Integration and analysis of phenotypic data from functional screens

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt

der Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden

von

Dipl.-Inf. Maciej Paszkowski-Rogacz

geboren am 20.01.1981 in Łódź, Polen

Eingereicht am 7.05.2010

Gutachter: Prof. Francis Stewart (BIOTEC, TU Dresden)
Prof. Michael Boutros (DKFZ, Uni Heidelberg)

Tag der Verteidigung: 29.11.2010

Die Dissertation wurde in der Zeit von 04/2006 bis 04/2010
im Max Planck Institute of Molecular Cell Biology and Genetics angefertigt.

to my parents

Acknowledgements

I would like to acknowledge my mentors, Frank Buchholz and Mayte Pisabarro, for giving me the opportunity to carry out my reasearch on the most exciting biological data and for supporting and guiding me through meanders of the projects. I am grateful to all current and past members of the Buchholz lab, who helped me to understand wonders of cell biology and who patiently coached me during my first experiments. I am also thankful to all current and past members of the Structural Bioinformatics lab for their professional and friendly support. I am indebted to Mandy Erlitz, who helped me settle down in Dresden, and to Ralf Gey, who was organizing and maintaining the top quality IT infrastructure that was so indispensable for my studies. Special thanks to Adriana Kłyszejko for introducing me the field of bioinformatics and encouraging to come to Dresden. Last, but not least, I would like to thank Anna Grzyb, without whose faith and support this work would not be possible.

Summary

Background: The functional annotation and identification of genes involved in various biological processes is a cumbersome and non trivial task that often requires investigating and understanding interactions and collective behaviour of hundreds of cellular components. Thanks to technological advances of recent years, it is now possible to simultaneously study complete genomes, transcriptomes and proteomes, or to image hundreds of organelles at a time, giving researchers an opportunities of making rapid discoveries in a highly parallel way.

Motivation: Although various high-throughput technologies provide a lot of valuable information, each of them is giving an insight into different aspects of cellular activity and each has its own limitations. Thus, a complete and systematic understanding of the cellular machinery can be achieved only by a combined analysis of results coming from different approaches. However, methods and tools for integration and analysis of heterogenous biological data still have to be developed.

Results: This work presents systemic analysis of basic cellular processes, *i.e.* cell viability and cell cycle, as well as embryonic stem cell pluripotency and differentiation. These phenomena were studied using several high-throughput technologies, whose combined results were analysed with existing and novel clustering and hit selection algorithms.

This thesis also introduces two novel data management and data analysis tools. The first, called DSViewer, is a database application designed for integrating and querying results coming from various genome-wide experiments. The second, named PhenoFam, is an application performing gene set enrichment analysis by employing structural and functional information on families of protein domains as annotation terms. Both programs are accessible through a web interface.

Conclusions: Eventually, investigations presented in this work provide the research community with novel and markedly improved repertoire of computational tools and methods that facilitate the systematic analysis of accumulated information obtained from high-throughput studies into novel biological insights.

Keywords: data integration, data analysis, bioinformatics, systems biology, high-throughput screening, RNA interference

Contents

Publications viii

List of figures ix

List of tables xi

Nomenclature xii

1 Introduction 1

1.1 High-throughput technologies 2

1.1.1 Gene expression profiling 2

1.1.2 Dissecting protein-DNA binding 3

1.1.3 RNA interference screenings 3

1.1.4 Meta-analysis 7

1.2 Data analysis methods 8

1.2.1 Databases 8

1.2.2 Analysis of systematic RNAi screens 10

1.2.3 Gene expression data analysis 16

1.2.4 ChIP-Chip Tiling Array Analysis 18

1.2.5 Enrichment analysis 19

1.2.6 Cluster analysis 22

1.3 Objectives 28

2 DSViewer–a database of systematic RNAi screens 29

2.1 Background 29

2.2 Implementation 30

2.2.1 Data management 30
2.2.2 User interface 32
2.3 Current content 36
2.4 Conclusions 36

3 Comparative profiling identifies C13orf3 as a component of the Ska complex required for mammalian cell division 37
3.1 Introduction 38
3.2 Comparative profiling of phenotypes from systematic RNAi screens . . 40
3.2.1 Data combination and normalization 41
3.2.2 Clustering of phenotypes 43
3.2.3 Cluster selection 43
3.2.4 Addition of cell viability data refines the clustering results . . . 45
3.3 Experimental results and discussion 46
3.3.1 Combining phenotypic profiling with protein localisation . . . 46
3.3.2 C13orf3 localises to prominent structures during mitosis 47
3.3.3 C13orf3 is required for anaphase onset 47
3.3.4 C13orf3 is differentially phosphorylated during mitosis 54
3.3.5 C13orf3 forms a complex with Ska1, Ska2 and PPP2R2B 55
3.4 Methods 59
3.4.1 Genome-wide cell-viability RNAi screen 59
3.4.2 BAC TransgeneOmics 59
3.4.3 Immunofluorescence microscopy 60
3.4.4 Live-cell imaging 60
3.4.5 Cell-based assays 61
3.4.6 Quantitative PCR and apoptosis assays 61
3.4.7 Western blot analysis 62
3.4.8 Immunoprecipitation and mass spectrometry 63
3.4.9 Quantification of phosphorylation 63
3.4.10 Yeast two-hybrid analysis 63
3.4.11 Sequence analysis and comparative modelling 64

4 A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity 65
4.1 Introduction 66
4.2 Analysis of the RNAi screen for Oct4 modulators 67
4.2.1 Primary screen 67
4.2.2 The screen identified known and discovered novel candidate pluripotency genes 70
4.2.3 Meta-analysis of phenotypic and gene expression data refines screening results 70
4.3 Experimental results 73
4.3.1 Self-renewal assays 73
4.3.2 The Paf1C is essential for ESC identity 75
4.3.3 Paf1C affects the expression of pluripotency and lineage-marker genes 76
4.3.4 Paf1C binds to promoters of pluripotency genes 79
4.3.5 Paf1C is required to maintain the chromatin structure of pluripotency genes in ESCs 80
4.3.6 Proteomic analyses of Paf1C in ESCs 83
4.3.7 Paf1C is downregulated during embryoid body formation . . . 84
4.3.8 Paf1C overexpression blocks ESC differentiation 86
4.4 Discussion 87
4.5 Methods 90
4.5.1 Mouse esiRNA Library 90
4.5.2 Cell Culture and High-Throughput Screen 91
4.5.3 Alkaline Phosphatase Staining 91
4.5.4 Western Blotting 91
4.5.5 RT-qPCR 92
4.5.6 Gene Expression Analyses 92
4.5.7 Cell-Cycle Profiling 92
4.5.8 Cross-Species RNAi Rescue 92
4.5.9 Chromatin Immunoprecipitation and Protein Identification by Mass Spectrometry 93
4.5.10 Overexpression Studies in Sox1-GFP ESCs 93

5 PhenoFam **95**
5.1 Background . . . 96
5.2 Implementation . . . 97
5.2.1 Data management . . . 97
5.2.2 Identifiers association . . . 97
5.2.3 Gene set enrichment analysis . . . 99
5.2.4 User interface . . . 100
5.3 Results . . . 100

6 Discussion **105**
6.1 High-throughput data integration . . . 105
6.2 Novel hit selection algorithms . . . 106
6.3 Enrichment analysis . . . 107
6.4 Outlook . . . 109
6.5 Conclusions . . . 110

References **111**

Publications

This thesis was based on the following publications:

Theis, M., Slabicki, M., Junqueira, M., **Paszkowski-Rogacz, M.**, Sontheimer, J., Kittler, R., Heninger, A.K., Glatter, T., Kruusmaa, K., Poser, I., Hyman, A.A., Pisabarro, M.T., Gstaiger, M., Aebersold, R., Shevchenko, A. & Buchholz, F. (2009). Comparative profiling identifies C13orf3 as a component of the Ska complex required for mammalian cell division. *The EMBO journal*, **28**, 1453–65.

Theis, M., **Paszkowski-Rogacz, M.** & Buchholz, F. (2009). Skanking with Ska3. *Cell Cycle*, **8**, 3435–3437.

Ding, L., **Paszkowski-Rogacz, M.**, Nitzsche, A., Slabicki, M.M., Heninger, A.K., de Vries, I., Kittler, R., Junqueira, M., Shevchenko, A., Schulz, H., Hubner, N., Doss, M.X., Sachinidis, A., Hescheler, J., Iacone, R., Anastassiadis, K., Stewart, A.F., Pisabarro, M.T., Caldarelli, A., Poser, I., Theis, M. & Buchholz, F. (2009). A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell stem cell*, **4**, 403–15.

Paszkowski-Rogacz, M., Slabicki, M.M., Pisabarro, M.T. & Buchholz, F. (2010) PhenoFam–gene set enrichment analysis through protein structural information. *BMC Bioinformatics*, accepted (in press).

List of Figures

1.1 Identification of protein-DNA interaction regions using ChIP-Chip . . 4
1.2 Mechanism of RNA interference 5
1.3 False positives and false negatives in threshold-based hits selection. . . 14
1.4 Comparison of hit selection thresholds based on z-scores and m-scores. 15
1.5 Hierarchical clustering dendrogram 23
1.6 Hierarchical clustering linkage criteria. 25

2.1 DSViewer database scheme. 31
2.2 DSViewer main page. 33
2.3 DSViewer properties selection panel. 34
2.4 DSViewer results panel. 35

3.1 Distribution of cell-cycle phenotypes. 41
3.2 Strategy of phenotypic profiling. 44
3.3 Enrichment of ontology terms after introducing cell viability data. . . . 45
3.4 Localisation of selected BAC transgenic cell lines. 46
3.5 Localisation pattern of C13orf3 in HeLa cells 48
3.6 RNAi phenotypes upon C13orf3 depletion 51
3.7 C13orf3 is phosphorylated during mitosis 56
3.8 Identification of minimal binding motifs and model of interaction . . . 58

4.1 Genome-Scale esiRNA Screen 68
4.2 Combined analysis of RNAi and time-course microarray data 71
4.3 Functional Analysis of Validated Hits 74
4.4 Paf1C Affects the Expression of Pluripotency and Lineage-Marker Genes 77

4.5 Ctr9 Binds to Promoters and Coding Regions of Pluripotency Genes and Is Required to Maintain the Chromatin Structure in ESCs 81
4.6 Paf1C Is Downregulated during Embryoid Body Formation, and Overexpression Blocks ESC Differentiation. 85

5.1 PhenoFam database scheme. 98
5.2 Identifier mapping procedure. 98
5.3 PhenoFam web interface. 101
5.4 Plexins are enriched for polyploidy RNAi phenotypes. 102

List of Tables

2.1 Systematic genome-wide RNAi screens accessible through DSViewer interface 36

3.1 Skewness and kurtosis of the cell-cycle phenotypes 42
3.2 Results of pulldown assays and mass spectrometry 57

4.1 Genes nominated as ‘hits’ by the combined analysis but not detected by the primary RNAi screen. 73
4.2 Proteomic Analyses of the Paf1C in ESCs 84

5.1 Normalized values of polyploidy RNAi phenotypes of plexins, from the primary cell cycle progression screen 103

Nomenclature

Roman Symbols

c	Scale factor
m	m-score (median-based z-score)
P	Probability density function (pdf)
Q_i	i-th quartile of the population
s_i	Silhouette of an observation i
x	Raw score
$\tilde{x}$	Median of a variable x
Z	Z-factor
z	Standard score (z-score)

Greek Symbols

α	Statistical significance level
μ	Mean
Φ^{-1}	Quantile function of a normal distribution
σ	Standard deviation
Θ	Probability function parameter set

Acronyms

AJAX	Asynchronous JavaScript and XML
ANOVA	Analysis of variance
APC	Anaphase-promoting complex
BAC	Bacterial artificial chromosome
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
dsRNA	Double-stranded RNA
ESC	Embryonic stem cell
esiRNA	Endoribonuclease-prepared siRNA
GO	Gene ontology
GSEA	Gene set enrichment analysis
HT	High-throughput
HTS	High-throughput screening
IgG	Immunoglobulin G
IP	Immunoprecipitation
IQR	Inerquartile range
J2EE	Java 2 platform, enterprise edition
JDBC	Java database connectivity
LAP	Localization and purification
MAD	Median absolute deviation
MGSA	Model-based gene set analysis

mRNA	Messenger RNA
PCR	Polymerase chain reaction
RISC	RNA-induced silencing complex
RNAi	RNA interference
SAC	Spindle assembly checkpoint
MEA	Modular enrichment analysis
SEA	Singular enrichment analysis
SILAC	Stable isotope labeling with amino acids in cell culture
siRNA	Short/Small interfering RNA
SOAP	Simple object access protocol
SQL	Structured query language
TF	Transcription factor
TM	Trimean of the population
TSS	Transcription start site
UPGMA	Unweighted pair group method with arithmetic mean
UTR	Untranslated region
UPGMC	Unweighted pair group method using the centroid average
WPGMA	Weighted pair group method with arithmetic mean
WPGMC	Weighted pair group method using the centroid average
XML	Extensible markup language

Chapter 1

Introduction

Cells are the basic building blocks of all living organisms. They are very complex and dynamic systems. To understand the interactions between cellular elements (molecules and organelles), biological research employs two principal strategies. The first, reductionist approach focuses on selected components of the cell at a time, studied in a greater detail. While molecular biology has proved fundamental to our understanding of basic laws governing interactions between single cellular components, it does not decode their collective behaviour. Answering this need, biology adopts the second, holistic approach. Here, multiple cellular components are studied systematically, leading to a quantitative and integrated description of biological processes.

Systems biology has begun to flourish thanks to technological advances of recent years. It is now possible to simultaneously study complete genomes, transcriptomes and proteomes, or to image hundreds of organelles. Performing massive amounts of experiments, known under the term high-throughput biology, gives researchers opportunities of making rapid discoveries in a highly parallel way. However, with the increase of experimental throughput, the amount of generated data rises as well and brings new challenges associated with data handling and computational analysis.

In this thesis, I present systemic analysis of basic cellular processes, *i.e.* cell viability and cell cycle, as well as embryonic stem cell pluripotency and differentiation. These phenomena were studied using several high-throughput technologies, namely gene expression profiling, chromatin immunoprecipitation and RNA interference screening, which are described in the first part of the *Introduction*. In the second part, I review current methodology that I applied, and further developed, to analyse the experimental results.

1.1 High-throughput technologies

1.1.1 Gene expression profiling

Gene expression profile

The collection of genes that are actively transcribed, called a 'transcriptome', is a major determinant of a cellular state. Differences in gene expression are responsible for morphological and physiological differences between various cell types of a multicellular organism. Also, alterations in gene expression underlie responses of cells to environmental stimuli or regulation of its temporal transitions, such as progression through the cell cycle. Hence, knowing expression pattern of a gene often provides a strong clue as to its biological role. Time-course measurements of gene expression can also give an insight into dynamics of cellular processes such as differentiation (Lu *et al.*, 2009), or even whole organism development (Arbeitman *et al.*, 2002).

On a lower scale relative levels of gene expression, or expression profiles, can be assayed by quantitative PCR that measures relative abundance of transcribed mRNAs. However, to be able to identify all genes whose transcription is influenced by a stimulus, high-throughput DNA microarray technology was developed (DeRisi *et al.*, 1997; Schena *et al.*, 1995).

Microarray technology

Microarray techniques have become one of the most widely used functional genomics tools (reviewed by Hegde *et al.*, 2000; Lockhart & Winzeler, 2000; Young, 2000). A variety of microarray platforms exist that have been developed to systematically measure gene expression. The basic idea of this technology is simple: a glass or a silicon slide (known as gene chip) is spotted or 'arrayed' with DNA fragments or oligonucleotides that represent specific gene coding regions. Purified RNA from cells is then reversely transcribed to cDNA, fluorescently labeled and hybridized to the slide. Each cDNA strand should preferably bind to a complementary spot on the array. In some cases, hybridization is done simultaneously with reference RNA (labeled with a different dye) to facilitate comparison of data across multiple experiments. Measured intensity of dye fluorescence at each spot reflects the amount of respective mRNA in the input sample (Slonim & Yanai, 2009).

1.1.2 Dissecting protein-DNA binding

Gene expression microarrays allow us to measure the outcome of the transcriptional regulation, which, in turn, is achieved by interactions between nuclear proteins and genomic DNA. Chromatin immunoprecipitation (ChIP) is widely used for determining the location of DNA binding sites in the genome for a protein of interest (reviewed by Collas, 2010). Genome-wide, ChIP can be combined with microarray technology, being referred to as ChIP-on-Chip or simply ChIP-Chip (Birney *et al.*, 2007; Blais & Dynlacht, 2005; Carey *et al.*, 2009; Horak *et al.*, 2002; Ren *et al.*, 2000). The principle underlying this technique is described in Figure 1.1. Importantly, the identification of a genomic fragment that the investigated protein was bound to does not necessarily imply direct regulatory functions of that protein on the expression of ‘neighbouring’ genes. A deeper insight into transcriptional regulation can only be obtained by integration of ChIP data with gene expression profiling.

1.1.3 RNA interference screenings

One possible way of investigating functions of genes is by analyzing effects caused by their inactivation. A technology that allows rapid investigation of phenotypes caused by gene silencing is RNA interference (RNAi) screening, also called loss-of-function screening.

RNA interference

RNAi, first discovered as an ancient anti-viral response (Fire *et al.*, 1998), is now recognized as a conserved biological mechanism of inhibiting gene expression at a post-transcriptional level. It is triggered by short double-stranded RNA (dsRNA) from endogenous or exogenous origin (the mechanism is described in Figure 1.2). Since it was

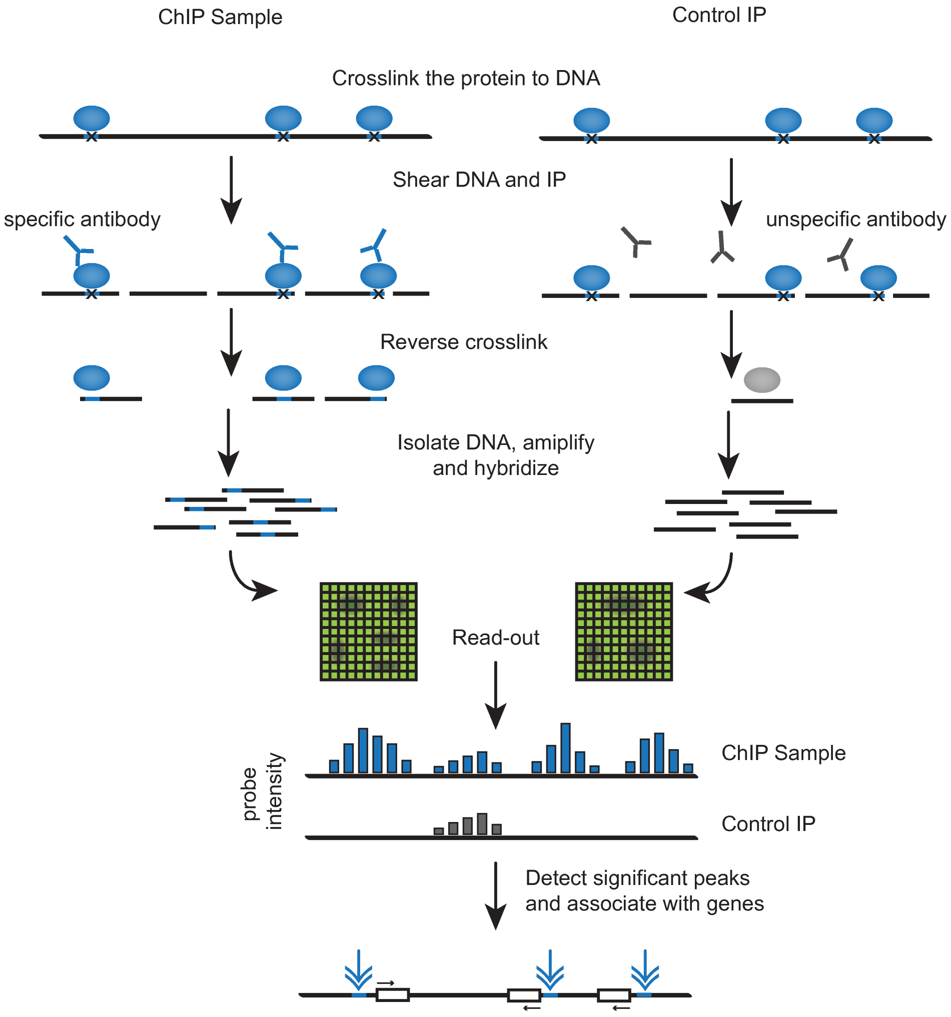


Figure 1.1: Identification of protein-DNA interaction regions using ChIP-Chip

DNA-binding proteins (including transcription factors and histones) in living cells are cross-linked to the DNA which they are bound to. Following crosslinking, the cells are lysed and the DNA is broken into pieces of approximately 0.2 kb. Using an antibody that is specific to a given DNA binding protein, one can immunoprecipitate the protein-DNA complex out of the cellular lysates. Afterwards, the cross-linking is reversed, allowing the DNA to be separated from proteins. As the control condition, an unspecific antibody can be used for chromatin precipitation.
The identity and quantity of the isolated DNA fragments can then be determined by specially designed microarrays, called tiling arrays (Nègre *et al.*, 2006; Yoder & Enkemann, 2009). Classical DNA microarrays contain up to a few probes per transcript or per gene, whereas tiling arrays cover the complete genome or its selected parts (*e.g.* promoter regions) with resolution of approximately 35 bp (Mockler *et al.*, 2005).

