Economics. 1

Deretic, D., Schmerl, S., Hargrave, P. A., Arendt, A., and McDowell, J. H. (1998). Regulation of sorting and post-Golgi trafficking of rhodopsin by its C-terminal sequence QVS(A)PA. *Proceedings of the National Academy of Sciences*, 95(18):10620–10625. 41

Deribe, Y. L. L., Pawson, T., and Dikic, I. (2010). Post-translational modifications in signal integration. *Nature Structural & Molecular Biology*, 17(6):666–672. 10, 36

D'Eustachio, P. (2011). Reactome knowledgebase of human biological pathways and processes. *Methods in Molecular Biology (Clifton, N.J.)*, 694:49–61. 24, 137

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Research*, 39(Database issue):D261–7. 25

Dinkel, H., Michael, S., Weatheritt, R. J., Davey, N. E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A., Jodicke, L., Dammert, M. A., Schroeter, C., Hammer, M., Schmidt, T., Jehl, P., McGuigan, C., Dymecka, M., Chica, C., Luck, K., Via, A., Chatr-Aryamontri, A., Haslam, N., Grebnev, G., Edwards, R. J., Steinmetz, M. O., Meiselbach, H., Diella, F., and Gibson, T. J. (2012). ELM–the database of eukaryotic linear motifs. *Nucleic Acids Research*, 40(Database issue):D242–51. 21

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128. 13

Dosztnyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434. 24

Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, 18(6):756 – 764. 10

Durrant, J. and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1):71. 4

Eden, E., Geva-Zatorsky, N., Issaeva, I., Cohen, A., Dekel, E., Danon, T., Cohen, L., Mayo, A., and Alon, U. (2011). Proteome half-life dynamics in living human cells. *Science*, 331(6018):764–768. 4

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–10. 16, 112

Eikenboom, J., Matsushita, T., Reitsma, P., Tuley, E., Castaman, G., Briet, E., and Sadler, J. (1996). Dominant type 1 von Willebrand disease caused by mutated cysteine residues in the D3 domain of von Willebrand factor. *Blood*, 88(7):2433–2441. 41

Ernst, J. and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7. 16, 136

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49. 4

Evans, T., Rosenthal, E. T., Youngblom, J., Distel, D., and Hunt, T. (1983). Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell*, 33(2):389 – 396. 6

Feeney, D. (2007). *Caesar's calendar: ancient time and the beginnings of history (Sather Classical Lectures)*. University of California Press. 12

Ferguson, S. (1991). *The 1753 Carte chronographique of Jacques Barbeu-Dubourg.* Friends of the Princeton University Library. 12

Finocchiaro, G., Wang, T., Hoffmann, R., Gonzalez, A., and Wade, R. (2003). DSMM: a database of simulated molecular motions. *Nucleic Acids Research*, 31(1):456–457. 5

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S. M. J. (2013). Ensembl 2013. *Nucleic Acids Research*, 41(D1):D48–D55. 99

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R., and Futreal, P. A. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*, 39(Database issue):D945–D950. 26, 88

Foster, I. (2008). Cancer: A cell cycle defect. *Radiography*, 14(2):144 – 149. 6

Fourest-Lieuvin, A., Peris, L., Gache, V., Garcia-Saez, I., Juillan-Binard, C., Lantez, V., and Job, D. (2006). Microtubule regulation in mitosis: tubulin phosphorylation by the cyclin-dependent kinase Cdk1. *Molecular Biology of the Cell*, 17(3):1041–1050. 32

Friendly, M. (2002). Visions and re-visions of Charles Joseph Minard. *Journal of Educational and Behavioral Statistics*, 27(1). 12

Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96(8):1254–1265. 16

Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1(5):159 – 162. 16, 67

Furusato, B., Shaheduzzaman, S., Petrovics, G., Dobi, A., Seifert, M., Ravindranath, L., Nau, M. E., Werner, T., Vahey, M., McLeod, D. G., Srivastava, S., and Sesterhenn, I. A. (2008). Transcriptome analyses of benign and malignant prostate epithelial cells in formalin-fixed paraffin-embedded whole-mounted radical prostatectomy specimens. *Prostate Cancer Prostatic Diseases*, 11(2):194–7. 91