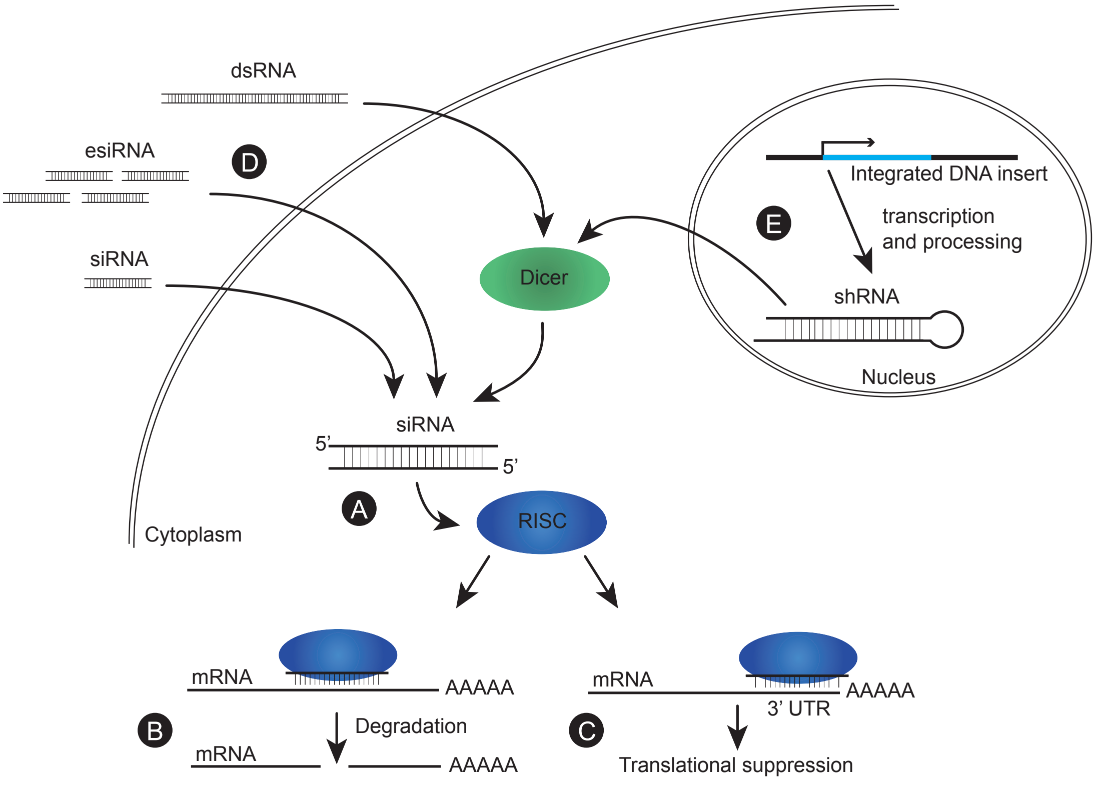


Figure 1.2: Mechanism of RNA interference

The key components of RNAi are small (≈20 bp long) double-stranded small interfering RNA molecules (siRNA). One of the two strands (a 'guide strand') is incorporated into the RNA-induced silencing complex (RISC) (**A**). The RISC complex with a bound siRNA strand recognizes complementary messenger RNA (mRNA) molecules and degrades them, resulting in substantially decreased levels of protein translation and effectively turning off the gene (**B**). If the homology between siRNA and mRNA fragment is not perfect, the mRNA is not degraded. However, the RISC complex remains bound to mRNA and inhibits the mRNA translation (**C**).

Transient effect of RNAi can be obtained by delivering siRNA into cells in various forms. Directly, by chemically synthetised siRNA or endoribonuclease-prepared pools of siRNA (esiRNA) that target common transcript (Buchholz *et al.*, 2006; Yang *et al.*, 2002a) (**D**). In invertebrates, which are lacking interferon reponse mechanism, it is possible to deliver long double-stranded RNA (dsRNA) molecules that are then digested by Dicer into a pool of siRNA (Chen *et al.*, 2007). Permanent and inheritable silencing can be achieved by integrating a fragment of DNA encoding for short hairpin RNA (shRNA) into the genome. The expressed RNA fragment forms a hairpin structure and after being exporting into cytoplasm and processed by Dicer, it forms a double-stranded siRNA (Paddison *et al.*, 2002) (**E**).

discovered that RNAi can be applied selectively, in experimental conditions, it became a powerful tool for silencing gene expression in eukaryotes (reviewed in He *et al.*, 2009).

Systematic screening

The high-throughput application of RNAi allows systematical searches for genes whose silencing leads to a specific phenotype, such as impaired development or reduced viability.

Such studies can be carried out by employing two different paradigms. The first is called a selection-based screening in which cells are infected with a pooled library of shRNA (Figure 1.2) where optimally each cell should integrate a single copy of shRNA. Further selection steps enrich for cells with a desired phenotype and by sequencing of the DNA fragments with the integrated shRNA, it is possible to identify genes whose silencing results in the phenotype. The second paradigm is termed systematic screening. Such screens are typically carried out in arrayed formats using microwell plates with 96 or 384 wells, where cells in each well are transfected with RNAi triggers targeting individual genes. Compared to selection-based, systematic screening offers the possibility of observing phenotypes of all knockdowns individually.

Depending on the assay, phenotypes are quantified by using different technologies (*e.g.* flow or laser-scanning cytometry, fluorimetry, microscopy). The typical RNAi high-throughput screening project starts with a genome-wide primary screen and the effective RNAi silencers, which are called ‘hits’, are investigated using further validation procedures, for example secondary screens.

Depending on the number of variables that compose a phenotype, systematic screens can be divided into two subcategories. If a phenotype carries a single or small number of values, for example quantified information about cell viability, growth rate or measured fluorescence intensity of a reporter protein, then such screens are termed low-content or low-dimensional. If a screen measures morphological features of cells and subcellular compartments, a number of values per phenotype can easily reach 100. RNAi screenings resulting in such high-dimensional phenotypes are called high-content screens (reviewed by Zanella *et al.*, 2010).

Regardless of the class of a systematic screen, the main challenge of the further data analysis is the optimal classification of phenotypes, which should result in a reliable identification of hits for follow-up studies. Due to technical and biological variations, phenotypes are subject to strong noise and to identify genes involved in a process of interest, data analysis techniques must cope with high rates of false positive and false negative results. Another intrinsic problem of a systematic high-throughput screen is connected with a structure of protein-protein interaction networks and cellular regulatory mechanisms. Due to redundancy of functions or presence of regulatory feedback loops, silencing of a gene that originally takes part in a process of interest may not lead to a significant change of a phenotype. And in consequence, such a gene may not be identified as a hit.

1.1.4 Meta-analysis

As explained above, each high-throughput methodology is providing an insight into different aspects of cellular activity. Since each of the single experimental techniques has its limitations, a complete and systematic overview can be achieved only by a combined analysis of results coming from different approaches.

By definition, meta-analysis is the statistical synthesis of the results of several studies that address a shared research hypotheses (O'Rourke, 2007). It is usually applied to increase statistical power to detect an effect and to reduce any bias or noise in the underlying data sets.

Previous studies have shown that meta-analysis of various gene expression microarray datasets helps in forming biological hypothesis and leads to novel discoveries (Grützmann *et al.*, 2005; Ramasamy *et al.*, 2008; Rhodes *et al.*, 2004; Silva *et al.*, 2007). Similar principles of combined analysis were also sucessfully applied to large-scale gene expression and ChIP data (Foltz *et al.*, 2009; Wu *et al.*, 2008).

Applications of meta-analysis for numerous RNAi screening data combined with protein localization, gene expression and chromatin immunoprecipitation are described in the following chapters of this thesis.

1.2 Data analysis methods

1.2.1 Databases

Recent years have seen an explosion in the amount of available biological data. More and more genomes are being sequenced and annotated, information about transcriptomes and proteomes are accumulating. Biological databases are indispensable and invaluable tools for managing these data and for making them accessible (Cochrane & Galperin, 2010). Modern database systems provide efficient mechanisms for storing, managing and querying large amount data, making them exceptionally useful for the high-throughput analysis. Below, I mention some biological databases relevant to my work.

Ensembl

Ensembl (Hubbard *et al.*, 2009) is a joint project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute to develop a software system that produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl automatically annotates genes and predict new ones, by integrating data from other biological data sources (*i.e.* InterPro, OMIM, SAGE). Currently, Ensembl contains sequences and annotations for over 50 metazoan genomes as well as genomes of plants, fungi and procaryotes.

Ensembl is a comprehensive and easily accessible source of data for any high-throughput approaches. It is available not only as an interactive website but also via Perl and Java programming interfaces, allowing simple scripts to be written to retrieve data of interest. In addition, all data is provided without any restriction in a form of an SQL database.

InterPro

InterPro (Hunter *et al.*, 2009) is a EBI-developed integrated database of predictive protein 'signatures' used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites.

InterPro constitutes a repository that integrates the most well established sources of data (referred to as member databases): PROSITE, HAMAP, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D and PANTHER.

Gene Ontology

The Gene Ontology (GO) project (Ashburner *et al.*, 2000) is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members. The use of GO terms by collaborating databases facilitates uniform queries across them.

Gene Ontology vocabulary contains three separate, hierarchically organized sets of annotations referring to cellular component, biological process and molecular function of any gene product. Each set forms a tree-like structure, in which on top of the tree there are the most general terms (*e.g.* 'developmental process') and the leafs of the tree contain more detailed annotations (*e.g.* 'positive regulation of embryonic development').

Controlled vocabularies of annotations are indespensable tools in any automated analysis of high-throughput data. Together with manual curration and other gene annotation methodologies (*e.g.* text mining), they allow introduction of previously gathered knowledge into computational algorithms.

PANTHER

The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System (Mi *et al.*, 2007; Thomas *et al.*, 2003) is a resource developed by the Evolutionary Systems Biology Group at SRI. PANTHER classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are classified by expert biologists into families and subfamilies of shared function, which are then categorized by molecular function and biological process ontology terms. The PANTHER ontology is a controlled vocabulary arranged in a similar fashion to the Gene Ontology, but greatly abbreviated and simplified to facilitate high-throughput analyses.

BioMart

BioMart (Haider *et al.*, 2009; Kasprzyk *et al.*, 2004; Smedley *et al.*, 2009) is a query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI). The system can be used with any type of data and is particularly suited for providing data mining searches of complex descriptive and sequential data.

BioMart web page offers a centralized access to a set of biological databases. These include major biomolecular sequence, pathway and annotation databases such as Ensembl, Uniprot, Reactome, HGNC, Wormbase and PRIDE (Haider *et al.*, 2009). Moreover, BioMart system can also be customized to manage any user-provided data, and thanks to a built-in support for data federation, it is possible to integrate it with other databases already configured for use with BioMart.

1.2.2 Analysis of systematic RNAi screens

Quality control

Quality control is a crucial step of any systematic RNAi screen analysis. It allows to identify problematic plates that need to be repeated and helps with making a choice of further analysis methods and is a necessary prerequisite for an integrative analysis of phenotypic profiles from multiple screens (Boutros *et al.*, 2006).

Usually each plate contains few wells with negative and positive controls (explicit controls). However, assuming that the majority of samples on a plate are irrelevant to the investigated biological process, it is possible to use samples from the whole plate for estimating the distribution of negative controls (implicit controls).

A popular measure to assess the quality of individual plates, or the whole screen, is called Z-factor (Zhang *et al.*, 1999). It is an estimate of a dynamic range of the assay, which is based on the values of explicit controls. Given the means and standard deviations of negative (n) and positive (p) controls, it is calculated as follows:

$$Z = 1 - \frac{3 \cdot \left(\sigma_p + \sigma_n\right)}{\left|\mu_p - \mu_n\right|} \tag{1.1}$$

According to Zhang *et al.* (1999), the values between 0.5 and 1.0 indicate an excellent assay and values below 0.0 suggest that the overlap between positive and negative controls is too big for a reasonable detection of phenotypes. However, it must be noted that this interpretation assumes a normal distribution of controls. If this criteria is not met, then despite low quality of an assay, the Z-factor may give misleadingly favourable results (Sui & Wu, 2007). Thus, it would be advisable to replace mean and standard deviation with more robust estimators, such as median and median absolute deviation (Zhang *et al.*, 2006).

Results normalization

Normalization is a data analysis step intended to remove systematic errors and to allow combination and comparison of results of different plates and replicates of the screen. The challenge of normalization is to remove as much of the technical variation as possible while leaving the biological variation untouched.

To perform a genome-wide screen targeting approximately 20 000 human genes and an assay performed in 384-well plates, it is necessary to use over 50 plates. Moreover, assuming two replicates of the screen, the total number of plates that must be processed is over 100. Even with current advances in automation and robotics, the first and the last plate of the screen might be processed in different batches separated by hours or even days. This might introduce a strong variability between results obtained from different plates (Coma *et al.*, 2009). A solution of this problem is to treat each plate as a separate unit and normalize results on different plates separately (normalization per-plate).

However, in cases of non-random distribution of phenotypes among plates, normalization per-plate may lead to an increased number of false positives (on plates lacking true-positives) and false negatives (on plates enriched with true positives). Thus, it might be necessary to treat the whole screen as one batch, and normalize all results together (per-batch normalization).

Standard score. The most frequent way of normalizing the high-throughput assay readouts is a conversion of raw values into their standard scores widely known as z-scores (Boutros *et al.*, 2006). A standard score of a raw value x is calculated as

$$z = \frac{x - \mu}{\sigma} \tag{1.2}$$

where μ is a mean and σ is a standard deviation of all raw values taken from the same plate (in case of per-plate normalisation) or the same batch (per-batch normalisation). In case of a screen in which explicit negative controls are used, the raw values of samples can be normalized to the mean and standard deviation of the negative controls.

This procedure is also called ‘zero mean and unit variance standardization’, because it ensures that all readouts are transformed in such a way that the mean value of the normalized readout is zero, whereas the standard deviation and the variance of whole sample are equal to unity. The value of z-score represents the distance between the raw score and the population mean in units of the standard deviation. In the RNAi screens, this value shows how significantly the knock-down result differ from negative controls (explicit or implicit).

One way of making the standard score normalization more robust against outliers or strong hits, which may bias the mean and standard deviation, is to calculate z-score using mean and standard deviation estimates based on the fraction of the sample (a trimmed mean and a trimmed standard deviation) (Weisberg, 1991). The fraction is build by discarding a certain number of the lowest and the highest values from the sample, usually 5 % of each end. In case of calculating a mean of 384 samples, one would discard 20 lowest and 20 highest values.

The trimmed mean and trimmed standard deviation produce unbiased estimators only if the underlying distribution is symmetric. An alternative and more robust estimator of the population mean is trimean (Tukey, 1977). It is calculated as an weighted average of the samples’s median and its two quartiles:

$$\mathrm{TM} = \frac{Q_1 + 2Q_2 + Q_3}{4} \tag{1.3}$$

An advantage of the trimean as a measure of the center of a distribution is that it combines the median’s emphasis on center values with the midhinge’s (average of the first and third quartiles) attention to the extremes (Weisberg, 1991).

Robust m-score. This method is an improvement on the z-score normalization approach by making it even more robust against the outliers that may significantly affect values of the mean and the standard deviation. A formula for an m-score (also known as 'robust z-score') is derived by replacing mean (μ) and standard deviation (σ) in formula 1.2 by median ($\tilde{x}$) and median absolute deviation (MAD), respectively (Chung *et al.*, 2008; Zhang *et al.*, 2006).

$$m = \frac{x - c\tilde{x}}{\text{MAD}} \tag{1.4}$$

where

$$\text{MAD} = \text{median}(|x - \tilde{x}|) \tag{1.5}$$

The equation contains an additional scale factor c, which is used for ensuring that m-scores calculated for normally distributed variables (perfect case) are equal to their z-scores. This criteria is met if

$$c\text{MAD} = \sigma \tag{1.6}$$

and

$$\tilde{x} = \mu \tag{1.7}$$

In case of normally distributed data ($\mu = 0$ and $\sigma = 1$), the median is equal to the mean (equation 1.7) and the criteria 1.6 is met if the scale factor $c \approx 1.4826$. This value can be obtained by the following calculation:

$$\begin{aligned} \sigma &= c\text{MAD} && (1.8)\\ c &= \frac{\sigma}{\text{MAD}} && (1.9)\\ c &= \frac{1}{\Phi^{-1}(0.75)} && (1.10)\\ c &\approx 1.4826 && (1.11) \end{aligned}$$

where Φ^{-1} is the quantile function of a normal distribution. By substituting c in equation 1.4 by value calculated in 1.11, we obtain the final equation for calculating m-scores:

$$m = \frac{x - 1.4826\tilde{x}}{\text{MAD}} \tag{1.12}$$

In general, for RNAi screens, m-score calculation is preferred over z-scores (Chung *et al.*, 2008).

Hits selection

The goal of any primary RNAi screen is to identify 'screening positives' or 'hits'. Hits selection is essentially a process of deciding which results differ significantly from negative controls. There are various techniques ranging from the simplest selection of a predefined number of top-scoring samples to the most elaborate, knowledge-based algorithms. A list of hits forms a basis for further validation experiments.

The aim of any hit selection algorithm is to minimize the false negative rate while keeping the false positive rate possibly low (Figure 1.3). The minimization of false negative rate is prioritized because the false positives can be filtered out by subsequent validation steps, whereas, any false negatives are already lost in the primary screen analysis.

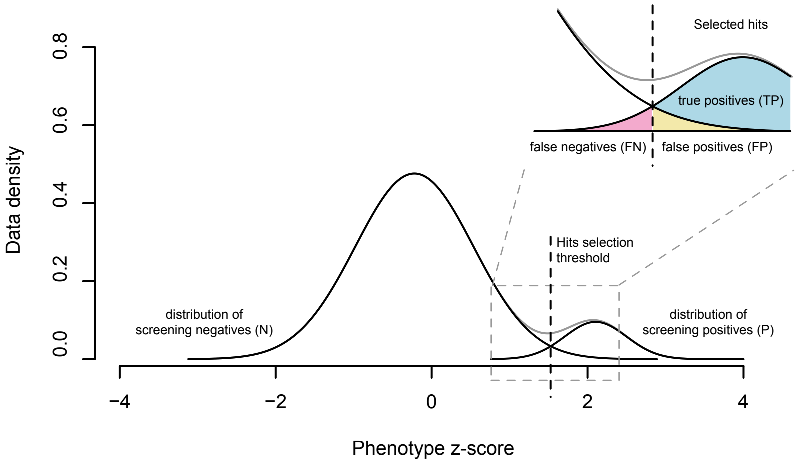


Figure 1.3: False positives and false negatives in threshold-based hits selection.

The plot contains an example of two hypothetical probability density functions representing screening negatives and screening positives (a standard normal distribution and a normal distribution with $\mu = 3$ and $\sigma = 0.5$ mixed with proportion $10 : 1$). By applying a hits selection threshold (dashed vertical line), a list of hits is generated. The list contains true positives and false positives. All screening positives that are not considered hits are false negatives. By adjusting the threshold, one can reduce the false negatives rate with a trade-off of increasing the false positives rate.

Threshold-based criteria. The most commonly used approach of hits selection involves selecting a z-score or m-score threshold and identifying positives as samples that surpass this threshold. An advantage of this method is that it is very easy to implement. Moreover, assuming a normal distribution of samples, the value of an upper or a lower cut-off can be derived from an expected statistical significance level (α), using the following formula:

$$z_{\text{cut}} = \pm\left|\Phi^{-1}(\alpha/2)\right| \tag{1.13}$$

For a commonly used significance level 0.05, a z-score or m-score threshold is approximately equal to 2. Figure 1.4 shows the difference between thresholds based on z-score and m-score normalization. Due to the fact that MAD is more robust against outliers, thresholds based on m-score normalization ensure a lower false negative rate (Fig. 1.4B).

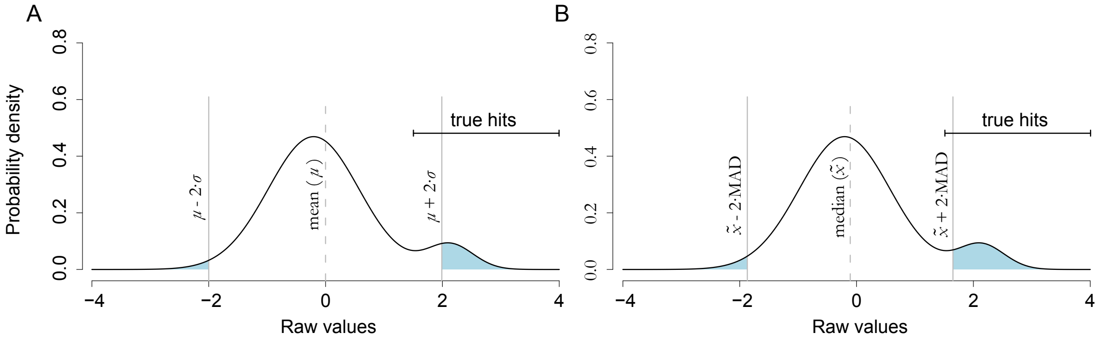


Figure 1.4: Comparison of hit selection thresholds based on z-scores and m-scores.

Both plots display the same probability density function build as a mixture of two normal distribution (Figure 1.3). Blue areas indicate values identified as hits. (A) Using the z-score normalization and threshold of 2 standard deviations. (B) Using m-score normalization and threshold of 2 MADs.

Robust methods. If a distribution of the screening results is highly asymmetrical, a simple threshold-based criteria cannot be applied. This is caused by to the fact that in such cases, the mean and the standard deviation cannot be estimated accurately. The quartile-based approach sets an upper or lower thresholds based on number of interquartile ranges (IQR) above or below the first and the third quartile of the data.

This method has been shown to outperform the threshold-based criteria. However, the improvement over the hits selection based on m-scores is rather moderate (Zhang *et al.*, 2006). Additionally, because the expected significance level can not be easily transformed into a quartile-based thresholds, the method has not been widely used.

Other techniques. Other hit selection methods rely on clustering techniques (Gagarin *et al.*, 2006) or Bayesian frameworks (Zhang *et al.*, 2008). Clustering algorithms, instead of relying on defined thresholds, identify hits based on their separation from the distribution of other values. Bayesian statistics use Bayes theorem to calculate the likelihood that a particular sample is better described by the positives-samples model *versus* the negative-samples model. A strength of this approach is that may incorporate both plate-wide and experiment-wide information as well as information from both implicit and explicit negative controls.

1.2.3 Gene expression data analysis

Background correction

Analysis of the microarray data begins with the acquisition of the fluorescence intensity values for each probe. Those probes on the array that are hybridized to a higher number of labeled fragments result in the higher intensity of the signal. The main source of variability caused by the microarray technology is introduced by nonspecific binding of oligonucleotides and optical noise (Sui *et al.*, 2009).

Various approaches have been taken to estimate and adjust for background noise in oligonucleotide arrays. The most popular algorithms include MAS 5.0 (Hubbell *et al.*, 2002), Robust Multiarray Analysis (RMA) (Irizarry *et al.*, 2003a,b) and Variance Stabilization Normalization (VSN) (Huber *et al.*, 2003).

MAS 5.0 is an algorithm designed especially for Affymetrix arrays containing additional 'mismatch probes' that differ from main probes by only one nucleotide. Intensities reported by mismatch probes are used by MAS to estimate non-specific binding and correct for it. Despite its strength, this background adjustment turned out to introduce a lot of variability in the log transformed gene expression measures (Irizarry *et al.*, 2003b).

RMA estimates the background noise by analyzing all probes across all arrays in the same hybridization batch and VSN extends the RMA methodology by introducing a simple model to predict the affinity of probes toward nonspecific binding based on the sequence composition of the probes. Presently, both methods became a standard for the background correction.

Microarray data normalization

Normalization of probe intensity values obtained from DNA microarrays is a critical step for obtaining data that are reliable and usable for subsequent analysis such as identification of differentially expressed genes and clustering. A variety of normalization methods have been proposed over the past few years (reviewed in Bolstad *et al.*, 2003; Slonim & Yanai, 2009).

Scaling. Scaling is the most straightforward normalization procedure that shifts trimmed mean or median of values (x) from each microarray (i) to a trimmed mean or median of the 'baseline' microarray j. The baseline microarray can be either one of the input microarrays or it can be calculated as an average of all microarrays from the study.

Normalized probe intensities (x') of a microarray are calculated in the following way:

$$\beta_i = \frac{\widetilde{x_j}}{\widetilde{x_i}} \tag{1.14}$$

$$x'_i = \beta_i x_i \tag{1.15}$$

Quantile normalization. The goal of the quantile method is to make the distribution of probe intensities for each array in a set of arrays the same. This result is achieved by first associating raw values from each microarray with their ranks within a microarray. Then, for each rank, an average raw value is calculated. This value is then used as a replacement for values in all microarrays associated with matching ranks. As a result, each microarray contains the same set of normalised values and thus, each array has the same distribution.

Spike-in normalization. Spike-in normalization is based on the presence of known and equally-abundant oligonucleotide (called spike-in) in all samples used for hybridization. After obtaining the raw data, results of different microarrrays are centered around the spike-in probe. This normalization method removes the error associated with shift of the dye intensity. The scale can be also corrected by introducing spike-ins of different concentrations.

Many other normalization techniques exist, including non-parametric methods (Troyanskaya *et al.*, 2002) that are especially useful for the analysis of two-color microarrays (Do & Choi, 2006). However, the analysis of two-color microarrays is not in the scope of this thesis.

1.2.4 ChIP-Chip Tiling Array Analysis

The analysis of ChIP-Chip data is a process that requires the identification of positive probes and ChIP-enriched binding regions, the mapping of those regions to the genes, and potential protein binding DNA sequence motif discovery (Yoder & Enkemann, 2009).

Peak detection

Peak detection is a proces of identifying genomic regions where ChIP-chip probes are bound by the protein of interest at levels significantly above background (control condition samples).

First methods developed to identify regions enriched by ChIP on Affymetrix tiling arrays are based on statistics that compare ChIP array data with one or more control samples. The Mann–Whitney U test is applied to ChIP-chip data by ranking of ChIP and control probe signals within 1 kb sliding windows (Cawley *et al.*, 2004).

However, the most effective algorithm for peak-finding on Affymetrix tiling microarrays is MAT (model-based analysis of tiling arrays) (Johnson *et al.*, 2006). By using a linear model, MAT estimates the baseline probe behavior based on probe sequence characteristics and genome copy number. Calculated baseline model is used then to standardize the probes and filter out the noise in the data.

Mapping of bound regions to genomic elements

Identification of genes, which potenially have their expression altered by bounding of a protein of interest, is an essential part of the ChIP-Chip analysis workflow. Association of identified peaks with genomic elements (*i.e.* transcription start sites, UTRs, introns) is performed by either automatic or manual inspection of chromosomes in the closes proximity of the peak summit.

For manual inspection, genome browsers are powerful tools that allow the visualization of ChIP-Chip experimental data as a track against an annotated genome. The

UCSC Genome Browser is a commonly used web-based genome browser that is available at `http://genome.ucsc.edu/` (Karolchik *et al.*, 2009). The proces of annotating peaks with genes can be automated by exploiting genomic databases and comparing chromosomal positions of peak summits with positions of transcription start sites (TSS) or other genomic elements of interest (Tompa *et al.*, 2005).

1.2.5 Enrichment analysis

Enrichment analysis is one of widely used bioinformatics methods for a systematic dissection of large genes lists, such as hits from systematic RNAi screens or lists of differentially expressed genes. Thanks to the biological knowledge gathered in various databases (*e.g.* Gene Ontology), the enrichment analysis makes it possible to assemble a summary of various biological annotations that could be associated with the given set of genes. Such annotations may include not only Gene Ontology terms but also pathways, transcription factor binding sites, epigenetic markers, structural properties of proteins or any other annotations derived from previous studies.

All currently available enrichment analysis tools can be classified into four categories, depending on the algorithm they use: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), modular enrichment analysis (MEA) (all three reviewed in Huang *et al.*, 2009) and model-based gene set analysis (MGSA). Some tools have implemented several algorithms so they may belong to more than one class.

Singular Enrichment Analysis (SEA)

This is the most commonly used approach in which a list of hits is iteratively tested for the enrichment of each annotation term one-by-one in a linear mode. Thereafter, the individual, enriched annotation terms passing the enrichment p-value threshold are reported in a tabular format ordered by the enrichment probability (enrichment p-value). The enrichment p-value calculation, *i.e.* number of genes in the list that are annotated with a given annotation as compared to pure random chance, can be performed with the aid of some common and well-known statistical methods, including χ^2 test, Fisher's exact test, Binomial probability and Hypergeometric distribution, *etc.*

Gene Set Enrichment Analysis (GSEA)

SEA approach strongly relies on a chosen hit selection algorithm and user-defined thresholds. Moreover, the experimental results (*i.e.* level of expression or phenotype strength) are not considered. To overcome these limitations, a gene set enrichment analysis (GSEA) method was developed (Mootha *et al.*, 2003). GSEA sorts a complete list of experimental results and searches for annotations enriched on its top or bottom. This allows even mild effects to contribute to the overall enrichment score. To calculate the significance of association, the Kolmogorov–Smirnov or the Mann–Whitney U-test are used (Keller *et al.*, 2008; Subramanian *et al.*, 2005).

However, tools in the GSEA class are also associated with some common limitations. First, the 'no-cutoff' strategy is the key advantage of GSEA, but is also becoming its major limitation in some biological studies. The GSEA method requires a summarized biological value (*e.g.* fold change) for each of the genes in the input. Despite that, in case of quantitative studies, such as cell-based RNAi screens, this limitation is not a problem. A more extensive introduction to GSEA methods is given in Chapter 5.

Modular Enrichment Analysis (MEA)

MEA inherits the basic enrichment calculation found in SEA and incorporates additional algorithms considering the relationships between annotation terms. An example of such algorithm is implemented in the ProfCom tool (Antonov & Mewes, 2006; Antonov *et al.*, 2008), which has an ability to profile enrichments of whole subgroups of available GO terms assembled in a Boolean fashion. This and other tools, such as Ontologizer (Bauer *et al.*, 2008), topGO (Alexa *et al.*, 2006), GENECODIS (Carmona-Saez *et al.*, 2007; Nogales-Cadenas *et al.*, 2009), ADGO (Nam *et al.*, 2006) claim to improve discovery sensitivity and specificity.

Model-based Gene Set Analysis (MGSA)

MGSA is a newly emerged class of methods that analyze all annotation categories at once by embedding them in a Bayesian network (Bauer *et al.*, 2010; Zhang *et al.*, 2010). Gene response is modeled as a function of the activation of biological categories. Probabilistic inference is used to identify the active categories. The Bayesian modeling approach naturally takes category overlap into account and avoids the need for multiple testing correction.

Multiple testing correction

Most enrichment analysis techniques require multiple statistical tests and thus, the probability of making at least one false discovery (*i.e.* that a given set of genes is enriched for a specific term) increases significantly. In statistics, this probability is called the family-wise error rate (FWER). In general, all multiple testing correction techniques attempt to reduce FWER while keeping the testing power at the same time. The reduction of FWER is achieved by requiring a stronger level of evidence to be observed in order for an individual enrichment to be called ‘significant’ (Lehmann & Romano, 2005).

The most commonly used multiple hypothesis correction method is the Bonferroni correction. Assuming n independent statistical tests and the given significance level for the whole family of tests to be (at most) α, each of the individual test should be tested at the level of α/n. So all individual p-values should be multipled by n before applying the significance threshold selection. It is considered to be the most conservative among all multiple testing correction techniques (Rice *et al.*, 2008).

This technique profides the maximum FWER control, but it is considered to be too restrictive for practical use in bioinformatics. An alternative to the Bonferroni correction was designed by Banjamini and Hochberg (Benjamini & Hochberg, 1995). It is the false discovery rate (FDR) control algorithm, which correct for the expected number of false discoveries (in contrast to Bonferroni method that assumes the worst-case scenario). It was shown (Benjamini & Yekutieli, 2001; Williams *et al.*, 1999) that their approach yields much greater power than the Bonferroni technique. The corrected p-value (p_{corr}) can be calculated from the following formula:

$$p_{\text{corr}} = \frac{p \cdot n}{r} \tag{1.16}$$

where n is the total number of statistical tests performed and r is a rank of the p-value (p) in a list of all obtained p-values sorted in the ascending order.

Other multiple testing correction algorithms are based on various permutation approaches (Boyle *et al.*, 2004). Boyle’s algorithm repeats the enrichment analysis on randomly picked lists of genes that are of the same size as the original list. Obtained results are used for generating *null* distributions of p-values for each annotation term. The *null* distribution can be constructed from at least 100 permutations. Finally, for a given term, a corrected p-value is calculated as a fraction of p-values from the *null* distribution that are the same or lower than the observed p-value.

Examples of application of enrichment analysis are given in following chapters, and especially in Chapter 5, which is exclusively describing a novel GSEA algorithm utilising protein structure-derived information as annotation terms.

1.2.6 Cluster analysis

Cluster analysis has become a standard computational method for gene function discovery as well as for more general explanatory data analysis. The objective of cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. In the analysis of RNAi screens, clustering is used to group genes giving similar phenotypes. In transcriptomics, cluster analysis is applied to group genes based on their expression patterns.

Clustering algorithms can be divided into three categories: hierarchical, partitional and model-based. Hierarchical algorithms find successive clusters using previously established clusters. These algorithms can be either agglomerative ('bottom-up') or divisive ('top-down'). Partitional algorithms typically determine all clusters at once. Model-based algorithms assume that the data were generated by a model and tries to recover the parameters describing it. Calculated parameters are then used to define clusters and the assignment of observations.

Regardless of the used method, the final number of clusters must be decided at some point. Algorithms helping in taking the right decision are described in the last part of this section.

Hierarchical clustering

Hierarchical clustering creates a hierarchy of clusters that may be represented in a binary tree data structure called a dendrogram. A 'root' of the tree consists of a single cluster containing all observations, and the 'leafs' correspond to individual observations. The final grouping of observations is obtained by cutting the dendrogram at a specified 'height' (se Figure 1.5).

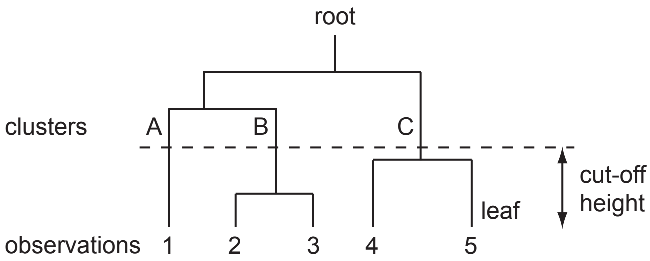


Figure 1.5: Hierarchical clustering dendrogram

Observations 1 to 5 belong to their own clusters - ‘leafs’ of the dendrogram. Observations 2 and 3 are the most similar to each and thus, they are merged at the lowest ‘height’ into one cluster (B). Sucesive merging of observations and formed clusters lead to the final single cluster called ‘root’. Application of a similarity cut-off allows a final definition of clusters. In this example, three clusters are defined: A, containing a single observation 1; B with observations 2 and 3; C with observations 4 and 5.

A dendrogram can be constructed either by iterative division of bigger clusters into two subclusters or by merging two subcluster into one bigger. In order to decide which clusters should be combined (for agglomerative clustering), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations), and a linkage criteria that specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

In the analysis of high-throughput data, the commonly used metrics include the Euclidean distance, Mahalonobis distance and other metrics based on correlation coeffitients (*i.e.* Pearson product-moment correlation, uncentered correlation).

A dissimilarity between sets of observations can be computed using different linkage criteria (for visual interpretation, see Figure 1.6):

Complete linkage (furthest neighbor). A distance between sets is defined as the maximum distance between any observation from one set, to any observation from the second set. This method usually performs quite well in cases when the samples are naturally separated into distinct groups.

Single linkage (nearest neighbor). As opposite to the previous method, single linkage takes the minimum distance between any two observations from two sets. This method is suitable for more diffused or elongated clusters.

Unweighted pair-group average (UPGMA). In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. UPGMA is a compromise between the two previous methods, performing well in cases of well defined and difused clusters of data. It is the most commonly used linkage technique.

Weighted pair-group average (WPGMA). This method is identical to the unweighted pair-group average method, except that in the computations, the size of the respective clusters (i.e., the number of observations contained in them) is used as a weight. Thus, this method (rather than the previous method) should be used when the cluster sizes are suspected to be greatly uneven.

Unweighted pair-group centroid (UPGMC). The distance between clusters is defined as the distance between their centroids, defined as an average of all objects in the cluster.

Weighted pair-group centroid (WPGMC). This method is identical to the previous one, except that weighting is introduced into the computations to take into consideration differences in cluster sizes (i.e., the number of objects contained in them). Thus, when there are (or we suspect there to be) considerable differences in cluster sizes, this method is preferable to the previous one.

Ward's method. This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step. This method is regarded as very efficient, however, it tends to create clusters of small size.

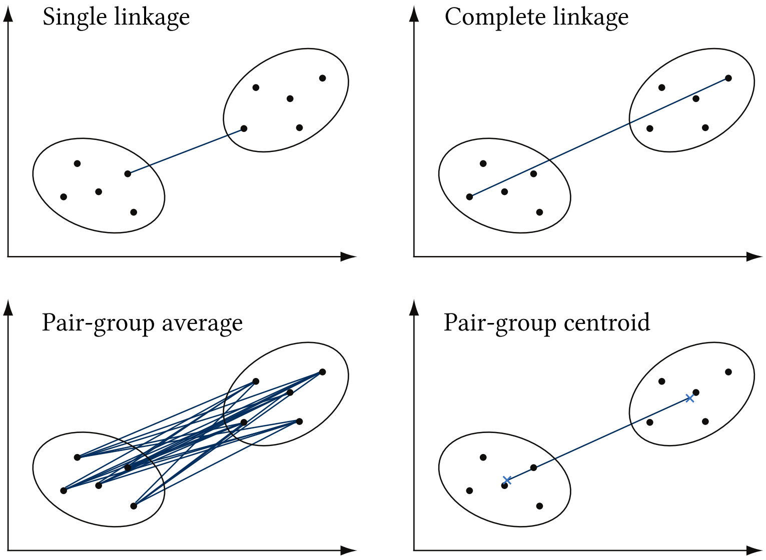


Figure 1.6: Hierarchical clustering linkage criteria.

Calculation of distances between two clusters according to different linkage criteria. Dots represent observations in two-dimensional feature space and ellipses indicate observations belonging to a single cluster. Lines show distances taken into consideration for calculating distances between clusters. Crosses in the 'Pair-group centroid' axis represent centroids of the clusters.

Partitional clustering

The most widely used partitional clustering method is a k-means algorithm (MacQueen, 1967). This algorithm divides a given set of observations $X = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, where each is a d-dimensional vector, into an *a priori* decided k sets $(k < n)$ $\{S_1, S_2, \dots, S_k\}$. The aim is to minimize the within-cluster sum of squares:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_j} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \tag{1.17}$$

where $\boldsymbol{\mu}_i$ is the mean of all observations in S_i and the $\| \bullet \|$ operator represents any distance metric (*e.g.* Euclidean distance).

A derivative of the k-means algorithm is the fuzzy c-means clustering (Bezdek, 1981). In the fuzzy clustering, each observation has a degree of belonging to clusters, rather than belonging completely to just one cluster. The fuzzy logic equations are used in the processes of finding the optimal solution. Then, the cluster space is discretized and each observation is associated with a single cluster.

Model-based clustering

In model-based methods (Villarroel *et al.*, 2009), each observation $\mathbf{x}_i$ is modelled by a finite mixture distribution with the prior probability α_j that every sample $\mathbf{x}_i$ is a member of only one mixture component j, and the conditional probability modelling each component j by the parametrized probability density function $P_j\left(\mathbf{x}_i|\mathbf{\Theta}_j\right)$ (usually the multivariate Gaussian distribution). The finite mixture model expresses the probability of observing the sample $\mathbf{x}_i$ as a sum of individual components:

$$P(\mathbf{x}_i|\mathbf{\Theta}) = \sum_{j=1}^{k} \alpha_j P_j\left(\mathbf{x}_i|\mathbf{\Theta}_j\right) \tag{1.18}$$

The aim of this clustering technique is to find such parameters $\mathbf{\Theta}_j$ and α_j, that maximize the joint probability of observing the data set $\mathbf{x}$. This goal is usually obtained by applying Expectation-Maximization (EM) algorithms. The number of mixture components is analogous to the number of clusters and the association between an observation $\mathbf{x}_i$ and a cluster is based on the final probability $P_j\left(\mathbf{x}_i|\mathbf{\Theta}_j\right)$ (which should be maximal for the cluster j).

The main advantage of all partitional and model-based clustering techniques, compared to hierarchical clustering, is that they attempt to find centers of natural clusters in the data set. However, their drawback is that an inappropriate choice of the number of clusters k may yield misleading results.

Optimizing number of clusters

The best method of determining the optimal number of clusters is to partition data into k clusters and measure the quality of this clustering. The final partitioning is found by comparing quality measures for different values of k. However, the goal of cluster analysis is not to find the best partitioning of the given sample, but to approximate the true partitions of the underlying space. In the analysis of biological data, usually, the best quality measures are obtained for very large number of clusters. This is the classical overfitting problem, which can be overcame by taking the smallest k yielding reasonably good clustering quality and for which an increase of k would not significantly imrpove the clustering. This criterion is called an 'elbow' method (Van Ryzin, 1995).

The most straightforward method of measuring how well the clusters are formed is by calculating the average silhouette of all observations (Rousseeuw, 1987). A silhouette s of an observation i is calculated with the following formula:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \tag{1.19}$$

where a_i denotes the average dissimilarity of i with all other observations belonging to the same cluster, and b_i is the average dissimilarity of i with all observations belonging to the closest cluster to i. The closets cluster is defined by having the smallest average dissimilarity with i. Values of s_i can range between -1 and 1, with -1 meaning that the observation i should rather be a member of the closest cluster and 1 meaning that the observation i fits well to its cluster. An average s_i for a single cluster tells how compact the cluster is, and the average s_i for all observations defines the whole clustering quality.

Evaluation of model-based clusters can be based on the likelihood that the obtained clustering model describes the data well. The likelihood is one of the parameter calculated by expectation-maximization (EM) clustering algorithms and reported as one of the clustering results.

1.3 Objectives

The main objective of this thesis was to develop and apply a set of methods and tools aiding in meta-analysis of phenotypic data obtained from systematic RNAi screens. The goal of this thesis is to show that combined analysis of this data with other sources of biological information (*e.g.* gene expression and chromatin immunoprecipitation data as well as existing functional and structural annotations), may bring a deeper insight into function of genes.

There are three specific aims of the systematic approaches presented in the following chapters. The first goal was to develop a tool and methodology for efficient organization and systematic integration of phenotypic data from multiple systematic RNAi screens. The software must provide a unified web-based interface to all high-throuput data generated in a laboratory. A novel application providing this functionality is described in Chapter 2.

The second goal was to establish meta-analysis algorithms for studying results of multiple genome-wide data sets and to apply them to investigations of such biological process as cell cycle, cell viability and self-renewal of embryonic stem cells. Results covering this objective of the thesis are presented in Chapter 3 and 4.

The last goal was to develop an enrichment analysis tool that facilitates analysis of results obtained from genome-wide studies, by employing structural annotation of proteins. A novel GSEA application and an example of its application are described in Chapter 5.

In summary, these investigations provide the research community with tools, methods and guidelines for analyzing various types of high-throughput data that ultimately lead to a better functional characterisation of genes.

Chapter 2

DSViewer–a database of systematic RNAi screens

With an increase of availability of large data-sets of RNAi-induced phenotypes, an important step is the organization and systematic integration of this functional information. Thus, a pressing need exists for tools simplifying storage and analysis of the large amount of phenotypic information generated. This chapter describes DSViewer, a new web-accessible and user-friendly database application designed for efficient storage, integration and querying of large-scale data sets generated by high-throughput experiments.

2.1 Background

The deluge of data generated by genome-scale RNAi screens resulted in a growing number of repositories of phenotypic data. Most of the databases tend to be species-specific, like RNAiDB (Gunsalus *et al.*, 2004), which stores phenotypic data from large-scale RNAi analyses in *C. elegans*, or *Drosophila*-specific repositories such as FlyBase (Drysdale, 2008), FLIGHT (Sims *et al.*, 2006) and FlyRNAi (Flockhart *et al.*, 2006), which contain a variety of curated data including annotated genomes, expression patterns, mutant phenotypes, genetic interactions and anatomy images. Databases providing access to data from mammalian systems include MPD (Bogue & Grubb, 2004), which contains phenotypic data on commonly used inbred mouse strains, and GenomeRNAi (Horn *et al.*, 2007), which is a database providing access to RNAi phenotypes obtained from cell-based screens in *Homo sapiens* and *Drosophila*.

Since the databases and resources containing phenotypic data tend to be species-specific, cross-species comparative phenomic analysis often remains an impossible task. Moreover, data filtering possibilities of their web-based interfaces are limited to single gene or single phenotype queries, making them difficult to be integrated into high-throughput analysis pipelines. Most of the publicly-available repositories are designed for serving results of finalised projects (*i.e.* fully analysed and published). Hence making them unavailable for integration into local laboratory information management systems.