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, 23(8):950–956. 23

Gardner, T. S., Dolnik, M., and Collins, J. J. (1998). A theory for controlling cell cycle dynamics using a reversibly binding inhibitor. *Proceedings of the National Academy of Sciences*, 95(24):14190–14195. 65

Gauthier, N. P., Jensen, L. J., Wernersson, R., Brunak, S., and Jensen, T. S. (2010). Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Research*, 38(Database issue):D699–702. 111, 120

Gauthier, N. P., Larsen, M. E., Wernersson, R., de Lichtenberg, U., Jensen, L. J., Brunak, S., and Jensen, T. S. (2008). Cyclebase.org–a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Research*, 36(Database issue):D854–9. 16, 109, 111, 120

Ge, S., Skaar, J. R., and Pagano, M. (2009). APC/C- and Mad2-mediated degradation of Cdc20 during spindle checkpoint activation. *Cell Cycle (Georgetown, Tex.)*, 8(1):167–171. 32

Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., and Gavin, A. C. (2010). Visualization of omics data for systems biology. *Nature Methods*, 7(3):S56–S68. 11

Ghosh, S., Matsuoka, Y., Asai, Y., Hsin, K.-Y. Y., and Kitano, H. (2011). Software for systems biology: from tools to integrated platforms. *Nature Reviews Genetics*, 12(12):821–832. 66

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Researchearch*, 15(10):1451–1455. 67

Gibson, T. J. (2009). Cell regulation: determined to signal discrete cooperation. *Trends in Biochemical Sciences*, 34(10):471 – 482. 9, 50

Glotzer, M., Murray, A. W., and Kirschner, M. W. (1991). Cyclin is degraded by the ubiquitin pathway. *Nature*, 349(6305):132–138. 29

Goecks, J., Nekrutenko, A., Taylor, J., and Team, T. G. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86+. 67

Greenspan, R. J. (2001). The flexible genome. *Nature Reviews Genetics*, (5):383387. 2

Gstaiger, M. and Aebersold, R. (2013). Genotype-phenotype relationships in light of a modular protein interaction landscape. *Molecular BioSystems*. 140

Guettler, S., LaRose, J., Petsalaki, E., Gish, G., Scotter, A., Pawson, T., Rottapel, R., and Sicheri, F. (2011). Structural basis and sequence rules for substrate recognition by tankyrase explain the basis for cherubism disease. *Cell*, 147(6):1340 – 1354. 29

Haber, D. A. and Settleman, J. (2007). Cancer: drivers and passengers. *Nature*, 446(7132):145–146. 50

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue). 26, 89

Hartwell, L. H. and Weinert, T. A. (1989). Checkpoints: controls that ensure the order of cell cycle events. *Science (New York, N.Y.)*, 246(4930):629–634. 6

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer, second edition. 102

Henry, C. S., Overbeek, R., Xia, F., Best, A. A., Glass, E., Gilbert, J., Larsen, P., Edwards, R., Disz, T., Meyer, F., Vonstein, V., DeJongh, M., Bartels, D., Desai, N., D'Souza, M., Devoid, S., Keegan, K. P., Olson, R., Wilke, A., Wilkening, J., and Stevens, R. L. (2011). Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1810(10):967 – 977. 3

Herraez, A. (2006). Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255–61. 16

Hester, R. L., Iliescu, R., Summers, R., and Coleman, T. G. (2011). Systems biology and integrative physiological modelling. *The Journal of Physiology*, 589(5):1053–1060. 13

Hidalgo, C., Blumm, N., Barabasi, A.-L., and Christakis, N. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):e1000353. 4

Hitchler, M. J. and Rice, J. C. (2011). Genome-wide epigenetic analysis of human pluripotent stem cells by ChIP and ChIP-Seq. *Methods in Molecular Biology*, 767:253–67. 91

Holmes, M. V., Shah, T., Vickery, C., Smeeth, L., Hingorani, A. D., and Casas,

J. P. (2009). Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. *PLoS ONE*, 4(12):e7960. 141

Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research*, 40(D1):D261–D270. 25

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., and Yon Rhee, S. (2008). Big data: the future of biocuration. *Nature*, 455(7209):47–50. 1