Although the analysis of results of a single RNAi screening experiment can be done by utilizing spreadsheet applications (*e.g.* Microsoft Office Excel, OpenOffice.org Calc), more complicated tasks involving data integration and meta-analsysis require an application of a professional statistical computing software, such as R/BioConductor (Durinck *et al.*, 2009), coupled with database management systems (*e.g.* MySQL, PostgreSQL, SQLite). These solutions require extensive knowledge about programming environments and are not intuitive to use. Therefore, an application that combines user-friendliness of web-accessible repositories with flexibility of statistical software is needed.

2.2 Implementation

DSViewer is a J2EE (Java 2, Enterprise Edition) web application running on a Tomcat 5.5 server and using MySQL as a database management system. To ensure stability and scalability of the tool, the model-view-controller is implemented using the Spring framework (Johnson, 2005). In order to ensure high responsiveness of the user interface, parts of the application are scripted using Asynchronous JavaScript and XML (AJAX) technology, which allows replacing fragments of the user interface without the need of reloading the complete web page.

2.2.1 Data management

In order to be able to store results of various systematic RNAi screening assays, the database scheme must be flexible enough to allow storage of different types of phe-

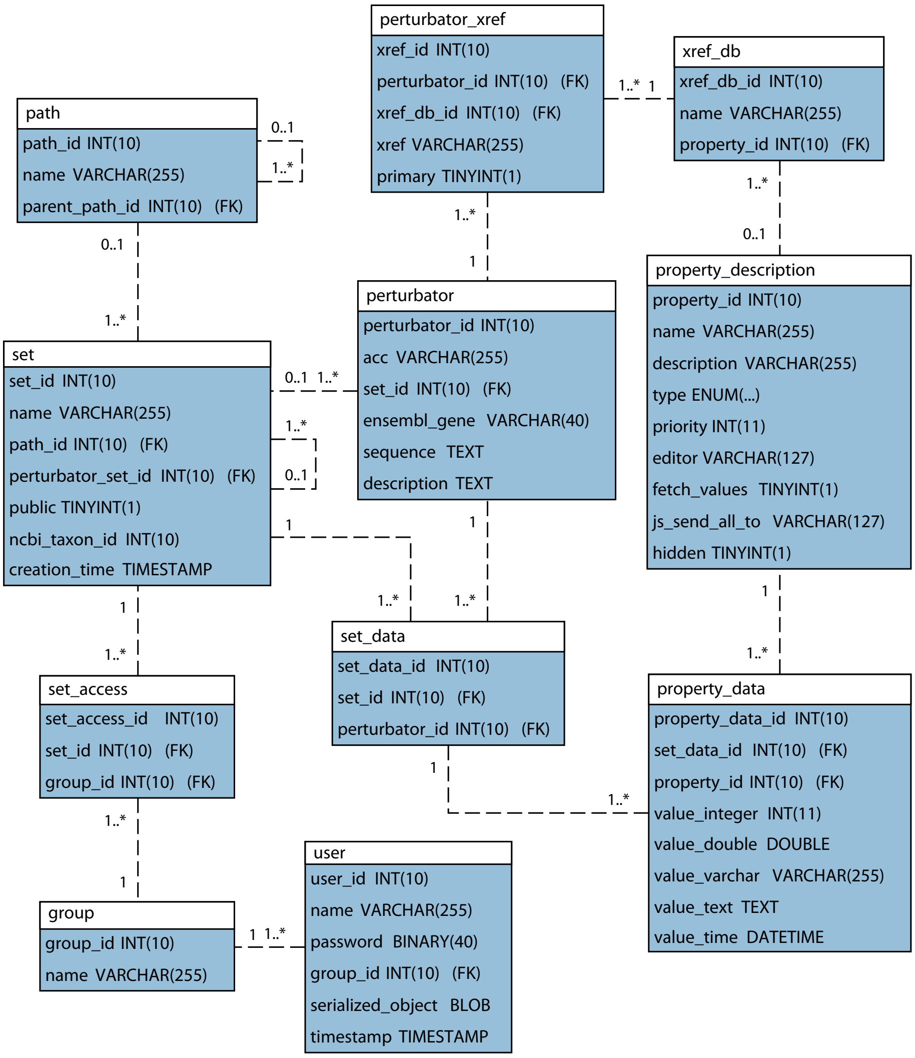


Figure 2.1: DSViewer database scheme.

Each data set stored in the database is represented by an entry in the `set` table. Data sets are organized as a tree, whose structure is managed by the `path` table. Each data set consists of a list of 'entries' (table `set_data`), each associated with an individual RNAi trigger (table `perturbator`). An 'entry' represents a set of experimental readouts or annotations. Each value, depending on its type, is stored in one of the `value_*` fields of the `property_data` table. Names and types of possible properties are stored in the `property_description` table. RNAi triggers are organized in libraries, which are also stored as separate data sets. Each perturbator is annotated with a gene name and other identifiers stored in the `perturbator_xref` table. To facilitate unified querying interface, identifiers are also recognized as properties. Access to different data sets is secured at a 'user' and a 'group' levels (tables `set_access`, `group` and `user`).

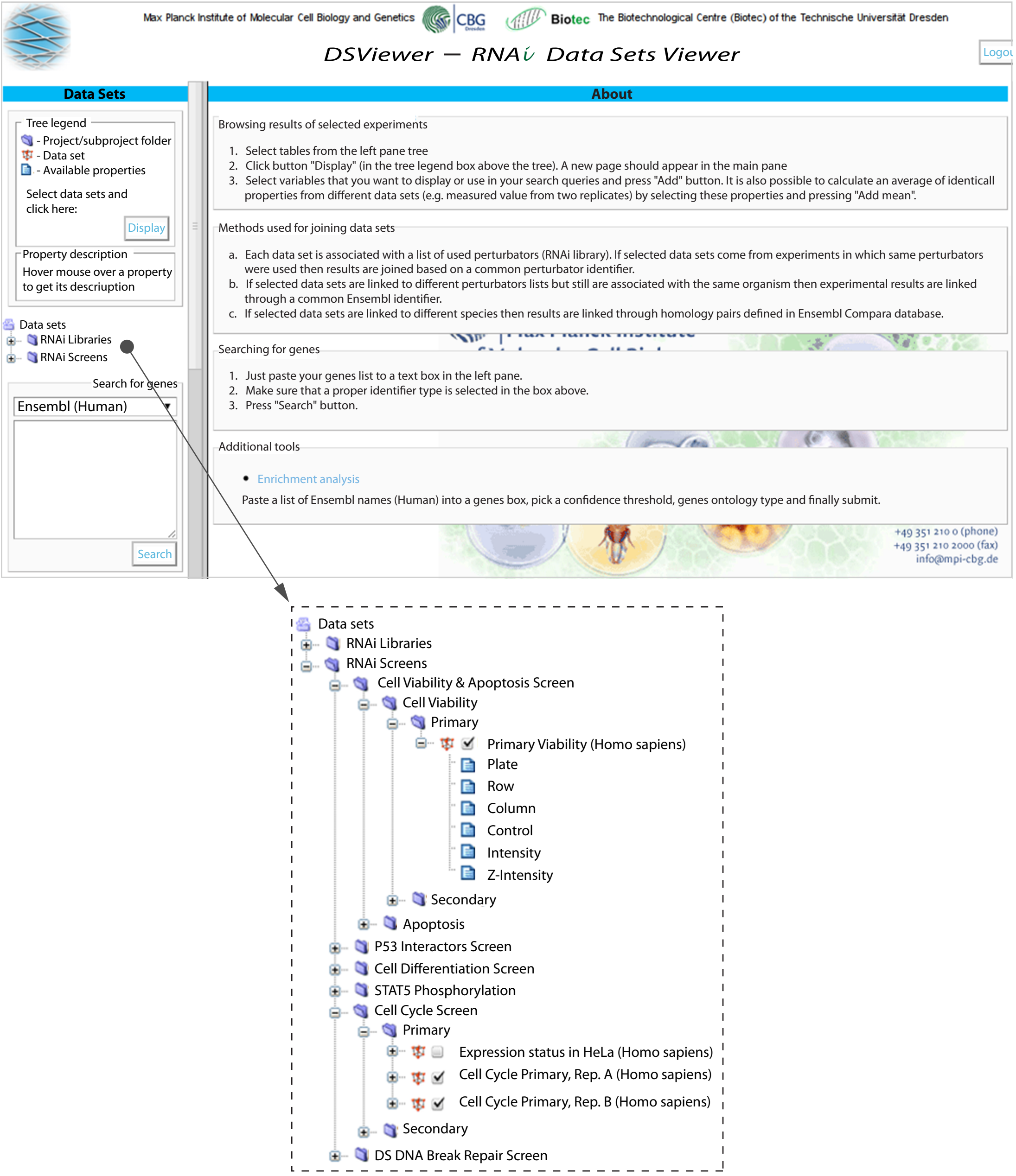
notypic information (*i.e.* quantitative and qualitative), as well as genetic annotations. Prioritizing this requirement, DSViewer was designed to allow storage of results of any gene-centered high-throuthput experiment.

The underlying database scheme (Figure 2.1) treats all numerical and textual values associated with a gene as its 'properties'. A property can be defined as any raw or normalized value reported by an experimental assay or any phenotypic or functional annotation (*i.e.* gene description). Properties provided by different data-sets are stored centrally, in a single database table (see `property_data` on Figure 2.1). Initial transfer of experimental results into the database is performed by a standalone application, which accepts Excel or comma- or tab-separated files as an input.

Different data-sets can be integrated with each other on different levels. If both data sets are results of RNAi screenings performed using the same library of silencing triggers, the results are automatically joined by a trigger identifier. If experiments were performed using different libraries but on the same species, the results are joined by the common gene identifiers. Results of experiments carried out on different species are joined by best-best orthology associations acquired from the Ensembl Compara database (Hubbard *et al.*, 2009).

2.2.2 User interface

DSViewer allows users to merge multiple data sets, filter them for combinations of phenotypes and display or save the results in a form of a table. A typical user session starts with an authorisation screen followed by a welcome page, which contains a short users' guide (Figure 2.2). In order to integrate and view results of multiple experiments, a user selects one or more data sets displayed in a navigation tree located in the left panel of the web page. After accepting the selection, the main panel is replaced by the properties selection window, which displays all variables provided by the selected data sets (Figure 2.3). Using the window controls, the user selects properties that should be displayed in the final results table. Numerical properties of the same type can be combined producing a new variable calculated as an average of the respective values of the combined variables. This feature is especially useful for generating new data sets that are summarizing multiple replicates of the same experiment. When the properties selection is confirmed, the selection panel is replaced by the results viewer (Figure 2.4).



Max Planck Institute of Molecular Cell Biology and Genetics
CBG
Biotec
The Biotechnological Centre (Biotec) of the Technische Universität Dresden
DSViewer — RNAi Data Sets Viewer
Logou
Data Sets
About
Tree legend
- Project/subproject folder
- Data set
- Available properties
Select data sets and click here:
Display
Property description
Hover mouse over a property to get its descriuption
Search for genes
Ensembl (Human)
Search
+49 351 210 0 (phone)
+49 351 210 2000 (fax)
info@mpi-cbg.de

Browsing results of selected experiments

1. Select tables from the left pane tree
2. Click button "Display" (in the tree legend box above the tree). A new page should appear in the main pane
3. Select variables that you want to display or use in your search queries and press "Add" button. It is also possible to calculate an average of identicall properties from different data sets (e.g. measured value from two replicates) by selecting these properties and pressing "Add mean".

Methods used for joining data sets

a. Each data set is associated with a list of used perturbators (RNAi library). If selected data sets come from experiments in which same perturbators were used then results are joined based on a common perturbator identifier.
b. If selected data sets are linked to different perturbators lists but still are associated with the same organism then experimental results are linked through a common Ensembl identifier.
c. If selected data sets are linked to different species then results are linked through homology pairs defined in Ensembl Compara database.

Searching for genes

1. Just paste your genes list to a text box in the left pane.
2. Make sure that a proper identifier type is selected in the box above.
3. Press "Search" button.

Additional tools

- Enrichment analysis

Paste a list of Ensembl names (Human) into a genes box, pick a confidence threshold, genes ontology type and finally submit.

- Data sets
  - RNAi Libraries
  - RNAi Screens
    - Cell Viability & Apoptosis Screen
      - Cell Viability
        - Primary
          - ✓ Primary Viability (Homo sapiens)
            - Plate
            - Row
            - Column
            - Control
            - Intensity
            - Z-Intensity
        - Secondary
      - Apoptosis
    - P53 Interactors Screen
    - Cell Differentiation Screen
    - STAT5 Phosphorylation
    - Cell Cycle Screen
      - Primary
        - ☐ Expression status in HeLa (Homo sapiens)
        - ✓ Cell Cycle Primary, Rep. A (Homo sapiens)
        - ✓ Cell Cycle Primary, Rep. B (Homo sapiens)
      - Secondary
    - DS DNA Break Repair Screen

Figure 2.2: DSViewer main page.

The main page of the DSViewer web application. It consists of the short users guide and a data set selection tree (expanded in the blow-up) with a genes search box located below the tree. All available data sets are organized in a structure resembling file system directories. Here, each branch represents a project or a set of replicates. Results of different experiments can be merged by selecting multiple data sets.

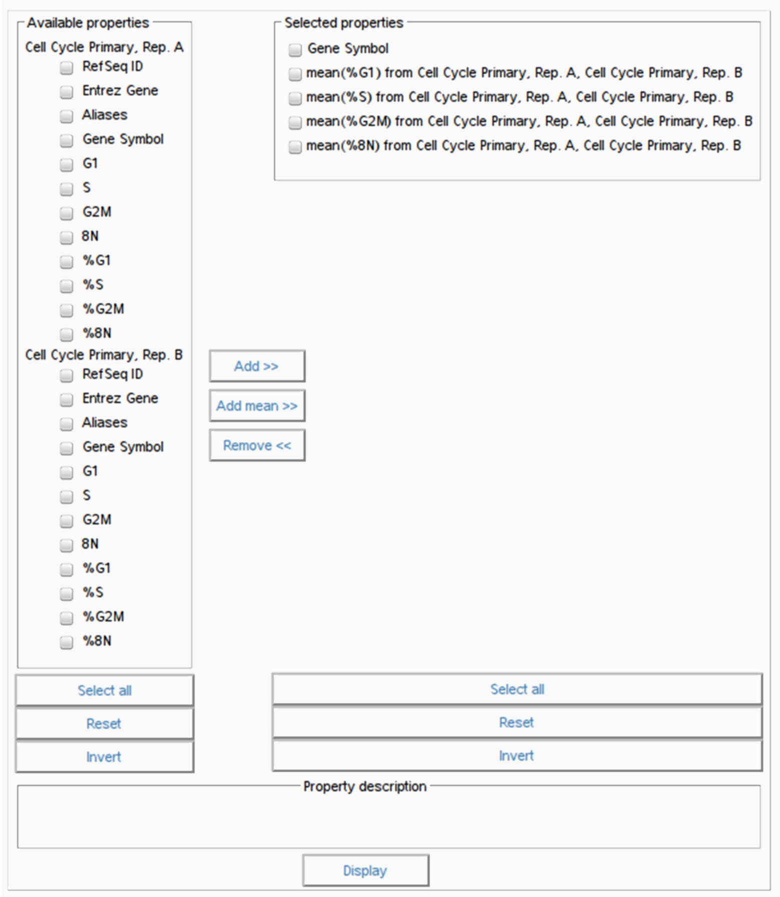
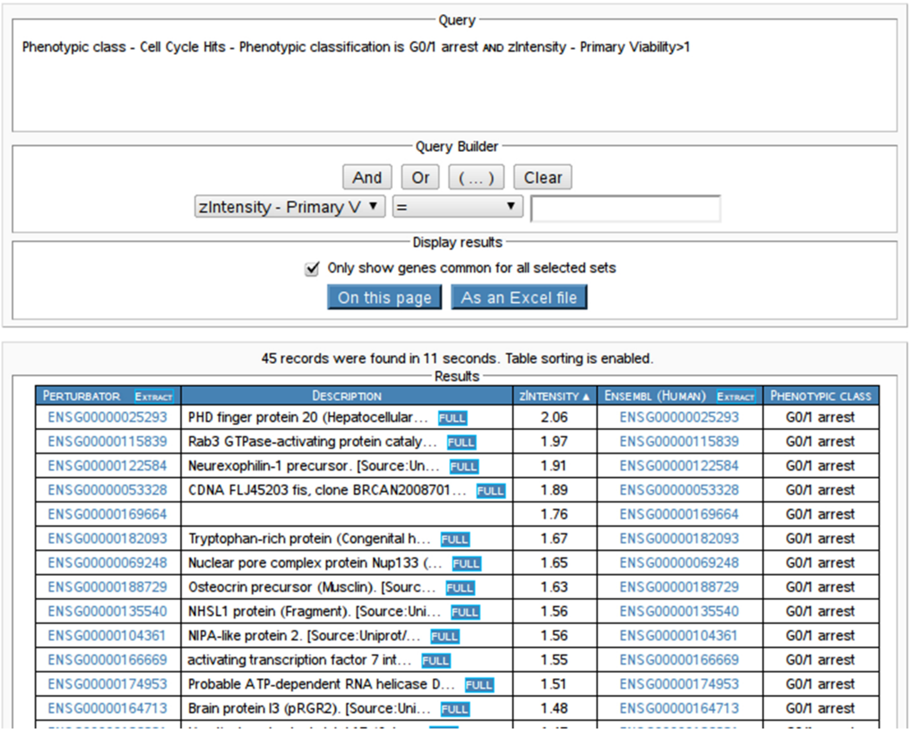

Available properties
Cell Cycle Primary, Rep. A
☐ RefSeq ID
☐ Entrez Gene
☐ Aliases
☐ Gene Symbol
☐ G1
☐ S
☐ G2M
☐ 8N
☐ %G1
☐ %S
☐ %G2M
☐ %8N
Cell Cycle Primary, Rep. B
☐ RefSeq ID
☐ Entrez Gene
☐ Aliases
☐ Gene Symbol
☐ G1
☐ S
☐ G2M
☐ 8N
☐ %G1
☐ %S
☐ %G2M
☐ %8N
Selected properties
☐ Gene Symbol
☐ mean(%G1) from Cell Cycle Primary, Rep. A, Cell Cycle Primary, Rep. B
☐ mean(%S) from Cell Cycle Primary, Rep. A, Cell Cycle Primary, Rep. B
☐ mean(%G2M) from Cell Cycle Primary, Rep. A, Cell Cycle Primary, Rep. B
☐ mean(%8N) from Cell Cycle Primary, Rep. A, Cell Cycle Primary, Rep. B
Add >>
Add mean >>
Remove <<
Select all
Reset
Invert
Select all
Reset
Invert
Property description
Display

Figure 2.3: DSViewer properties selection panel.

The panel contains two lists of properties. A list in the left side shows all variables provided by the selected data sets. Properties selected for displaying in the results panel are shown in the right side. Numerical variables of the same type can be averaged, allowing a summarization of multiple replicates of the same experiment.



Query

Phenotypic class - Cell Cycle Hits - Phenotypic classification is G0/1 arrest AND zIntensity - Primary Viability>1

Query Builder

Display results

☑ Only show genes common for all selected sets

45 records were found in 11 seconds. Table sorting is enabled.

Results

PERTURBATOR EXTRACT	DESCRIPTION	ZINTENSITY ▲	ENSEMBL (HUMAN) EXTRACT	PHENOTYPIC CLASS
ENSG00000025293	PHD finger protein 20 (Hepatocellular... FULL	2.06	ENSG00000025293	G0/1 arrest
ENSG00000115839	Rab3 GTPase-activating protein cataly... FULL	1.97	ENSG00000115839	G0/1 arrest
ENSG00000122584	Neurexophilin-1 precursor. [Source:Un... FULL	1.91	ENSG00000122584	G0/1 arrest
ENSG00000053328	CDNA FLJ45203 fis, clone BRCAN2008701... FULL	1.89	ENSG00000053328	G0/1 arrest
ENSG00000169664		1.76	ENSG00000169664	G0/1 arrest
ENSG00000182093	Tryptophan-rich protein (Congenital h... FULL	1.67	ENSG00000182093	G0/1 arrest
ENSG00000069248	Nuclear pore complex protein Nup133 (... FULL	1.65	ENSG00000069248	G0/1 arrest
ENSG00000188729	Osteocrin precursor (Musclin). [Sourc... FULL	1.63	ENSG00000188729	G0/1 arrest
ENSG00000135540	NHSL1 protein (Fragment). [Source:Uni... FULL	1.56	ENSG00000135540	G0/1 arrest
ENSG00000104361	NIPA-like protein 2. [Source:Uniprot/... FULL	1.56	ENSG00000104361	G0/1 arrest
ENSG00000166669	activating transcription factor 7 int... FULL	1.55	ENSG00000166669	G0/1 arrest
ENSG00000174953	Probable ATP-dependent RNA helicase D... FULL	1.51	ENSG00000174953	G0/1 arrest
ENSG00000164713	Brain protein I3 (pRGR2). [Source:Uni... FULL	1.48	ENSG00000164713	G0/1 arrest

Figure 2.4: DSViewer results panel.

The results panel contains a query builder, which allows data filtering, and a sortable table with the results of data sets integration.

The most important feature of the panel used for displaying results is the query builder (Figure 2.4), which allows constructing and editing complex selection criteria containing filtering statements joined by logical operators ‘and’ and ’or’. Precedence of the logical expressions can be altered by enclosing them with parenthesis. Results of the filtering are either presented as a table or can be downloaded as a spreadsheet file containing genes or RNAi reagents in rows and selected variables (phenotypes) in columns. In case of specific gene list query, the database retrieves all mapped RNAi triggers and information of phenotypes that were reported, dispaying them as separate tables, one for each data set.

Additional features of the web-based results viewer are table sorting, link-outs to external databases and identifiers extraction, which allows copying of selected gene identifiers into external applications (*i.e.* enrichment analysis tools).

2.3 Current content

Currently, the database integrates information about two in-house produced genome-wide libraries of esiRNA reagents targeting mouse and human genomes (Buchholz *et al.*, 2006; Ding *et al.*, 2009) and results of numerous primary and secondary systematic RNAi screens performed in our laboratory (Table 2.1).