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., the rest of the SBML Forum: Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J. H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531. 60, 61

Jackman, M., Lindon, C., Nigg, E., and J., P. (2003). Active cyclin B1-Cdk1 first appears on centrosomes in prophase. *Nature Cell Biology*, 5(2):143–8. 32

Johannsen, W. (1911). The genotype conception of heredity. *The American Naturalist*, 45(531):pp. 129–159. 2

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, 2nd edition. 60

Kaizu, K., Ghosh, S., Matsuoka, Y., Moriya, H., Shimizu-Yoshida, Y., and Kitano, H. (2010). A comprehensive molecular interaction map of the budding yeast cell cycle. *Molecular Systems Biology*, 6:415. 8, 64

Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. *Database*, 2011(0):bar049. 109, 111, 120

Kastan, M. B. and Bartek, J. (2004). Cell-cycle checkpoints and cancer. *Nature*, 432(7015):316–323. 7

Kawabata, T., Ota, M., and Nishikawa, K. (1999). The Protein Mutant Database. *Nucleic Acids Research*, 27(1):355–357. 50

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, 181(4610):662–666. 9

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., and Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846. 24

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(suppl 1):D767–D772. 24

Kholodenko, B. N., Hancock, J. F., and Kolch, W. (2010). Signalling ballet in space and time. *Nature Reviews Molecular Cell Biology*, 11(6):414–426. 4

Kincaid, R., Kuchinsky, A., and Creech, M. (2008). VistaClara: an expression browser plug-in for Cytoscape. *Bioinformatics*, 24(18):2112–4. 16, 86, 136

Koch, C. (2012). Modular biological complexity. *Science*, 337(6094):531–532. 1

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69. 87

Kruglyak, S. and Tang, H. (2001). A new estimator of significance of correlation in time series data. *Journal of Computational Biology*, 8:463–70. 58

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–45. 13, 94, 136

Kuhn, M., Szklarczyk, D., Franceschini, A., Campillos, M., von Mering, C., Jensen, L. J., Beyer, A., and Bork, P. (2010). STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Research*, 38(suppl 1):D552–D556. 47, 49

Lander, E. S. (1999). Array of hope. *Nature Genetics*, 21(1 Suppl):3–4. 3

Landstrom, A. P., Weisleder, N., Batalden, K. B., Bos, J. M., Tester, D. J., Ommen, S. R., Wehrens, X. H., Claycomb, W. C., Ko, J.-K., Hwang, M., Pan, Z., Ma, J., and Ackerman, M. J. (2007). Mutations in JPH2-encoded junctophilin-2 associated with hypertrophic cardiomyopathy in humans. *Journal of Molecular and Cellular Cardiology*, 42(6):1026 – 1035. 43

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J. L., and Hucka, M. (2006). BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*, 34(suppl 1):D689–D691. 65, 88

Lee, J.-H., You, J., Dobrota, E., and Skalnik, D. G. (2010). Identification and characterization of a novel human PP1 phosphatase complex. *Journal of Biological Chemistry*, 285(32):24466–24476. 32

Lee, M. G. and Nurse, P. (1987). Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2. *Nature*, 327(6117):31–35. 6

Lefrancois, P., Zheng, W., and Snyder, M. (2010). ChIP-Seq using high-throughput DNA sequencing for genome-wide identification of transcription factor binding sites. *Methods Enzymology*, 470:77–104. 91

Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17(4):183 – 189. 4

Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–8. 16, 111

Li, K. K. C. and Lee, K. A. W. (2000). Transcriptional activation by the Ewing's Sarcoma (EWS) oncogene can be cis-repressed by the EWS RNA-binding domain. *Journal of Biological Chemistry*, 275(30):23053–23058. 73

Li, S., Iakoucheva, L. M., Mooney, S. D., and Radivojac, P. (2010). Loss of post-translational modification sites in disease. *Pacific Symposium on Biocomputing*, pages 337–347. 36

Lopes, T., Schaefer, M., Shoemaker, J., Matsuoka, Y., Fontaine, J., Neumann, G., Andrade-Navarro, M., Kawaoka, Y., and Kitano, H. (2011). Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics*. 89