Table 2.1: Systematic genome-wide RNAi screens accessible through DSViewer interface

Investigated process	Species	Reference
Cell cycle	Human	Kittler *et al.* (2007b)
Pluripotency of mouse embryonic stem cells	Mouse	Ding *et al.* (2009)
Cell viability	Human	Theis *et al.* (2009)
Synthetic lethality of TP53 interactors in cancer cells	Human	Unpublished data
Double-stranded DNA repair through homologous recombination	Human	Unpublished data

2.4 Conclusions

DSViewer is a computational tool developed for rapid and effortless integration of results obtained from high-throughput experiments. Thanks to an advanced querying capabilities, DSViewer can facilitate selection of candidate genes that share user-specified combination of phenotypes (*e.g.* 'cell-cycle arrest in G1 phase' and 'increased metabolic activity' or 'increased cell size'). The main feature of DSViewer is flexibility of its database scheme, which allows storage of any type of the phenotypic data, as well as result of DNA microarrays or other gene-centered data-sets. DSViewer was a primary application employed by data analysis pipelines described in the following chapters of this thesis.

Chapter 3

Comparative profiling identifies C13orf3 as a component of the Ska complex required for mammalian cell division

Proliferation of mammalian cells requires the coordinated function of many proteins to accurately divide a cell into two daughter cells. Several RNAi screens have identified previously uncharacterised genes that are implicated in mammalian cell division. The molecular function for these genes needs to be investigated to place them into pathways. Phenotypic profiling is a useful method to assign putative functions to uncharacterised genes.

This chapter shows the meta-analysis of two loss-of-function screens combined with protein localisation data. The utility of this approach is shown by defining a function of the previously uncharacterised gene C13orf3 during cell division. C13orf3 localises to centrosomes, the mitotic spindle, kinetochores, spindle midzone, and the cleavage furrow during cell division and is specifically phosphorylated during mitosis. Furthermore, C13orf3 is required for centrosome integrity and anaphase onset. Depletion by RNAi leads to mitotic arrest in metaphase with an activation of the spindle assembly checkpoint and loss of sister chromatid cohesion.

Proteomic analyses identify C13orf3 (Ska3) as a new component of the Ska complex and show a direct interaction with a regulatory subunit of the protein phosphatase PP2A. All together, these data identify C13orf3 as an important factor for metaphase to anaphase progression and highlight the potential of combined RNAi screening and protein localisation analyses.

My main contribution to this work is presented in section 3.2 of this chapter, which describes a novel methodology for a combined hit selection analysis based on multiple RNAi screens. Results of my analysis established foundations for further experimental studies, whose results are reported in section 3.3.

3.1 Introduction

Cell division and mitosis of eukaryotic somatic cells require the coordinated function of many proteins in a temporally and spatially well-orchestrated process (Nasmyth, 2002; Nigg, 2001; Varetti & Musacchio, 2008). Mitosis can be subdivided into different phases mainly depending on morphological features. The purpose of early mitosis from prophase to metaphase is the establishment of a bipolar spindle with all kinetochores attached amphitelic to spindle microtubules (Musacchio & Salmon, 2007). Proper attachment and the creation of tension at the kinetochores are believed to be the key factors for silencing of the spindle assembly checkpoint (SAC).

Silencing of the SAC leads to the activation of the E3 ubiquitin–protein ligase, APC/C (Sullivan & Morgan, 2007). The APC/C is a multiprotein complex composed of at least 12 subunits, of which, among others, the subunits Cdc16 and Cdc27 are crucial for its activity (Peters, 2006; Thornton *et al.*, 2006); it promotes execution of anaphase by polyubiquitylation of its main substrates cyclin B1 and securin, thereby targeting them for destruction by the proteasome. Degradation of cyclin B1 results in a decrease in Cdk1 activity that is required for entry into the late phases of mitosis. Loss of securin allows activation of separase required for sister chromatid separation (Sullivan & Morgan, 2007).

In vertebrate cells, arm cohesion is largely lost during prophase and prometaphase in a separase-independent pathway requiring polo-like kinase 1 (Plk1) and aurora kinase B (AurkB) activity to facilitate sister chromatid resolution (Waizenegger *et al.*, 2000; Watanabe, 2005). In contrast, centromeric cohesion is preserved until anaphase by the protein shugoshin-like 1 (SGOL1), which recruits protein phosphatase 2A (PP2A) to centromeric cohesion, thereby counter-acting phosphorylation by Plk1.

Although parts of the genes and the mechanisms that guard mammalian cell division have been identified, others remain elusive. The complexity of mammalian cell division calls for a systems-level approach to understand the sophisticated interaction and regulation of proteins involved (Kittler *et al.*, 2008). Loss-of-function screening by RNAi is a valuable strategy for the systematic analysis of genes implicated in cell-cycle regulation, and different methods to carry out RNAi experiments in mammalian cells are available (Sachse & Echeverri, 2004). We and others have developed and successfully utilised endoribonuclease-prepared short interfering RNAs (esiRNAs) as mediator for RNAi (Fazzio *et al.*, 2008; Galvez *et al.*, 2007; Kittler *et al.*, 2004, 2007b; Yang *et al.*, 2002b). As esiRNAs are highly specific, they are well suited for RNAi experiments, especially for large-scale RNAi screens (Kittler *et al.*, 2007c). A previously conducted genome-scale esiRNA screen on cell-cycle progression in mammalian cells (Kittler *et al.*, 2007b) identified many previously uncharacterised genes implicated in this process. The use of multiparametric analysis in combination with hierarchical clustering allowed the placement of some of these genes into pathways (Kittler *et al.*, 2007b). However, it remains challenging to interpret multiparametric phenotypic data to build valid biological hypotheses, and for many uncharacterised genes the molecular role during cell division remains elusive.

Localisation is an independent indicator of gene function (Wang *et al.*, 2008c), which may provide valuable information in addition to loss-of-function data. Antibodies are useful to determine the localisation pattern of proteins, and are commercially available for many known cell-cycle proteins. However, generating antibodies for uncharacterised genes is time-consuming and cost-intensive. Tagging of genes with fluorescent proteins is a rapid and cost-effective alternative to antibodies, which also allows the dynamic localisation of proteins in living cells, and collections of tagged genes based on cDNA constructs have been assembled (Pepperkok & Ellenberg, 2006). However, expression of tagged genes from cDNA constructs can be problematic because the genomic context of the gene is not preserved. As a consequence, the gene is often expressed at nonphysiological levels, which can lead to mislocalisation of the protein. Recently, the TransgeneOmics approach has been developed to allow rapid tagging of many genes that preserves the genomic context (Poser *et al.*, 2008). Using recombineering technology (Muyrers *et al.*, 2001) to tag genes, encoded on a bacterial artificial chromosome (BAC), allows the expression of genes close to the endogenous level. As a BAC usually contains all cis-regulatory elements of the promoter, 3'-UTR, and the coding region, the transgene maintains its physiological expression levels and splicing pattern (Poser *et al.*, 2008), a feature especially interesting for genes implicated in cell-cycle control.

To test whether localisation data help to refine phenotypic profiles, I exemplified the analysis of a subcluster derived from two RNAi screens that were enriched for known regulator proteins of mitosis by using the TransgeneOmics approach. I nominated the previously uncharacterised protein, C13orf3 (also known as Rama1), to exhibit a similar localisation pattern and phenotypic features to the protein Ska1 (spindle and kinetochore associated protein 1). A detailed characterisation identified a direct interaction of C13orf3 with members of the Ska complex, described as a two-component complex, composed of Ska1 (C18orf24) and Ska2 (Fam33a), with a critical role in the maintenance of the metaphase plate and progression through mitosis (Hanisch *et al.*, 2006).

Investigation reported in this chapter showed that C13orf3 (Ska3) is an integral part of the Ska complex and localises to the mitotic spindle, kinetochores, and cleavage furrow during mitosis. In addition, it was shown that C13orf3 is required for the maintenance of a bipolar spindle. Depletion of C13orf3 leads to an arrest in a metaphase-like state with an activation of the SAC and sister chromatid separation. These findings underline the importance of C13orf3 in the mitotic progression of mammalian cells and show that the combination of phenotypic profiling and localisation data improves the predictive power helping to identify pathways for genes with important roles during cell division.

3.2 Comparative profiling of phenotypes from systematic RNAi screens

To predict functions of previously uncharacterised mitosis-related genes, I utilized a data set from a cell-cycle esiRNA screen carried out previously in my laboratory (Kittler *et al.*, 2007b). To refine this data, a genome-wide cell viability screen was performed. Both screens were done in the same cell line (HeLa) and with the same library of gene silencing triggers, thereby reducing between-experiment variability.

3.2.1 Data combination and normalization

Input data

In the cell-cycle screen, an analysis of genes implicated in the cell-cycle progression was carried out using DNA content analysis combined with laser-scanning cytometry. Phenotypes were expressed as proportions of cells in different phases of the cell cycle (G1, S, G2/M) and cells with polyploidy phenotype for each knockdown (Kittler *et al.*, 2007b). The final result was calculated as an average from two replicates. In the cell viability screen, cellular metabolic activity was measured with alamarBlue® assay. For each knockdown, it returned a single value corresponding to fluorescent intensity of a reporter dye. Since both screens were performed using the same library of esiRNA compounds, the data sets were joined by unique esiRNA identifiers. The final list comprised 16 363 combined phenotypes.

Data normalization

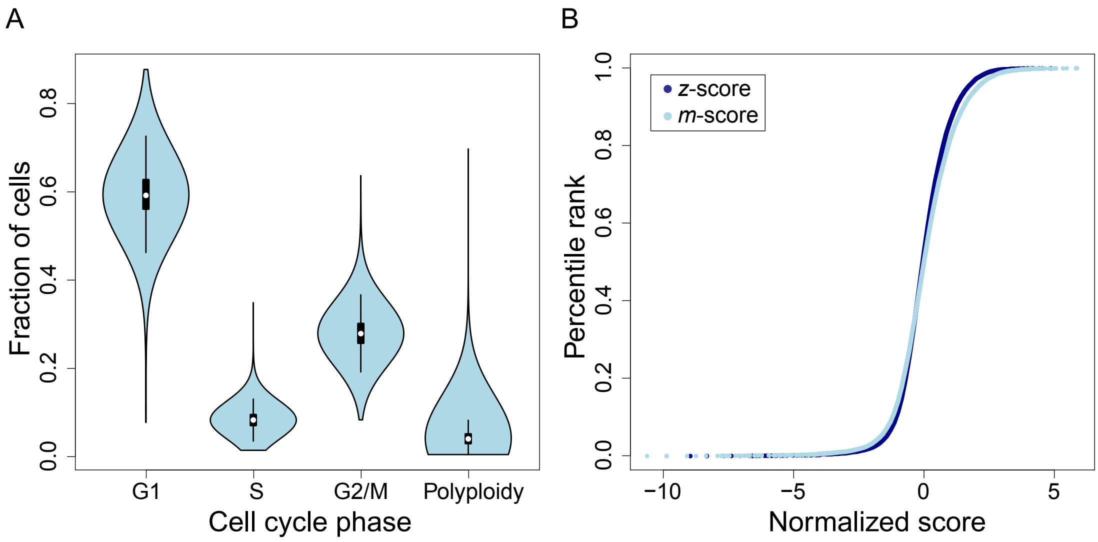


Figure 3.1: Distribution of cell-cycle phenotypes.

(A) The violin plot shows distributions of fractions of cells in cell-cycle phases after the genome-wide esiRNA treatment (average of two replicates). White dots represent median values, upper and lower bounds of black boxes represent 1st and 3rd quartiles, vertical lines extend 1. 5IQR beyond the box boundaries. (B) Comparison of z-score and m-score normalization *vs.* percentile rank normalization of fractions of cells in G1 phase. Each dot represent a phenotype of a single gene knock-down.

Table 3.1: Skewness and kurtosis of the cell-cycle phenotypes

	Raw values[1]		$\log_2$ transformed	
Variable	Skewness[2]	Kurtosis	Skewness	Kurtosis
G1	−0.36	7.10	−2.28	29.41
S	1.02	10.42	−0.54	5.16
G2/M	0.54	7.61	−0.53	6.48
Polyploid	7.51	198.42	−0.06	3.81

[1] Ratio of cells compared to total.

[2] The normal distribution has both skewness and kurtosis equal to zero.

Originally, each variable of the cell-cycle data set was normalized using z-scores (Kittler *et al.*, 2007b). However, the percentages of cells in each phase of te cell cycle are not normally distributed (Lillefors normality test) and their distributions are highly skewed (Table 3.1 and Figure 3.1A). The same results were obtained for a hypothesis of a log-normal distribution. Thus, the z-score normalization of the cell-cycle-related variables would lead to a strong bias in the further analysis and increase the false discovery rate. In addition, the analysis of variance (ANOVA) did not indicate a significant between-plate variability making the within-plate normalization not necessary.

To normalize all variables in the study, I applied a non-parametric approach and converted all variables from the study into their respective percentile ranks. The percentile rank of a score is the percentage of scores that are the same or lower (Crocker & Algina, 2006) and can be easily calulated from the empirical distribution as a rank of the score divided by the total number of observations:

$$percentile\ rank(x) = \frac{B + 0.5E}{n} \cdot 100\% = \frac{r - 0.5}{n} \cdot 100\% \tag{3.1}$$

where B is a number of scores lower than x, E is the number of scores equal to x and r is rank of the score.

Compared to z-score or m-score normalization, percentile ranks reduce the influence of extreme values and introduce higher separation of mild scores (Figure 3.1B). This effect gives an advantage during detection of weakly-scoring genes, which due to lower z-scores would be missed during classical threshold-based hits selection.

3.2.2 Clustering of phenotypes

Phenotypic signature of a gene can be used as a reporter of the gene function in the cell cycle and, as a consequence, genes showing similar phenotypes upon knockdown might participate in the same biological processes (Kittler *et al.*, 2007b). To group genes by their phenotypic signature, I applied hierarchical clustering.

The essential part of the hierarchical clustering algorithm is an appropriate choice of a metric used for measuring distances between phenotypic signatures of two genes. Because a complete phenotypic signature consists of variables obtained from two experiments, I used a weighted euclidean metric:

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n} w_i \left(x_i - y_i\right)^2} \tag{3.2}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{3.3}$$

Weights were chosen to ensure equal influence of both experiments in the calculated distance: $w = 0.5$ for the viability variable and $w = 0.125$ for each of the four cell-cycle variables. The final tree was constructed with the UPGMA linkage.

3.2.3 Cluster selection

To identify a set of genes potentially involved in progression through mitosis, I searched for a cluster with a significant enrichment of known genes annotated with Gene Ontology terms ‘mitosis’ and ‘cell cycle’. Such an enrichment suggests that uncharacterized genes within the same cluster may also be related to biological processes described by the selected GO terms.

For each branch of the tree, I selected genes belonging to its all subclusters. If the number of genes exceeded 10, I calculated an enrichment of the two GO terms. The main advantage of my enrichment-based approach of cluster selection procedure is that it does not require the tree-cutting step. The optimal number of cluster does not have to be calculated because the algorithm is analyzing the whole dendrogram in search for a cluster of interest.

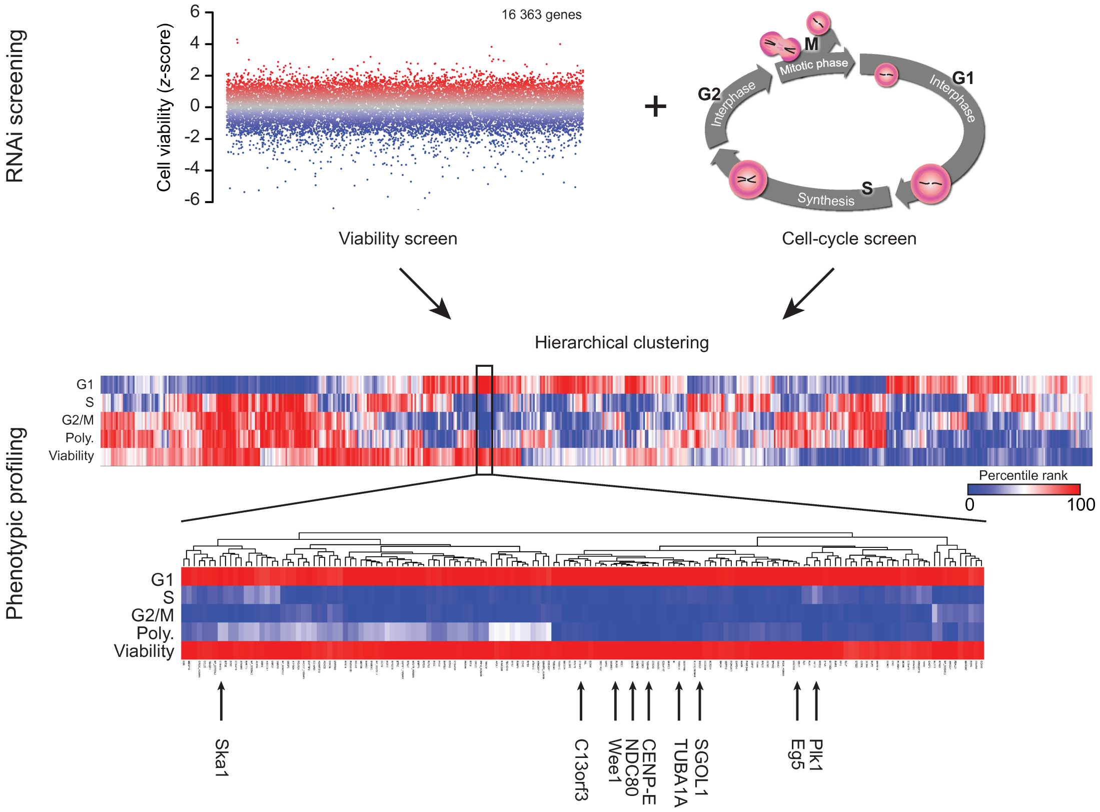


Figure 3.2: Strategy of phenotypic profiling.

Genome-scale RNAi data from a cell cycle and a viability screen were combined for multiparametric hierarchical clustering. Values for five parameters, that is, proportions of cells in phases G1, S, G2/M, and aneuploid cells (8N), as well as cell viability (decreased: blue, increased: red), were converted in percentile ranks and used to build a hierarchical tree sorting the genes by similarity of their phenotype. A subcluster enriched with genes with important mitotic functions is shown in the blow-up.

As a result, the highest scores were obtained for the cluster containing 154 genes and the enrichment values were 3.0-fold of cell-cycle genes and 5.5-fold of mitosis-related genes (*e.g.* CENP-E, SGOL1, Eg5, Ska1, Plk1). This cluster was selected for further experimental validation (Figure 3.2).

3.2.4 Addition of cell viability data refines the clustering results

To test wether combination of different RNAi screens increases the predictive power to nominate annotations for uncharacterized genes, I performed an analogous analysis for the cell-cycle data alone. In the clustering without the viability data, the enrichment of selected GO terms in the highest scoring cluster, which contained 1069 genes, was only 1.6-fold and 2.2-fold for cell cycle and mitosis associated genes, respectively, compared to enrichment of 3.0-fold and 5.5-fold in the cluster obtained from the analysis of combined data (Figure 3.3). This result implied that combination of data derived from independent RNAi screens using different assays reduced the noise and improve the data quality.

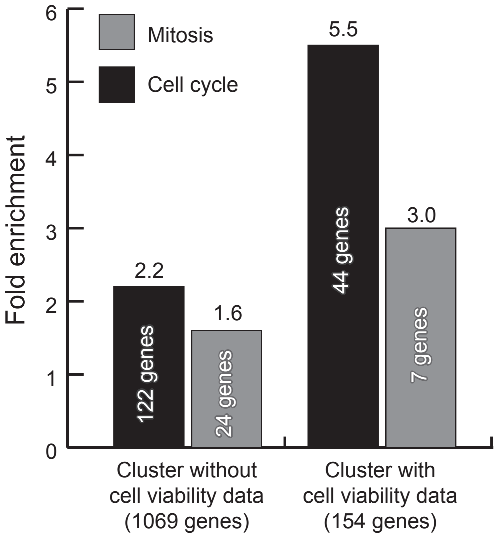


Figure 3.3: Enrichment of ontology terms after introducing cell viability data.

Hierarchical clustering of cell-cycle data alone followed by the cluster selection procedure resulted in a cluster comprising 1069 genes, among which 122 (1%) were annotated with the 'cell cycle' and 24 (2%) with the 'mitosis' GO term. After merging with the cell viability data, the clustering algorithm reported a refined cluster comprising 154 genes, among which 44 (29%) and 7 (5%) were annotated with the 'cell cycle' and 'mitosis' terms respectively.

3.3 Experimental results and discussion

3.3.1 Combining phenotypic profiling with protein localisation

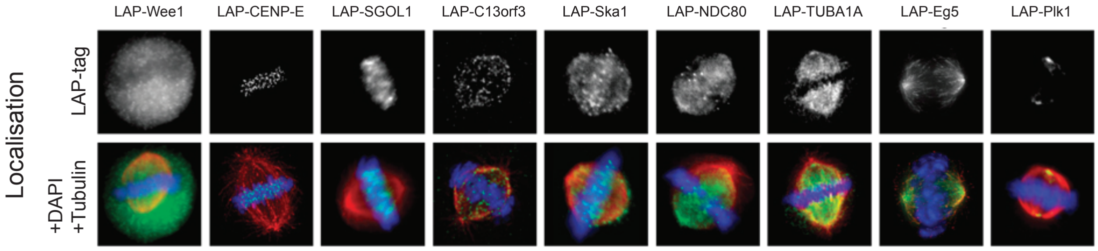


Figure 3.4: Localisation of selected BAC transgenic cell lines.

Selected examples of imaged BAC transgenic cell lines stained for α-tubulin (red), DNA (blue), and LAP-tag (green).