Lu, R., Markowetz, F., Unwin, R., Leek, J., Airoldi, E., MacArthur, B., Lachmann, A., Rozov, R., Ma'ayan, A., Boyer, L., Troyanskaya, O., Whetton, A., and Lemischka, I. (2009). Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature*, 462(7271):358–62. 69, 152

MacArthur, B. D., Lachmann, A., Lemischka, I. R., and Ma'ayan, A. (2010). GATE: software for the analysis and visualization of high-dimensional time series expression data. *Bioinformatics*, 26(1):143–4. 16, 86, 136

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–9. 109

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 35(Database issue). 99

Malumbres, M. and Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. *Nature Reviews Cancer*, 9(3):153–66. 113

Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176. PMID: 20647212. 3

Marian, A. J. (2012). Molecular genetic studies of complex phenotypes. *Translational Research*, 159(2):64 – 79. 3

Marwan, N., Romano, M. C., Thiel, M., and Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(56):237 – 329. 13

Masry, E. (2011). The estimation of the correlation coefficient of bivariate data under dependence: convergence analysis. *Statistics & Probability Letters*, 81:1039–45. 58

May, K. M. and Hardwick, K. G. (2006). The spindle checkpoint. *Journal of Cell Science*, 119(20):4139–4142. 6

Mayr, E. (1963). *Animal species and evolution.* Belknap Press of Harvard University Press. 2

Mazzocchi, F. (2008). Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory. *EMBO Reports*, 9(1):10–14. 1

Mendel, G. (1965). *Experiments in plant hybridisation*. Harvard University Press, twenty-sixth printing, 1994 edition. 2

Meyer, M., Wong, B., Styczynski, M., and Pfister, H. (2010a). Pathline: a tool for comparative functional genomics. *Computer Graphics Forum (Proc. EuroVis)*, 29(3):1043–52. 86

Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., Carrillo, O., Camps, J., Fenollosa, C., Repchevsky, D., Gelpí, J. L., and Orozco, M. (2010b). MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure*, 18(11):1399 – 1409. 16

Minguez, P., Parca, L., Diella, F., Mende, D., Kumar, R., Helmer-Citterich, M., Gavin, A.-C., van Noort, V., and Bork, P. (2012). Deciphering a global network of functionally associated post-translational modifications. *Molecular Systems Biology*, 8(599). 48

Mir, S. E., Hamer, P. C. D. W., Krawczyk, P. M., Balaj, L., Claes, A., Niers, J. M., Tilborg, A. A. V., Zwinderman, A. H., Geerts, D., Kaspers, G. J., Vandertop, W. P., Cloos, J., Tannous, B. A., Wesseling, P., Aten, J. A., Noske, D. P., Noorden, C. J. V., and W T. (2010). In silico analysis of kinase expression identifies WEE1 as a gatekeeper against mitotic catastrophe in glioblastoma. *Cancer Cell*, 18(3):244 – 257. 6

Montojo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics*, 26(22):2927–8. 110, 117

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D., and Ferrin, T. E. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, 12(1):436. 86

Mosca, R., Pache, R. A., and Aloy, P. (2012). The role of structural disorder in the rewiring of protein interactions through evolution. *Molecular & Cellular Proteomics*, 11(7). 10

Mosconi, F., Julou, T., Desprat, N., Sinha, D. K., Allemand, J.-F., Croquette, V., and Bensimon, D. (2008). Some nonlinear challenges in biology. *Nonlinearity*, 21(8):T131. 1

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9 Suppl 1:S4. 110, 117

Musacchio, A. and Hardwick, K. (2002). The spindle checkpoint: structural insights into dynamic signalling. *Nature Reviews Molecular Cell Biology*, 3:731–741. 7

Nachtomy, O., Shavit, A., and Yakhini, Z. (2007). Gene expression and the concept of the phenotype. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 38(1):238 – 254. 3

Nakayama, K.-i. and Nakayama, K. (1998). Cip/Kip cyclin-dependent kinase inhibitors: brakes of the cell cycle engine during development. *BioEssays*, 20(12):1020–1029. 29

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., Masi, F. d., Gibson, T. J., Lewis, J., Serrano, L., and Russell, R. B. (2005). Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biology*, 3(12):e405. 10

Neduva, V. and Russell, R. B. (2005). Linear motifs: Evolutionary interaction switches. *FEBS Letters*, 579(15):3342 – 3345. 10

Neumann, B., Walter, T., Heriche, J.-K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., Cetin, C., Sieckmann, F., Pau, G., Kabbe, R., Wuensche, A., Satagopam, V., Schmitz, M. H. A., Chapuis, C., Gerlich, D. W., Schneider, R., Eils, R., Huber, W., Peters, J.-M., Hyman, A. A., Durbin, R., Pepperkok, R., and Ellenberg, J. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464:721–727. 8, 9, 17, 20, 21, 22, 47, 65, 74, 75, 92, 103, 109, 113, 139

Niida, H. and Nakanishi, M. (2006). DNA damage checkpoints in mammals. *Mutagenesis*, 21(1):3–9. 6

Novikoff, A. B. (1945). The concept of integrative levels and biology. *Science*, 101(2618):pp. 209–215. 1

Nurse, P. (2000). A long twentieth century of the cell cycle and beyond. *Cell*, 100(1):71 – 78. 6

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999).

KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1):29–34. 88, 99, 137

Ostaszewski, M., Eifes, S., and del Sol, A. (2012). Evolutionary conservation and network structure characterize genes of phenotypic relevance for mitosis in human. *PLoS ONE*, 7(5):e36488. 34

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stmpflen, V., Mewes, H.-W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834. 24

Palacios, E. H. and Weiss, A. (2004). Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. *Oncogene*, 23(48):7990–8000. 34

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N., and Brazma, A. (2011). ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research*, 39(Database issue):D1002–D1004. 88

Passani, L. A., Bedford, M. T., Faber, P. W., McGinnis, K. M., Sharp, A. H., Gusella, J. F., Vonsattel, J.-P., and MacDonald, M. E. (2000). Huntingtin's WW domain partners in Huntington's disease post-mortem brain fulfill genetic criteria for direct involvement in Huntington's disease pathogenesis. *Human Molecular Genetics*, 9(14):2175–2182. 29

Pavlopoulos, G., Secrier, M., Moschopoulos, C., Soldatos, T., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. (2011). Using graph theory to analyze biological networks. *BioData Mining*, 4(1):10. 111

Pavlopoulos, G., Soldatos, T., Barbosa-Silva, A., and Schneider, R. (2010). A reference guide for tree analysis and visualization. *BioData Mining*, 3(1):1. 13

Pavlopoulos, G. A., O'Donoghue, S. I., Satagopam, V. P., Soldatos, T. G., Pafilis, E., and Schneider, R. (2008). Arena3D: visualization of biological networks in 3D. *BMC Systems Biology*, 2:104. 55

Petsalaki, E. and Russell, R. B. (2008). Peptide-mediated interactions in biological systems: new discoveries and applications. *Current Opinion in Biotechnology*, 19(4):344 – 350. 49

Pfleger, C. M. and Kirschner, M. W. (2000). The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes & Development*, 14(6):655–665. 29

Piechota, M., Korostynski, M., Solecki, W., Gieryk, A., Slezak, M., Bilecki, W., Ziolkowska, B., Kostrzewa, E., Cymerman, I., Swiech, L., Jaworski, J., and Przewlocki, R. (2010). The dissection of transcriptional modules regulated by various drugs of abuse in the mouse striatum. *Genome Biology*, 11(5):R48. 109, 112, 125, 130

Przytycka, T. M., Singh, M., and Slonim, D. K. (2010). Toward the dynamic interactome: it's about time. *Briefings in Bioinformatics*, 11(1):15–29. 4

Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D. M., Ausiello, G., Brannetti, B., Costantini, A., Ferre, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, L., Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, L., Kuster, B., Helmer-Citterich, M., Hunter, W. N., Aasland, R., and Gibson, T. J. (2003). ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Research*, 31(13):3625–30. 10

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–427. 60, 87

Radivojac, P., Baenziger, P. H., Kann, M. G., Mort, M. E., Hahn, M. W., and Mooney, S. D. (2008). Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, 24(16):i241–i247. 36