To further refine the profile within the cluster, the protein localisation in mitotic cells for known cell-cycle genes and previously uncharacterised genes was determined using the BAC-based TransgeneOmics approach (Kittler *et al.*, 2005b; Poser *et al.*, 2008). BAC constructs maintain the genomic context of a gene and usually contain all cis-regulatory elements for gene expression. Furthermore, they typically integrate at low copy number, thereby allowing physiological expression of the tagged proteins (Kittler *et al.*, 2005b; Poser *et al.*, 2008). A modified version of the 'localisation and affinity purification' (LAP) tag (Cheeseman & Desai, 2005) consisting of an EGFP and an S-peptide sequence was used. The analysis of images from these BAC-transgenic HeLa cell lines during mitosis confirmed the known localisation of cell-cycle-relevant proteins such as CENP-E (kinetochores), SGOL1 (centromeres), Eg5 (centrosomes, mitotic spindle), and Ska1 (kinetochores, mitotic spindle) (Figure 3.4), showing the utility of the TransgeneOmics approach. To improve the phenotypic profiling, we determined the localisation of 52 known and uncharacterised BAC-tagged proteins from the subcluster comparing their localisation with each other to identify proteins with similar localisation patterns. Interestingly, the analysis of the uncharacterised protein, C13orf3, showed a spindle and kinetochore localisation during mitosis, most similar to the localisation pattern of Ska1 in the subcluster (Figure 3.4) (Hanisch *et al.*, 2006; Rines *et al.*, 2008). Although the phenotypic data alone were insufficient to link C13orf3 to Ska1, the combination with the localisation data predicted that Ska1 and C13orf3 might physically and/or genetically interact. To test this hypothesis, C13orf3 was selected for an in-depth analysis. Comparative sequence analysis identified putative C13orf3 orthologues in mammals, birds, amphibians, and bony fish (e.g., mouse: ENSMUSG00000021965, chicken: ENSGALG00000017128, frog: ENSXETG00000009595, and zebrafish: ENSDARG00000067746, respectively), but not in invertebrates. Structure-based bioinformatics analyses (Godzik, 2003; Sippl & Flöckner, 1996) identified a Gle2-binding sequence motif (GLEBS motif) at the C-terminal region of C13orf3 (aa 345–382). Interestingly, an additional putative GLEBS motif has been proposed in the N-terminal region (aa 17–51) of C13orf3 by Gaitanos *et al.* (E Nigg, personal communication, 2008). GLEBS motifs are present in Bub1 and BubR1 and have been structurally characterised to mediate binding to Bub3 (Larsen *et al.*, 2007), substantiating a potential role of C13orf3 in mitosis.

3.3.2 C13orf3 localises to prominent structures during mitosis

To analyse the dynamic localisation of C13orf3 during cell division and to substantiate the overlapping localisation with Ska1 observed in the profiling study (Figure 3.4), the C13orf3 BAC-transgenic HeLa cell line was used and dividing cells were imaged using fluorescence time-lapse microscopy. The protein was predominantly cytoplasmatic during interphase, with a noticeable concentration around the nuclear envelope (Figure 3.5Aa, j). In prophase, just before the nuclear envelope breakdown, an accumulation of C13orf3 at the centrosomes was readily detectable (Figure 3.5Ab and B). During prometaphase and metaphase, C13orf3 localised to the mitotic spindle as well as to the kinetochores (Figure 3.5Ac, d and B). Upon entry into anaphase, the fusion protein was enriched at the spindle (Figure 3.5Ae and B) and at the spindle midzone in late anaphase and telophase (Figure 3.5Af, g, and B). During cytokinesis, C13orf3 was found at the cleavage furrow (Figure 3.5Ah, i, and B). The localisation to important mitotic structures and the overlap with the localisation pattern of Ska1 (Hanisch *et al.*, 2006) underlines a potential interaction of these proteins.

3.3.3 C13orf3 is required for anaphase onset

The prominent localisation of C13orf3 suggested to carry out a more detailed phenotypic analysis. To confirm and validate findings from the genome-wide RNAi screens

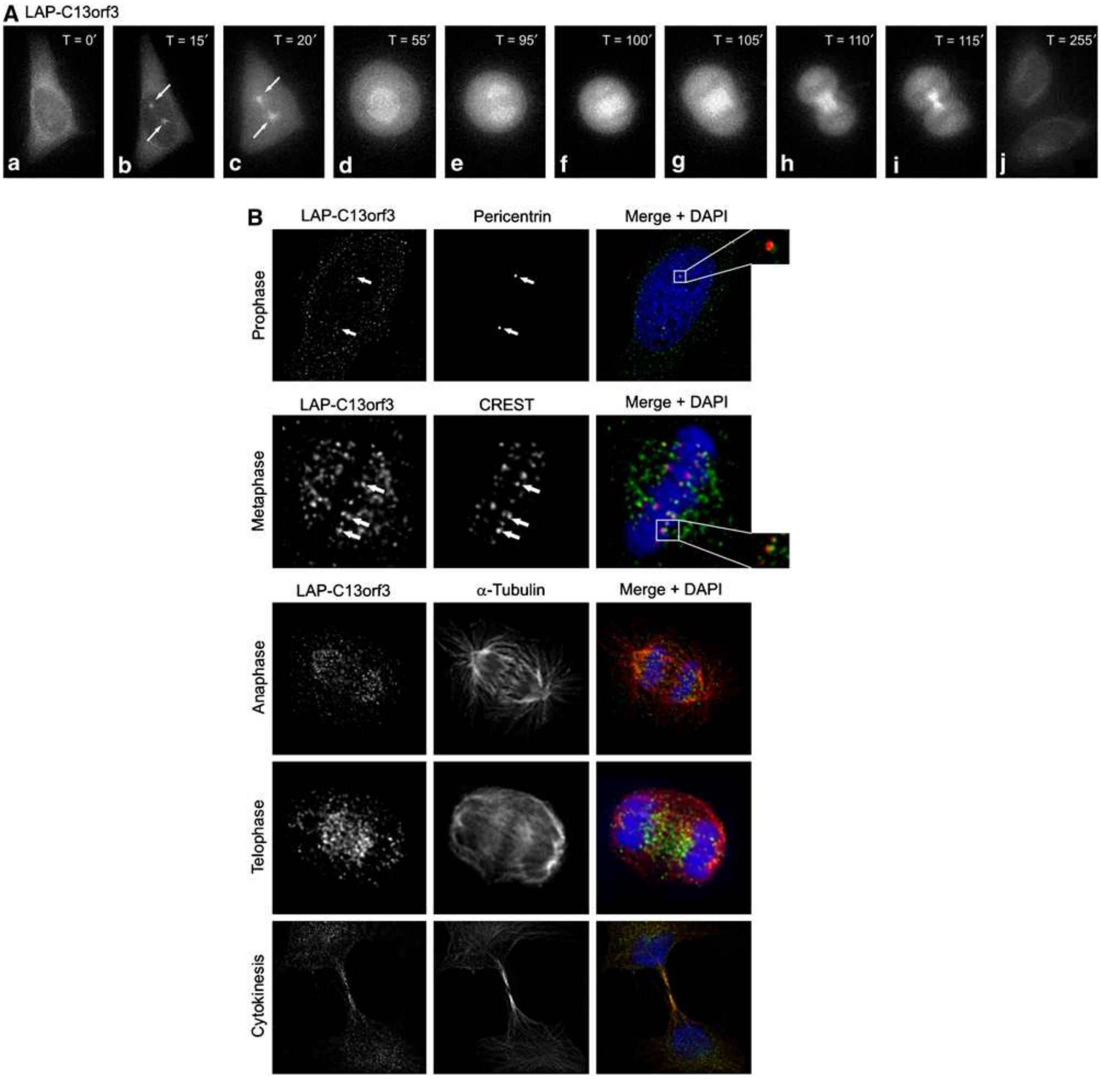


Figure 3.5: Localisation pattern of C13orf3 in HeLa cells

(A) Selected frames from fluorescence time-lapse microscopy of HeLa cells expressing LAP-tagged C13orf3 at indicated time points are shown in interphase (a, j), prophase (b), prometaphase (c), metaphase (d), early anaphase (e), late anaphase (f), telophase (g), early cytokinesis (h), and late cytokinesis (i). (B) Immunofluorescence microscopy of LAP-C13orf3 co-stained with pericentrin, CREST, and alpha-tubulin antibodies during indicated cell cycle phases. Arrows and blow-ups point to areas of colocalisation.

(Figure 3.2), first a repeated DNA content analysis by flow cytometry was performed with two independent esiRNAs targeting C13orf3. For both esiRNAs, a significant cell-cycle arrest in the G2/M phase was observed 42 h post transfection. To ensure efficient knockdown of the intended target genes by RNAi, Q-PCR and western blot analyses were conducted. For all esiRNAs, a knockdown of at least 80 % was achieved at the mRNA level 24 h post transfection and a knockdown of at least 85 % at the protein level 42 h post transfection, showing the efficacy of the employed esiRNAs. As the DNA content for cells in G2 phase and mitosis is 4 N, it is not possible to distinguish these two phases by DNA content measurements. To distinguish between a G2- and M-phase arrest, C13orf3 was depleted by RNAi and, 42 h post transfection, the mitotic index was determined by staining for the phosphorylation of serine 10 of histone H3, a mitotic marker for chromatin condensation (Goto *et al.*, 1999). Immunofluorescence microscopy showed an increase in the mitotic index to 26.2 % upon C13orf3 depletion (3.6 % for mock control), of which 80.5 % of the cells were arrested in a metaphase-like state (36.1 % for mock control) (Figure 3.6A and C). These data suggest that C13orf3 is required for anaphase onset. In addition to the metaphase arrest, we observed a significant increase in cells with tripolar and tetrapolar spindles (19 % *versus* 2 % for mock-transfected cells) (Figure 3.6D and E), indicating that C13orf3 might also have a function in centrosome duplication or maintenance. To obtain a dynamic description of the phenotypic consequences upon C13orf3 depletion, time-lapse microscopy analyses was carried out in a cell line expressing a histone(H2B)–GFP fusion protein. These analyses showed an apparently normal chromosome congression and a proper establishment of the metaphase plate (Figure 3.6F, H). However, as early as 30 h post transfection, cells failed to maintain the metaphase plate, with individual chromosomes exiting from the aligned chromosomes (Figure 3.6F) causing a mitotic arrest, which ultimately led to cell death through caspase-dependent apoptosis. Closer analysis of the first mitosis after C13orf3 depletion in the histone(H2B)-GFP cell line showed that affected cells first form a straight metaphase plate, indicative of a bipolar spindle with two centrosomes. At later stages, the metaphase suddenly became kinked, indicating that the bipolar spindles had reverted to a tripolar spindle, likely through fragmentation of one centrosome. This observation confirms the results from the phospho-histone H3 (pS10) and anti-pericentrin immunofluorescence stains (Figure 3.6A, C, D, and E) and provides a possible explanation for the frequent appearance of tripolar and tetrapolar spindles. Staining of C13orf3-depleted cells with antibodies against the mitotic checkpoint proteins, Bub1 (Figure 3.6G) and Mad2 (data not shown), showed that the detached kinetochores were positive for both proteins, suggesting that the mitotic arrest is caused by the activation of the SAC. Other kinetochore proteins such as CASC5, Mis12, or NDC80 did not loose their localisation, showing that the overall structure of the kinetochores is not affected upon C13orf3 depletion. Co-depletion of Mad2 together with C13orf3 rescued the mitotic arrest (Figure 3.6B, C), showing that the activation of the SAC is indeed the primary cause of the mitotic arrest. Together, the comparison of the RNAi phenotype with the described Ska1 RNAi phenotype, that is, mitotic arrest in metaphase, SAC activation (Hanisch *et al.*, 2006), and mitotic centrosome fragmentation, substantiated a possible functional link between these two proteins.

Next, we wanted to investigate whether C13orf3 is required for chromosome segregation. Onset from metaphase to anaphase with chromosome segregation during anaphase requires, among others, the inhibition of the kinase activity of Cdk1 (Sullivan & Morgan, 2007). Furthermore, Cdk1 activity is necessary for mitotic entry and maintenance of the mitotic state in early mitosis (Vassilev *et al.*, 2006). Consequently, an inhibition of Cdk1 kinase activity during early mitosis, for example, by the small molecular inhibitor, RO-3306, results in mitotic exit(Vassilev *et al.*, 2006). In contrast, inhibition of Cdk1 activity in interphase prevents mitosis entry and induces an arrest in the G2 phase. Accordingly, mitotic arrest in prometaphase by nocodazole could be released by RO-3306 treatment, leading to mitotic exit without chromosome segregation. To investigate possible differences of RO-3306-induced mitotic exit in metaphase, cells were arrested through RNAi against the APC/C subunits Cdc16 and C13orf3. After treatment with RO-3306, the exiting cells were imaged by fluorescence time-lapse microscopy. Although Cdc16 depletion leads to a metaphase arrest with the formation of a metaphase plate and a bipolar spindle, the release by RO-3306 treatment does not result in chromosome segregation (Figure 3.6I), inter alia because sister chromatids are still held together at the centromeres by cohesin. In contrast, a release from the metaphase arrest after C13orf3 depletion by RO-3306 resulted in anaphase onset and chromosome segregation with high statistical significance (Figure 3.6I, J). Consequently, C13orf3 is not per se required for anaphase execution or chromosome segregation. To exclude effects caused by tagging with GFP and Cherry, we repeated these assays with unlabelled cells leading to the same conclusion.

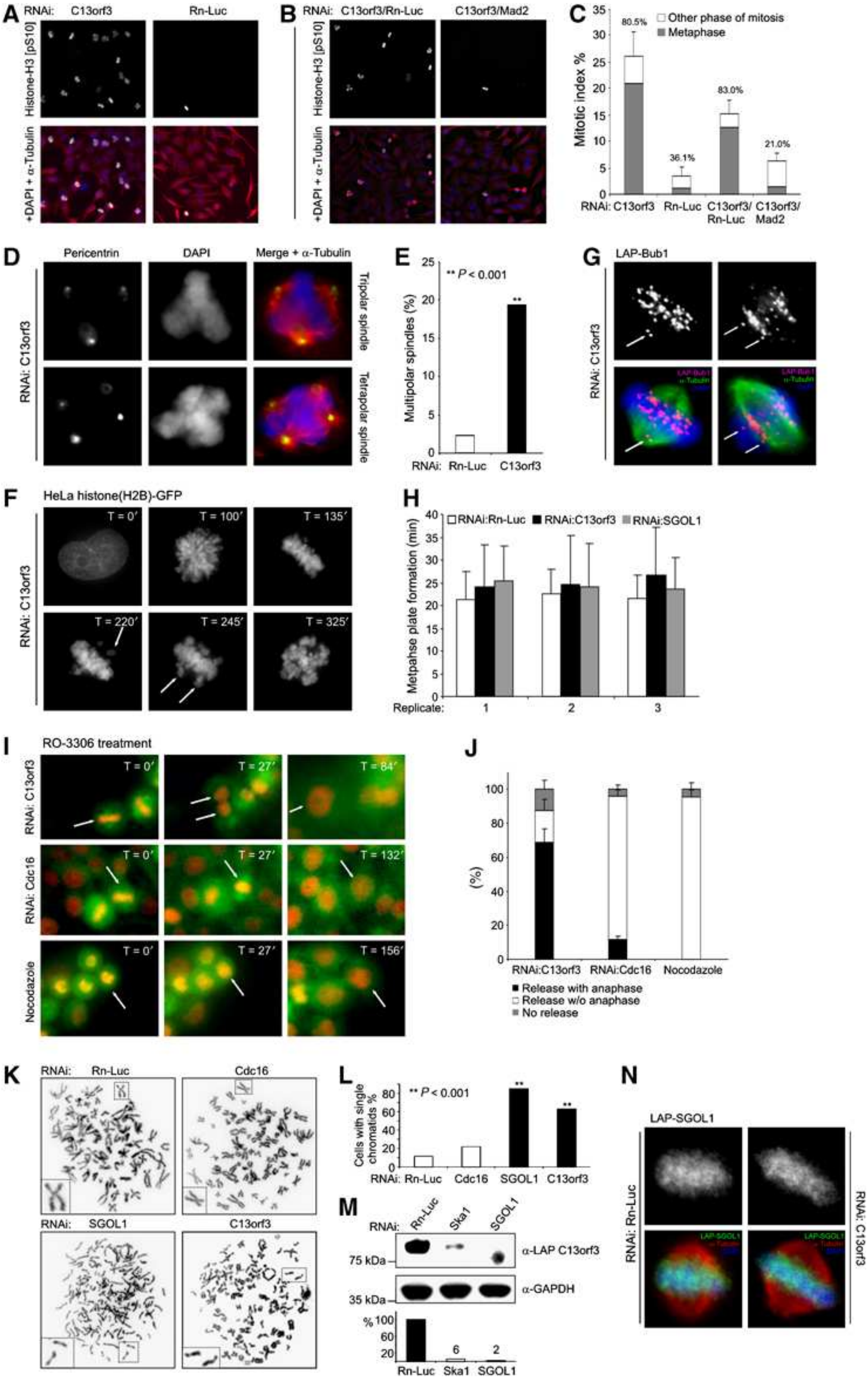


Figure 3.6: RNAi phenotypes upon C13orf3 depletion

Figure 3.6: cont.

(A) C13orf3-depleted cells arrest in metaphase. Anti-phospho-histone H3 (pS10) (green), DAPI (blue), and α-tubulin (red) stains are shown of HeLa cells treated with esiRNAs as indicated. Representative images of three independent experiments are shown. (B) Metaphase arrest upon C13orf3 depletion is dependent on spindle-assembly checkpoint integrity. Anti-phospho-histone H3 (pS10) (green), DAPI (blue), and α-tubulin (red) stains of HeLa cells treated with indicated mixtures of esiRNAs are shown. (C) C13orf3 depletion results in a SAC-dependent metaphase arrest with high statistical significance ($p < 0.001$). Quantitative evaluation of phospho-histone H3 (pS10) stains (as in panels A and B) are shown. The mitotic index is shown with the percentages of cells arrested in metaphase indicated above the bars. At least 200 cells were counted for each experiment. Error bars indicate the standard deviation of three independent experiments. Significance tests were carried out with differences calculated between mitotic and metaphase indices of cells treated with C13orf3 or C13orf3/Mad2 esiRNAs versus controls by a two-tailed t-test. (D) Cells arrested by C13orf3 depletion show an increased frequency of multipolar spindles compared with control cells. Representative cells stained with antibodies against α-tubulin (red), pericentrin (green), and DAPI (blue) are shown. (E) Depletion of C13orf3 increases the number of multipolar spindles with high statistical significance ($p < 0.001$). Quantitative evaluations of pericentrin stains (as in panel D) are shown. At least 80 metaphases were evaluated for each value. Significance tests were carried out with differences calculated between percentages of multipolar spindles of cells treated with C13orf3- versus Rn-Luc esiRNAs by a Pearson's χ^2 test with Yates' correction for continuity (**). (F) C13orf3 is necessary for metaphase plate maintenance. HeLa cells stably expressing histone(H2B)-GFP depleted of C13orf3 followed by fluorescence time-lapse microscopy for the indicated periods are shown. Arrows indicate unaligned chromosomes. Representative images from a total of 45 cells filmed by time-lapse microscopy are presented. (G) Immunofluorescence staining of LAP-Bub1 in mitotic HeLa cells after depletion of C13orf3 by RNAi. The single arrow (left panel) points to detached paired sister kinetochores, and the two arrows (right panel) point to separated sister kinetochores. Representative images are depicted from a total of 56 mitotic cells evaluated from two independent experiments. (H) C13orf3 depletion leaves the timing of metaphase plate formation unaltered. Measurement of the time from nuclear envelope breakdown to metaphase plate formation is shown. For every RNAi treatment, the average time and standard deviation of 15 cells are shown. The experiment was repeated three times independently. Average and standard deviations are given for every replicate separately. (I) Cdk1 inhibition promotes anaphase entry in arrested cells after C13orf3 depletion. HeLa cells stably expressing Cherry-histone(H2B) and GFP-tubulin arrested by RNAi against C13orf3, Cdc16, or nocodazole are shown ($T = 0$). Selected frames from time-lapse microscopy after the addition of the Cdk1 inhibitor, RO-3306, are presented ($T = 27\text{–}156$). Arrows track cells exiting from mitosis. (J) Statistical quantification shows significant differences in mitotic exit upon C13orf3-RNAi, Cdc16-RNAi, or nocodazole treatments. Cells arrested in mitosis after indicated treatments were analysed with respect to mitotic release by RO-3306 with progression to anaphase (black) or without anaphase (white). Arrested cells that could not be released by RO-3306 are shown in grey. Error bars indicate standard deviation for 200 evaluated cells from three independent experiments. (K) Sister chromatids are separated after mitotic arrest by C13orf3 or SGOL1 depletion. Representative chromosome preparations for HeLa cells treated with indicated esiRNAs are shown. Cells treated with the negative control Rn-Luc were arrested by nocodazole treatment before harvesting. RNAi against the APC/C component, Cdc16, was used as a control for metaphase-arrested cells with X-shaped chromosomes. Individual chromosomes are shown as blow-ups. (L) Quantitative evaluation of metaphase spreads shows significant increase in single chromatids upon treatment with esiRNA for SGOL1 and C13orf3 compared with Cdc16 or Rn-Luc RNAi ($p < 0.001$). For each treatment, 40–60 metaphases from two independent experiments were evaluated. Significance tests were carried out with differences calculated between the percentage of cells showing single chromatids treated with C13orf3, Cdc16, or SGOL1 and Rn-Luc esiRNA by a Pearson's χ^2 test with Yates' correction for continuity (**).

Figure 3.6: cont.

(M) C13orf3 protein stability is dependent on Ska1 and SGOL1. Western blot analysis of extracts from mitotic cells transfected with indicated esiRNAs and stained with indicated antibodies are shown. Size standards are depicted on the left. Quantifications of the band intensities by densitometry are shown below the western blot. Numbers indicate the protein levels in percent after RNAi treatment, normalised to GAPDH. The intensity for the mock control (i.e., Rn-Luc esiRNA) was used as reference. (N) SGOL1 protein stability and localisation is not dependent on C13orf3. Immunofluorescence stains of LAP-SGOL1 in mitotic HeLa cells after treatment with esiRNAs against C13orf3 or Rn-Luc are shown. Representative images are depicted from a total of 48 mitotic cells evaluated from two independent experiments.

Together, these data identify C13orf3 as an essential protein to satisfy the SAC but not for chromosome segregation during anaphase. Beside the inhibition of Cdk1 activity, the separation of sister chromatids is an important step before execution of anaphase (Sullivan & Morgan, 2007; Yamagishi *et al.*, 2008). Therefore, we asked whether cells arrested in metaphase after C13orf3 depletion have intact sister chromatid cohesion. The protein SGOL1 is implicated in the protection of centromeric sister chromatid cohesion during early mitosis mainly by recruiting PP2A to the centromeric region of chromosomes. Hence, knockdown of SGOL1 leads to mitotic arrest due to premature sister chromatid separation (Riedel *et al.*, 2006; Waizenegger *et al.*, 2000). Observation of chromosome spreads prepared from cells arrested by nocodazole treatment or RNAi against Cdc16, C13orf3, and SGOL1 revealed X-shaped chromosomes in the cases of Cdc16 RNAi and nocodazole treatment (Figure 3.6K). It indicates that the sister chromatid cohesion is still intact. In contrast, single sister chromatids were observed with high statistical significance for RNAi of C13orf3 and SGOL1 (Figure 3.6K and L), indicating that the phenotypic consequences of the depletion of these two proteins may be similar. Interestingly, on closer inspection of the RNAi phenotype of SGOL1, other striking similarities to the C13orf3 depletion were seen, for example, exiting chromosomes from the metaphase plate and centrosome fragmentation without alteration on timing of early mitosis (Nakajima *et al.*, 2007; Wang *et al.*, 2008b) (Figure 3.6H). Similar to the C13orf3 knockdown, the RNAi depletion of SGOL1 does not alter the localisation of kinetochore proteins such as CASC5, Mis12, or NDC80, showing that the overall kinetochore structure stays intact. To further investigate a potential connection of SGOL1 and Ska complex, Ska1 and SGOL1 were depleted in the C13orf3 BAC-tagged HeLa cell line and possible changes in C13orf3 protein levels were monitored by western blot analysis. Strikingly, both knockdowns greatly reduced protein levels of C13orf3 in mitotic cells (Figure 3.6M), corroborating the link between C13orf3, Ska1 and SGOL1 and indicating that C13orf3 requires Ska1 and SGOL1 for its stability. To test whether the loss of sister chromatid cohesion phenotype upon C13orf3 depletion (Figure 3.6K) might be due to loss of SGOL1 protein or mislocalisation, C13orf3 was depleted in a BAC transgenic HeLa cell line expressing LAP-tagged SGOL1. No significant loss in SGOL1 protein level or mislocalisation was observed upon C13orf3 depletion compared with the mock control (Figure 3.6N). These results suggest that there is no mutual dependency of C13orf3 and SGOL1 protein levels and places C13orf3 downstream of SGOL1.

3.3.4 C13orf3 is differentially phosphorylated during mitosis

Western blot analysis of lysates isolated from asynchronously growing LAP-tagged C13orf3 cells identified a single band of the predicted size. However, cell extracts prepared from mitotic cells showed an additional band of higher molecular weight (Figure 3.7A), suggesting that C13orf3 is modified in mitosis. Treatment of mitotic extracts with calf intestine phosphatase led to the disappearance of the slower migrating band (Figure 3.7A), showing that C13orf3 is phosphorylated during mitosis. Furthermore, the mass spectrometry analysis identified a C13orf3 peptide that was specifically phosphorylated at threonine 190 or 193 during mitosis (Figure 3.7B). To study phosphorylation of C13orf3 during mitosis in more detail, the protein Eg5 (kinesin 11, depletion produces monopoles and causes a prometaphase arrest) and the APC/C subunit Cdc27 (depletion of important APC/C subunits cause a metaphase arrest) were depleted by RNAi. Protein extracts isolated from these cells showed that C13orf3 is phosphorylated during prometaphase and metaphase (Figure 3.7C). To identify potential kinases implicated in C13orf3 phosphorylation, the selected kinases with prominent roles during mitosis were depleted by RNAi in the LAP-tagged C13orf3 BAC-transgenic HeLa cells and mitotic cell lysates were analysed by western blot. This analysis showed that depletion of AurkB, but not Plk1, abolished phosphorylation of C13orf3 (Figure 3.7D), indicating that C13orf3 phosphorylation is AurkB-dependent. Interestingly, in addition to the reduced protein levels, no band of higher molecular weight was visible in mitotic extracts upon Ska1 RNAi treatment (Figure 3.6M), indicating that C13orf3 phosphorylation was also dependent on Ska1. Given the differential phosphorylation of C13orf3 in interphase and mitosis, it appeared likely that C13orf3 is dephosphorylated by a protein phosphatase at the end of mitosis. To test this hypothesis, HeLa cells stably expressing LAP-tagged C13orf3 were released from nocodazole arrest and treated with okadaic acid, an inhibitor of the phosphatase activity of PP2A and PP1 (Mailhes *et al.*, 2003). The phosphorylation of C13orf3 persisted in the presence of okadaic acid (Figure 3.7E), indication that either PP2A or PP1 activity is required to remove C13orf3 phosphorylation at the end of mitosis. The conclusions is that C13orf3 is differentially phosphorylated during the cell cycle, with AurkB and PP2A or PP1 being potential candidates that phosphorylate and dephosphorylate the protein, respectively.

3.3.5 C13orf3 forms a complex with Ska1, Ska2 and PPP2R2B

To identify protein interaction partners of C13orf3, immunoprecipitation assays were performed and followed by mass spectrometry using the LAP-tagged C13orf3 HeLa cell line. These analyses showed interactions of C13orf3 with the Ska complex proteins Ska2 (Fam33A) and Ska1 (C18orf24) (Table 3.2). Mass spectrometry analyses utilising LAP-tagged BAC-transgenic Ska1 and Ska2 cell lines validated the physical interaction of these three proteins (Table 3.2). Hence, these studies identify C13orf3 as a new member of the Ska complex. Based on this data, we propose to rename C13orf3 into Ska3. In contrast to the Ska protein interactions, a direct interaction of C13orf3 with SGOL1 was not detected by mass spectrometry. However, a global proteomic study with subunits of PP2A showed a physical interaction of C13orf3 and Ska1/2 with PPP2R2B (Glatter *et al.*, 2009) (Table 3.2). The interaction with PPP2R2B indicates that PP2A is the phosphatase that dephosphorylates C13orf3 at the end of mitosis (Figure 3.7E) and also provides a possible link to SGOL1 through the regulation of PP2A activity. To map the interaction domains of Ska proteins, different protein fragments were tested in yeast two-hybrid assays (Boxem *et al.*, 2008). These analyses showed an interaction of the N-terminal part of C13orf3 (aa 1–159) with the N-terminal part of Ska1 (aa 1–84) (Figure 3.8A and B), as well as with Ska2 at the two terminal regions of C13orf3 (aa 1–87 and aa 303–412). In addition, these assays defined a minimal motif for binding of Ska1 and Ska2 at the N-terminal side of Ska1 (aa 1–63) (Figure 3.8A and B). Hence, all Ska complex proteins interact directly with each other (Figure 3.8B). Bioinformatic analysis (threading) of Ska1 predicts a three-helical bundle at the N-terminus that is structurally homologous to a Spectrin repeat-like fold (SCOP-ID: 46965), followed by a KEN box (Figure 3.8B), a short

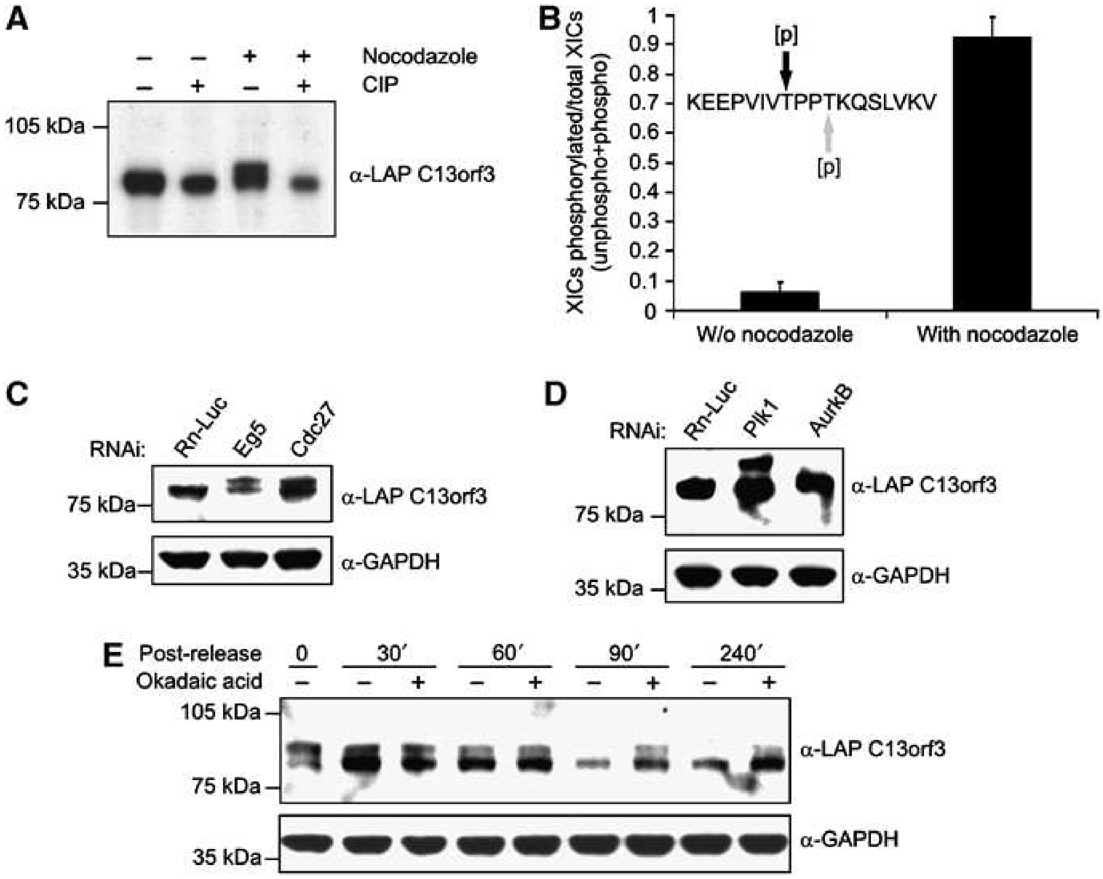


Figure 3.7: C13orf3 is phosphorylated during mitosis

(A) C13orf3 is phosphorylated in mitotic cells but not in interphase. Extracts from cells stably expressing LAP-tagged C13orf3 and treated with indicated reagents were analysed by western blotting with an anti-GFP antibody (CIP, calf intestine phosphatase). (B) Identification of a cell-cycle-dependent phosphorylation site in C13orf3. HeLa cells expressing LAP-C13orf3 were treated with nocodazole and harvested for analysis by mass spectrometry. Asynchronously growing cells served as reference. The monophosphorylated peptide detected and quantified by mass spectrometry is shown (either position T190 (black) or T193 (grey) is phosphorylated). Peak areas from the ion chromatogram for the phosphorylated and unphosphorylated peptides were used for given ratios. Error bars represent the standard deviation of three independent replicates. (C) C13orf3 is phosphorylated in different phases of mitosis. BAC-transgenic HeLa cells expressing LAP-C13orf3 arrested in prometa- (RNAi: Eg5) or metaphase (RNAi: Cdc27) analysed by western blotting are presented. (D) C13orf3 is not phosphorylated in AurkB-depleted cells. HeLa cells expressing LAP-C13orf3 treated with esiRNA against Plk1 or AurkB are shown. Mitotic cells were harvested by mechanical shake-off and analysed by western blot. (E) C13orf3 phosphorylation persists in the presence of okadaic acid. BAC-transgenic HeLa cells expressing LAP-C13orf3 released from nocodazole arrest and treated with okadaic acid for indicated time periods are shown. Cell lysates were analysed by western blotting with indicated antibodies. Markers are shown in the left of western blots.

motif found in many proteins involved in cell-cycle regulation and mitosis (Michael *et al.*, 2008). Three helical bundles are known to mediate protein–protein interactions (Fridmann-Sirkis *et al.*, 2006). Hence, the observed interaction of Ska1 with C13orf3 might be mediated through the three-helical bundle region. In summary, the two-hybrid assays validate the observed interaction of C13orf3 with Ska1 and Ska2, and define the minimal motifs for these interactions (Figure 3.8A and B).

Table 3.2: Results of pulldown assays and mass spectrometry

Bait	Interaction partner	Ensembl Gene Identifier	International Protein Identifier	Detected peptides:	
				unique	coverage (%)
C13orf3	**–**	**ENSG00000165480**	**IPI00333014**	**20**	**46**
C13orf3	Ska1	ENSG00000154839	IPI00059912	13	40
C13orf3	Ska2	ENSG00000182628	IPI00103149	10	66
Ska1	**–**	**ENSG00000154839**	**IPI00059912**	**13**	**41**
Ska1	C13orf3	ENSG00000165480	IPI00333014	14	39
Ska1	Ska2	ENSG00000182628	IPI00103149	9	58
Ska2	**–**	**ENSG00000182628**	**IPI00789882**	**4**	**58**
Ska2	C13orf3	ENSG00000165480	IPI00333014	10	29
Ska2	Ska1	ENSG00000154839	IPI00059912	6	21
PPP2R2B	**–**	**ENSG00000156475**	**IPI00020850**	**70**	**55**
PPP2R2B	C13orf3	ENSG00000165480	IPI00333014	10	28
PPP2R2B	Ska1	ENSG00000154839	IPI00059912	2	20
PPP2R2B	Ska2	ENSG00000182628	IPI00103149	4	29

Baits are shown in bold characters.

In conclusion, the data on C13orf3 show that the combination of phenotypic profiling with protein localisation data is a useful approach to predict functions of uncharacterised genes. Large-scale tagging of proteins at endogenous expression levels is possible in yeast, and comprehensive protein localisation (Huh *et al.*, 2003) and protein interaction network studies (Gavin *et al.*, 2006; Krogan *et al.*, 2006) have been carried out in this organism. To broaden this approach, systematic BAC tagging (Poser *et al.*, 2008) of most proteins would allow similar studies in mammalian cells. In a test study, applying this technology in small scale, we identified C13orf3 as a new interaction partner of the Ska complex. This study provides a first link between the Ska complex and regulation of sister chromatid cohesion possibly through SGOL1 and PP2A pathways.

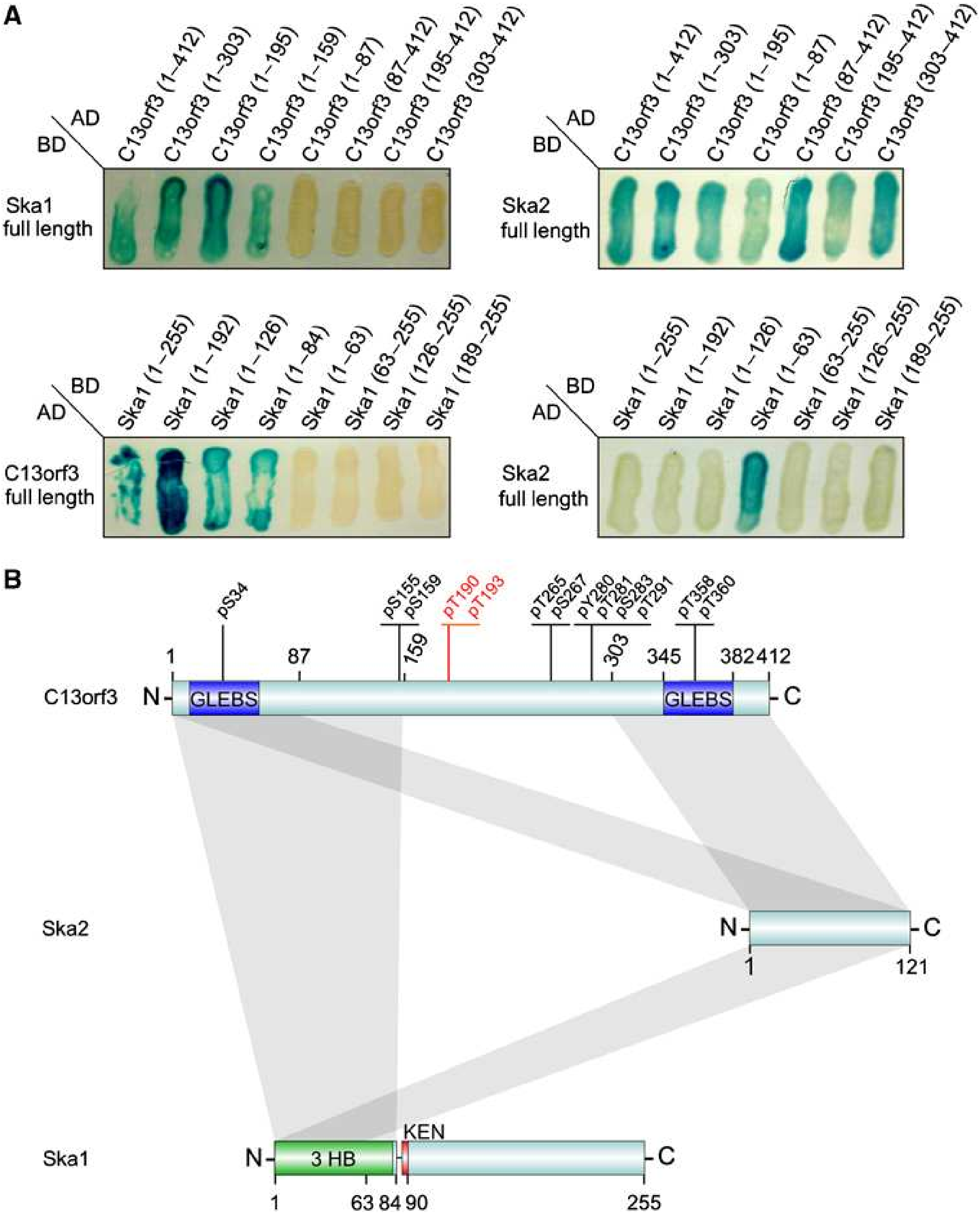


Figure 3.8: Identification of minimal binding motifs and model of interaction

Figure 3.8: cont.

(A) Yeast two-hybrid analysis of Ska complex protein interactions. Ska proteins were shortened from both termini (fragment sizes are given as numbers of amino acids) and combined for yeast two-hybrid analyses (AD, activation domain; BD, binding domain). (B) Model for Ska complex interactions. Minimal binding motifs derived from panel A are indicated as grey bars. Phosphorylated residues are indicated (Brill *et al.*, 2004; Cantin *et al.*, 2008; Nousiainen *et al.*, 2006; Rush *et al.*, 2005). Newly identified phosphorylation at threonine 190 or threonine 193 are highlighted in red. Predicted GLEBS motifs, three-helical bundle (3 HB), or KEN-box motifs are shown in blue, green, and red, respectively. Numbers indicate the amino acid positions.

3.4 Methods

3.4.1 Genome-wide cell-viability RNAi screen

EsiRNAs were synthesised as described previously (Kittler *et al.*, 2005a, 2007c) and, after normalisation, arrayed into 384-well plates. Briefly, esiRNAs are silencing triggers for RNAi in mammalian cells prepared by enzymatic digestion (bacterial RNase III) of long dsRNA (300 bp–600 bp, derived from target mRNA). esiRNAs were chosen in this study in favour of chemically synthesised siRNAs because they were shown to produce less off-target effects (Kittler *et al.*, 2007c). All esiRNAs used in this study were designed to target all splicing variants of their target genes, respectively. For the genome-scale viability screen, esiRNAs (15 ng each) were reversely transfected into HeLa cells with Oligofectamine (Invitrogen) in black, tissue culture plates (Greiner) and incubated for 72 h. For viability analysis, the cell culture medium was supplemented with AlamarBlue dye (Serotec) and after 3 h incubation the fluorescence intensities (excitation: 535 nm; emission: 590 nm) were measured using a plate reader (GENiosPro, Tecan).

3.4.2 BAC TransgeneOmics

BACs harbouring the genes of interest were obtained from the BACPAC Resource Center (`http://bacpac.chori.org`). A LAP cassette (Cheeseman & Desai, 2005) was inserted as a C-terminal fusion using recombineering (Zhang *et al.*, 2000) (Gene Bridges). Isolated BAC DNA was transfected and selected for stable integration as described (Poser *et al.*, 2008).

3.4.3 Immunofluorescence microscopy

Cells were grown on coverslips in 12-well plates, transfected with 300 ng esiRNAs, fixed in cold methanol at −20 °C for 8 min, and blocked with 0.2 % gelatin from cold-water fish skin (Sigma) in phosphate buffered saline (PBS) (PBS/FSG) for 10 min, 36 h–48 h post transfection for Ska complex esiRNAs and SGOL1 esiRNA. Staining was carried out by incubation with the following primary antibodies for 20 min in PBS/FSG: goat anti-GFP (1:4000, MPI-CBG Antibody Facility), mouse anti-α-tubulin (1:2000, MPI-CBG Antibody Facility), human anti-CREST (1:500, Cortex Biochem), rabbit anti-pericentrin (1:5000, Abcam). After washes with PBS/FSG, the cells were incubated with fluorescently labelled secondary antibodies (donkey anti-mouse Alexa594, Molecular Probes; donkey anti-goat FITC, Molecular Probes; donkey anti-rabbit Alexa594, Molecular Probes; and goat anti-human Alexa594, Invitrogen). After washing with PBS/FSG, the coverslips were mounted on glass slides by inverting them in the mounting solution containing 4',6-diamidino-2-phenylindole (DAPI, ProLong Gold antifade, Invitrogen). Images were taken on an Axioplan II Microscope (Zeiss) operated by MetaMorph (Molecular Devices) or on Olympus IX70 (Olympus) equipped with the imaging system DeltaVision RT using x40/1.00 or x63/1.40 Plan-Apochromat oil immersion objectives. Z-stacks (0.2 µm optical sections) were collected, deconvolved, and projected into one picture using softWoRx software (Applied Precision). Acquired images were cropped and contrast adjusted in Adobe Photoshop 8.0 (Adobe Systems) and then sized and placed in figures using Corel Draw 11.633 (Corel Corporation).

3.4.4 Live-cell imaging

HeLa cells stably expressing LAP-tagged proteins or histone(H2B)-GFP or Cherry-histone(H2B) and GFP-tubulin were grown in 96-well tissue culture plates and transfected with 40 ng esiRNAs. Images were obtained 12 h–36 h after transfection with a ScanR system (Olympus) placed in a heated chamber (37 °C) with 5 % CO_2 and filmed for 1 h–36 h as indicated. If appropriate, cells were arrested by treatment with 50 ng/ml nocodazole (Sigma) for 12 h and/or treated with 9 µM of RO-3306 (Merck Biosciences). For high-resolution time-lapse imaging, the cells were grown on eight-well LabTek chambered cover glasses (Nalge Nunc). Before imaging, the medium was changed to CO_2-independent medium (Invitrogen), and the cell culture chamber was placed onto a heated sample stage within a heated chamber (37 °C). Images were acquired with an Olympus IX70 DeltaVision RT system (Olympus).