Ramakrishnan, N., Tadepalli, S., Watson, L. T., Helm, R. F., Antoniotti, M., and Mishra, B. (2010). Reverse engineering dynamic temporal models of biological processes and their relationships. *Proceedings of the National Academy of Sciences*, 107(28):12511–12516. 2

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics : TIG*, 13(4). 73, 75, 116, 152

Rifkin, S. A., Houle, D., Kim, J., and White, K. P. (2005). A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature*, 438(7065):220–223. 4

Rosenberg, D. and Grafton, A. (2010). *Cartographies of time: a history of the timeline*. Princeton Architectural Press. 12

Saez-Rodriguez, J., Kremling, A., and Gilles, E. (2005). Dissecting the puzzle of life: modularization of signal transduction networks. *Computers & Chemical Engineering*, 29(3):619 – 629. 1

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(suppl 1):D449–D451. 24

Sangurdekar, D., Srienc, F., and Khodursky, A. (2006). A classification based framework for quantitative description of large-scale microarray data. *Genome Biology*, 7. 87, 88

Sauer, F., Rivera-Pomar, R., Hoch, M., and Jackle, H. (1996). Gene regulation in the Drosophila embryo. *Philosophical Transactions: Biological Sciences*, 351(1339):pp. 579–587. 4

Sawyers, C. L. (2008). The cancer biomarker problem. *Nature*, 452(7187):548–552. 2

Scales, T., Derkinderen, P., Leung, K.-Y., Byers, H., Ward, M., Price, C., Bird, I., Perera, T., Kellie, S., Williamson, R., Anderton, B., and Reynolds, C. H. (2011). Tyrosine phosphorylation of Tau by the Src family kinases Lck and Fyn. *Molecular Neurodegeneration*, 6(1):12. 34

Schneider, M. and Orchard, S. (2011). Omics technologies, data and bioinformatics principles. In *Bioinformatics for Omics Data*, volume 719 of *Methods in Molecular Biology*, pages 3–30. Humana Press. 3

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):337–42. 73

Schwartz, W. J., Carpino, A., J., de la Iglesia, H. O., Baler, R., Klein, D. C., Nakabeppu, Y., and Aronin, N. (2000). Differential regulation of fos family genes in the ventrolateral and dorsomedial subdivisions of the rat suprachiasmatic nucleus. *Neuroscience*, 98(3):535–47. 129

Secrier, M., Pavlopoulos, G. A., Aerts, J., and Schneider, R. (2012). Arena3D: visualizing time-driven phenotypic differences in biological systems. *BMC Bioinformatics*, 13:45. 54, 63, 70, 72, 76, 79

Secrier, M. and Schneider, R. (2013). Visualizing time-related data in biology, a review. *Briefings in Bioinformatics*, (in press). 5, 14, 16

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Researchearch*, 13(11):2498–2504. 16, 26, 67, 110

Shaw, P. H. (1996). The role of p53 in cell cycle regulation. *Pathology - Research and Practice*, 192(7):669 – 675. 6

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311. 26

Shi, Y., Seto, E., Chang, L.-S., and Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein. *Cell*, 67(2):377 – 388. 73

Silvia D. M. Santos, J. E. F. (2008). Systems biology: on the cell cycle and its switches. *Nature*, (7202):288289. 19

So, L. H., Ghosh, A., Zong, C. H., Sepulveda, L. A., Segev, R., and Golding, I. (2011). General properties of transcriptional time series in Escherichia coli. *Nature Genetics*, 43(6):554–U84. 4

Stankiewicz, P. and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455. 3

Stein, A., Pache, R. A., Bernad, P., Pons, M., and Aloy, P. (2009). Dynamic interactions of proteins in complex networks: a more structured view. *FEBS Journal*, 276(19):5390–5405. 4

Strass, W. (1849). *Stream of time, or chart of universal history.* C. Smith. 12

Strogatz, S. H. (1995). Nonlinear dynamics: ordering chaos with disorder. *Nature*, 378. 13

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., and von Mering, C. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database issue):D561–8. 96, 98

Tan, C., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M., Jrgensen, C., Bader, G., Aebersold, R., Pawson, T., and Linding, R. (2009). Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Science Signaling*, 2(81):39. 85

Tavanti, M. and Lind, M. (2001). 2D vs 3D, implications on spatial memory. In *INFOVIS '01: Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, pages 139+, Washington, DC, USA. IEEE Computer Society. 84

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073. 26

The UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 40(D1):D71–D75. 99, 156, 158

Theocharidis, A., van Dongen, S., Enright, A. J., and Freeman, T. C. (2009). Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nature Protocols*, 4(10):1535–1550. 16, 86

Tominski, C. and Schumann, H. (2008). Visualization of gene combinations. In *Information Visualisation, 2008. 12th International Conference*, pages 120–126. IEEE. 135

Tory, M., Kirkpatrick, A., Atkins, M., and Möller, T. (2006). Visualization task performance with 2D, 3D, and combination displays. *IEEE Trans Vis Comput Graph*, 12(1):2–13. 84, 135

Toussaint, M. and von Seelen, W. (2007). Complex adaptation and system structure. *Biosystems*, 90(3):769 – 782. 2

Tufte, E. (2001). *The visual display of quantitative information*. Graphics Press, 2nd edition. 103

Tyson, J. J. and Novak, B. (2008). Temporal organization of the cell cycle. *Current Biology*, 18(17):R759 – R768. 6

Vacic, V., Markwick, P. R. L., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., and Iakoucheva, L. M. (2012). Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Computational Biology*, 8(10):e1002709. 41

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799. 7

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–61. 24, 96, 98

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue):D433–7. 24

Wagner, A. (2012). The role of robustness in phenotypic adaptation and innovation. *Proceedings of the Royal Society: Biological Sciences*. 2

Warsow, G., Greber, B., Falk, S., Harder, C., Siatkowski, M., Schordan, S., Som, A., Endlich, N., Scholer, H., Repsilber, D., Endlich, K., and Fuellen, G. (2010). ExprEssence - revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Systems Biology*, 4(1):164. 16

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191. 16

Weil, D., El-Amraoui, A., Masmoudi, S., Mustapha, M., Kikkawa, Y., Lain, S., Delmaghani, S., Adato, A., Nadifi, S., Zina, Z. B., Hamel, C., Gal, A., Ayadi, H., Yonekawa, H., and Petit, C. (2003). Usher syndrome type I G (USH1G) is caused by mutations in the gene encoding SANS, a protein that associates with the USH1C protein, harmonin. *Human Molecular Genetics*, 12(5):463–471. 41

Westenberg, M. A., Roerdink, J. B., Kuipers, O. P., and van Hijum, S. A. (2010). SpotXplore: a Cytoscape plugin for visual exploration of hotspot expression in gene regulatory networks. *Bioinformatics*, 26(22):2922–3. 16, 86, 136

Wilson, E. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212. 59

Wilson, E. (1987). *The cell in development and heredity*. Genes, cells, and organisms. Garland Publishing. 5

Wright, P. E. and Dyson, H. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321 – 331. 10

Wu, S., Shi, Y., Mulligan, P., Gay, F., Landry, J., Liu, H., Lu, J., Qi, H., Wang, W., Nickoloff, J., Wu, C., and Shi, Y. (2007). A YY1-INO80 complex regulates genomic stability through homologous recombination-based repair. *Nature Structural & Molecular Biology*, 14(12):1165–72. 73

Xuan, J., Yu, Y., Qing, T., Guo, L., and Shi, L. (2012). Next-generation sequencing in the clinic: promises and challenges. *Cancer Letters*, (0):–. 3

You, H., Jang, Y., You-Ten, A. I., Okada, H., Liepa, J., Wakeham, A., Zaugg, K., and Mak, T. W. (2004). p53-dependent inhibition of FKHRL1 in response to DNA damage through protein kinase SGK1. *Proceedings of the National Academy of Sciences*, 101(39):14057–62. 129

Zhivotovsky, B. and Orrenius, S. (2010). Cell cycle and cell death in disease: past, present and future. *Journal of Internal Medicine*, 268(5):395–409. 6

Zhou, R. and Snyder, P. M. (2005). Nedd4-2 phosphorylation induces serum and glucocorticoid-regulated kinase (SGK) ubiquitination and degradation. *Journal of Biological Chemistry*, 280(6):4518–4523. 32