3.4.5 Cell-based assays

For cellular DNA content analysis, esiRNA-transfected cells were fixed and stained with propidium iodide (Molecular Probes) and scanned with a FACSCalibur flow cytometer (BD Biosciences) 42 h post transfection. The resulting DNA content histograms were manually gated to determine the percentage of cells in G2/M phase. For determining the mitotic index, esiRNA-transfected cells were fixed and incubated with mouse anti-α-tubulin (1:2000; MPI-CBG Antibody Facility) and rabbit anti-phospho-histone H3 Ser10 antibodies (1:10, conjugated to Alexa488, Cell Signaling Technologies) 42 h post transfection. Subsequently, the cells were incubated with fluorescently labelled donkey anti-mouse Alexa594 antibody and DAPI; images were obtained as described above. For chromosome preparations, HeLa cells were treated with esiRNA (42 h: C13orf3 and SGOL1; 96 h: Cdc16) and/or nocodazole (50 ng/ml for 14 h). After harvesting, the cells were resuspended in hypertonic solution (30 mM sodium citrate) and incubated for 35 min–45 min at 37 °C. Subsequently, the cells were fixed with ethanol/acetic acid (3:1), spread onto a coverslip, rehydrated for 15 min with PBS and fixed again with formaldehyde (4 %). After washing for three times with PBS, the cells were dehydrated by washing with stepwise increasing concentrations of ethanol (70 %–100 % in four steps). After drying, the coverslips were mounted on glass slides by inverting them in the mounting solution containing DAPI (ProLong Gold antifade, Invitrogen). Images were taken on an Axioplan II Microscope (Zeiss) as described above.

3.4.6 Quantitative PCR and apoptosis assays

To ensure an efficient silencing for all prominent esiRNAs used in this study, we conducted mRNA quantification by quantitative-PCR (Q-PCR). HeLa cells were transfected in 12-well cell culture dishes using 300 ng esiRNA and 4.2 µl Oligofectamine (Invitrogen) per well. After 24 h incubation, the cells were harvested and total mRNA was extracted using the RNeasy Mini Kit (Qiagen) including an on-column DNaseI digest as given in the manufacturer's manual. Subsequently, total mRNA was reverse transcribed using SuperScript III reverse transcriptase (Invitrogen) and oligo (dT)12–18 primers (Invitrogen). Quantification of the targeted mRNA was conducted using the Absolute qPCR SYBR Green Kit and an Mx3000p (Stratagene) real-time PCR machine. For apoptosis assays, HeLa cells were transfected with esiRNA (30 ng) in 96-well cell culture dishes. After 24 h incubation, the caspase inhibitor, z-VAD-FMK (Merck Biosciences), was added to the cell culture supernatant (50 nM) or to 1 % DMSO as vehicle control. The cells were harvested and stained 48 h after transfection with fluorescently labelled Annexin-V (APC-conjugated, Becton Dickinson) and propidium iodide (Molecular Probes) and analysed by flow cytometry on a FACSCalibur system (BD Biosciences). Cells that stained positive for Annexin-V but negative for propidium iodide were considered apoptotic.

3.4.7 Western blot analysis

Whole-cell lysates or mitotic cells (mechanical shake-off) stably transfected with different BAC constructs and treated with esiRNAs, nocodazole (50 ng/ml, Sigma), or okadaic acid (250 nM; Sigma) were subjected to SDS–PAGE (NuPage 4 %–12 % Bis-Tris; Invitrogen), blotted to nitrocellulose (Protran, Schleicher & Schuell) and incubated with primary antibodies (mouse anti-GFP, 1:5000, Roche; or mouse anti-GAPDH, 1:20 000, Acris Antibodies). Subsequently, the membranes were incubated with goat anti-mouse antibody conjugated to horseradish peroxidase (1:4000, Bio-Rad); bands were visualised with enhanced chemiluminescence Western Blotting Detection Reagents (GE Healthcare). As a molecular weight standard, the Full-Range Rainbow ladder 10 kDa–250 kDa (GE Healthcare) was used. Films were scanned and images were cropped and contrast adjusted in Adobe Photoshop 8.0 (Adobe Systems) and then sized and placed in figures using Corel Draw 11.633 (Corel Corporation). For phosphatase assays, nocodazole-arrested cells were harvested by mechanical shake-off, and lysates were incubated with calf intestinal phosphatase (New England Biolabs) for 15 min at 37 °C or left untreated. Subsequently, all lysates were analysed by western blotting as described above. All band intensities of western blot images were quantified with ImageJ 1.40 g (National Institutes of Health).

3.4.8 Immunoprecipitation and mass spectrometry

Transgenic cells expressing LAP-tagged or Strep-HA-tagged versions of the proteins of interest were harvested and, after lysis, cleared from insoluble material by ultracentrifugation (100 000 g for 20 min at 2 °C). Immunoprecipitation was carried out by incubation with goat anti-GFP antibody (MPI-CBG Antibody Facility, 1 h at 4 °C) immobilised on G-protein sepharose (FastFlow, GE Healthcare, 200 µg antibody per 100 µl matrix) or on 200 µl Strep-Tactin beads (IBA TAGnologies). Specificity of the goat anti-GFP antibody in immunoprecipitation assays was extensively validated (Poser *et al.*, 2008). After washing, elution from the affinity beads was carried out with 100 µl of glycine (100 mM, pH 2.0), which was subsequently neutralised with 1.5 M Tris at pH 8.0 or 100 mM NH_4HCO_3. Strep-HA-tagged proteins were eluted with TNN-HS buffer with 2 mM biotin and immunoprecipitated with 100 µl anti-HA agarose (Sigma) before glycine elution. Modified porcine trypsin (Promega) was added (16 ng/µl) and proteins were digested overnight. The tryptic peptides were analysed by mass spectrometry.

3.4.9 Quantification of phosphorylation

Peak areas of extracted ion chromatograms (XICs) corresponding phosphorylated and non-phosphorylated precursors at 2+/3+ charge states were determined using Xcalibur 2.0 software (Thermo Fisher Scientific), assuming better than 10 ppm mass accuracy and <1 min retention time variation within multiple runs. Survey spectra were examined to make sure no peaks of the same mass were co-eluted with the quantified peptides. Phosphorylation was calculated as a ratio of the peak areas of phosphorylated precursors to the sum of peak areas of phosphorylated and non-phosphorylated forms and averaged between four repetitive runs for arrested cells and duplicated samples for non-arrested cells. When calculating phosphorylation of the peptide, the peak areas of the partially miscleaved form (both in phosphorylated and in non-phosphorylated states) were considered.

3.4.10 Yeast two-hybrid analysis

Yeast two-hybrid analysis was carried out using a system described previously (Vidal *et al.*, 1996). Full-length proteins and fragments of Ska1, Ska2, and C13orf3 were fused to

Gal4 DNA-binding domain (BD, aa 1–147) or Gal4 activation domain (AD, aa 768–881) as indicated in Figure 3.8A. Clones growing on media lacking uracil were streaked out on selective media lacking histidine and containing 50 mM 3-amino-1,2,4-triaziole (MP Biomedicals). Positive clones were further analysed for β-galactosidase activity.

3.4.11 Sequence analysis and comparative modelling

Sequence analyses were carried out with the ELM server (Puntervoll *et al.*, 2003). Secondary structure predictions were carried out with Jpred (Cole *et al.*, 2008). Threading analysis of the Ska protein sequences was carried out with the program Prohit (Proceryon GmbH) and a fold library containing 20 008 chains from the Brookhaven Protein Data Bank (PDB). GO terms according to the observed RNAi phenotype were used to distinguish possible true hits from false positives in the fold hit list. The N-terminal region of Ska1 was modelled as a Spectrin repeat-like fold using the sequence–structure alignment obtained from threading and the X-ray structure of SNARE Tlg1 at 2.05 Å resolution as template (PDBId 2c5k, chain T; Fridmann-Sirkis *et al.*, 2006). The C-terminal region of Ska3 was modelled as a GLEBS motif using the X-ray structure of the yeast Bub1-GLEBS/Bub3 at 1.9 Å resolution as template (PDBId 2i3s, chain B; Larsen *et al.*, 2007). The human Bub3 was modelled by homology based on the chain A of the yeast complex X-ray structure. The Discovery Studio package (v1.7, Accelrys, San Diego, CA) was used for model construction and refinement. The RosettaDock Server (Lyskov & Gray, 2008) was used to dock the modelled human Ska3-GLEBS and Bub3 structures, and the complex with the highest score was selected after visual inspection.

Chapter 4

A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity

Pluripotent embryonic stem cells (ESCs) maintain self-renewal while ensuring a rapid response to differentiation cues. The identification of genes maintaining ESC identity is important to develop these cells for their potential therapeutic use. This chapter reports a genome-scale RNAi screen for a global survey of genes affecting ESC identity via alteration of Oct4 expression. Factors with the strongest effect on Oct4 expression include components of the Paf1 complex, a protein complex associated with RNA polymerase II. Using a combination of proteomics, expression profiling, and chromatin immunoprecipitation, it was demonstrated that the Paf1C binds to promoters of key pluripotency genes, where it is required to maintain a transcriptionally active chromatin structure. The Paf1C is developmentally regulated and blocks ESC differentiation upon overexpression, and the knockdown in ESCs causes expression changes similar to Oct4 or Nanog depletions. As an outcome of this study, it was proposed that the Paf1C plays an important role in maintaining ESC identity.

My main contribution to this work is described in section 4.2, which contains a novel methodology and results of combined analysis of phenotypic and gene expression data. Additional gene expression and transcription factor binding analyses, which I performed as crucial parts of the follow-up studies, are presented in subsections 4.3.3, 4.3.5 and 4.3.7 of this chapter.

4.1 Introduction

Embryonic stem cells (ESCs) have unlimited capacity for self-renewal and can be kept undifferentiated for many passages under appropriate conditions while maintaining the competence to generate a wide range of cell types upon differentiation (Chambers & Smith, 2004). Because of these distinctive properties, ESCs are widely used for studies of developmental processes (O'Shea, 2004). The potential to differentiate into many cell types also makes ESCs a starting point for potential cell-based therapies (Keller, 2005). However, a systematic molecular understanding of self-renewal and differentiation is required to harness the full potential of ESCs.

Several transcription factors that contribute to the regulation of self-renewal and differentiation of ESCs have been identified. Oct4, Nanog, and Sox2 form a transcriptional core unit upon which ESC pluripotency is critically dependent (Silva & Smith, 2008). Recently, an RNAi analysis of 70 candidate transcription factors identified roles for additional genes, including Tbx3, Esrrb, Tcl1, and Dppa4, in the maintenance of ESC pluripotency (Ivanova *et al.*, 2006). Depletion of these genes negatively affected self-renewal, and induced ESC differentiation.

Several signaling pathways including the LIF/Stat3, PI3K, Wnt, and Bmp/Smad pathways also contribute to the regulation of self-renewal and differentiation of ESCs. For example, the PI3K pathway regulates multiple cascades including Ras/MAPK and mTOR pathways, which are essential for proliferation of mouse ESCs (Takahashi *et al.*, 2005).

In addition to transcription factors and signaling pathways, a defined epigenetic state has been shown to be essential to maintain ESC identity (Pietersen & van Lohuizen, 2008). Accordingly, many promoters of key developmental genes including Sox, Hox, Pax, and Pou gene family members display both an activating (H3K4me) and a repressive (H3K27me) histone mark on the nucleosomes at their promoters in ESCs (Bernstein *et al.*, 2006). This 'bivalent' histone code silences lineage-control gene expression due to the dominant effect of H3K27me over H3K4me, while preserving their potential to be rapidly activated upon differentiation stimuli via the removal of the H3K27me mark. It is therefore not surprising that proteins, which regulate chromatin structure, are important for ESC identity, as recently shown by a focused RNAi screen of not similar 1000 chromatin proteins (Fazzio *et al.*, 2008).

The work performed on transcription factors, signaling, and chromatin has substantially improved our understanding of ESCs. However, our knowledge of how these processes are connected is still limited. A global survey of genes essential for ESC self-renewal and identity would not only advance our understanding of this fundamental biological process but should also help to develop better protocols for directed differentiation of ESCs for their potential therapeutic use.

To obtain a more systematic understanding of the genes associated with ESC identity, a genome-scale RNAi screen in mouse ESCs was performed by using an Oct4 reporter assay as a surrogate for ESC identity. It identified many novel genes that affected Oct4 expression and therefore ESC identity and one group of genes, which form the Paf1 complex (Paf1C), was analyzed in more detail. This work therefore expands the inventory of genes required to maintain ESC identity and defines a role of the Paf1C in maintaining active transcription of pluripotency genes in ESCs.

4.2 Analysis of the RNAi screen for Oct4 modulators

4.2.1 Primary screen

To perform an RNAi screen in mouse ESCs, a genome-scale mouse endoribonuclease prepared (e)siRNA library containing 25 057 esiRNAs was generated using the established esiRNA synthesis protocol (Kittler *et al.*, 2005a). EsiRNAs have proven efficacy and specificity in human cells (Kittler *et al.*, 2004, 2007b,c), suggesting that this resource should be useful for RNAi screening in mouse cells. To identify genes essential for the maintenance of ESC identity, a rapid and robust assay amenable to high-throughput detection of differentiation triggered by RNAi was developed (Figure 4.1A). Oct4 expression is a hallmark of ESC identity, and constant expression levels of Oct4 are required for self-renewal of ESCs (Niwa *et al.*, 2000). Depletion of Oct4 in ESCs via knockout or RNAi leads to exit from self-renewal and to differentiation (Carpenter & Zernicka-Goetz, 2004; Nichols *et al.*, 1998). In turn, the expression of Oct4 is rapidly switched off in differentiating cells during embryonic development (Pesce & Schöler, 2001). Hence, Oct4 expression levels can be used to monitor the differentiation status of ESCs.

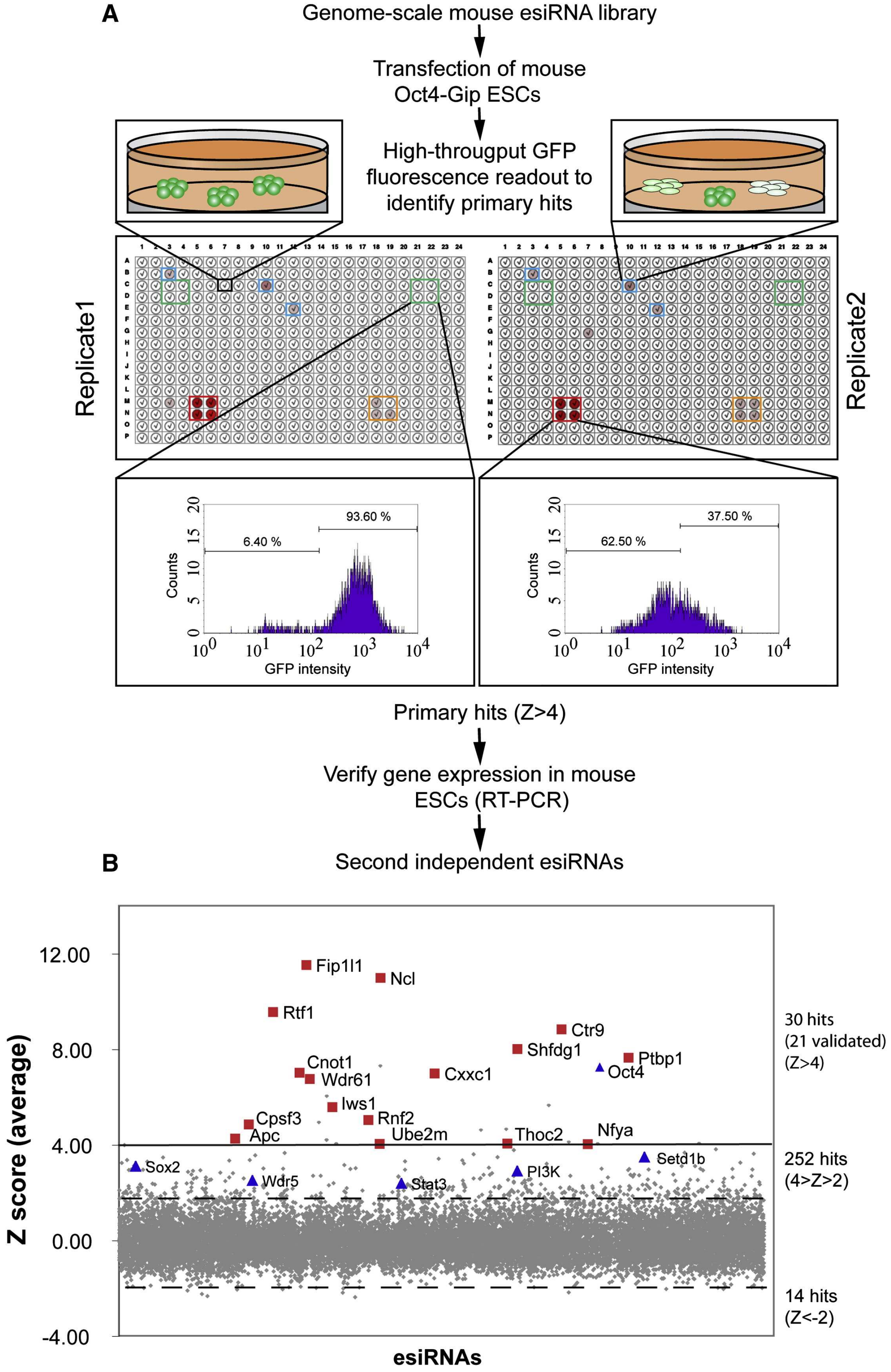


Figure 4.1: Genome-Scale esiRNA Screen

Figure 4.1: cont.

(A) Flow diagram of the screening strategy to identify genes essential for mouse ESC identity. Eight negative controls (boxed in green), eight positive controls (four GFP esiRNA, red; and four Sox2 esiRNA, orange), three primary hits (blue), and one random well without a phenotypic consequence (black) are highlighted. The red color intensity reflects the strength of phenotypes observed in each well, which is illustrated by FACS readouts for two exemplified wells (boxes below the plates). (B) Dot plot of the primary screen results. The average z-scores of the GFP readouts are shown. The dotted lines indicate z-score > 2 or < -2. The solid line marks z-scores > 4. Validated genes with z-scores > 3 with a second, independent esiRNA are shown as red squares. Selected pluripotency genes that scored with a z-score > 2 are shown as blue triangles.

An Oct4 reporter cell line (Oct4-Gip) was used in which the expression of GFP is controlled by Oct4 regulatory elements to establish an assay for analyses of ESC identity (Ying *et al.*, 2002). Quantification of GFP fluorescence faithfully reflects the self-renewal and differentiation status in individual cells and was thus used as a rapid and accurate readout to identify genes required for ESC identity (Figure 4.1A). To test the reliability of this assay, we transfected the Oct4-Gip cells with control esiRNA (Luc) and esiRNAs directed against known pluripotency genes, including Sox2, Stat3, and Oct4. The ratio of GFP-positive *versus* GFP-negative cells was determined by microscopy and by flow cytometry 96 h posttransfection. In both cases, a loss of GFP expression was readily detected upon Sox2, Stat3, and Oct4 depletion, which coincided with differentiation of these cells. In contrast, no loss of GFP expression or differentiation was observed when the cells were transfected with the control esiRNA. To rule out the possibility that essential genes that are not required for ESC self-renewal, but are required for general cell growth and viability, would score in this assay, the Oct4-Gip cells were transfected with esiRNAs targeting genes with housekeeping functions, such as ribosomal, proteasomal, mitochondrial, and Pol II subunits. As expected, depletion of housekeeping genes strongly affected cell viability. However, none of these esiRNAs caused a significant loss of GFP-positive cells, demonstrating that this assay is highly specific to identify genes affecting ESC identity.

The screen was carried out in duplicate using a high-throughput FACS-based readout. I nominated 296 esiRNAs, which significantly regulated GFP expression (z-score $>$ 2, or z-score < -2) from the primary screen (Figure 4.1B). GO term analysis indicated that transcription factors (25 genes) and gene expression regulators (51 genes) were significantly enriched in the primary hit list ($p < 0.05$), suggesting that this subset of genes can be a rich starting point to dissect the regulation of self-renewal and early differentiation of mouse ESCs.

4.2.2 The screen identified known and discovered novel candidate pluripotency genes

Many known pluripotency-controlling genes including Oct4, Sox2, Stat3, PI3K, Set1b, and Wdr5 scored with significant z-scores, demonstrating the effectiveness of the screen. The screen also identified many additional genes that have been linked to ESC biology, such as direct Oct4 and Nanog target genes (Cxxc1, Vti1a, Nfix, Etv1, Rad21, Ina, Cdh4, Spred2, Bach2, Myh9, Map3k7, Tcf12) (Fouse *et al.*, 2008) and genes that are regulated by Oct4 and Nanog (Rnf2, Ncl, Ina, Spred2, Thbs3, Ell2, Tmem4, Etv1, Foxp1) (Loh *et al.*, 2006). Some known pluripotency genes, including Esrrb, Tbx3, and Klf4, did not score significantly in the primary screen, possibly reflecting insufficient knockdown, redundancy, or that their roles in ESC identity are not reflected in Oct4 expression levels.

Further study concentrated on the genes with the strongest phenotypic scores (z-score > 4) and validated the phenotype of these candidates with a second, independent esiRNA. Twenty-nine independent esiRNAs were successfully synthesized and transfected into the Oct4-Gip ESCs. Twenty-one of these caused a reduction of GFP expression greater than two times the standard deviation, validating their role in maintaining Oct4-driven GFP expression. Sixteen hits with z-scores above three were further analyzed. Gene ontology analyses placed these genes into different functional classes, including transcription regulation (Nfya, Ptbp1, Ctr9, Rtf1, Wdr61, Cpsf3, Fip1l1, Iws1, Thoc2), chromatin modulation (Rnf2, Cxxc1, Cnot1, Rtf1, Ctr9, Wdr61, Ncl), signaling (Apc), and protein degradation/DNA repair (Ube2m, Shfdg1).

4.2.3 Meta-analysis of phenotypic and gene expression data refines screening results

Because perturbations of many pathways in ESC lead to differentiation, many hits from the RNAi screen may not be directly involved in pluripotency maintenance. To refine results of the RNAi screen, I utilized a microarray data that measured time-course gene expression during the first 7 days of ESC differentiation into embryonic bodies. Genes, whose silencing leads to decreased levels of Oct4 and are being downregulated during the phases of early development, are more likely to be implicated in pluripotency maintenance.

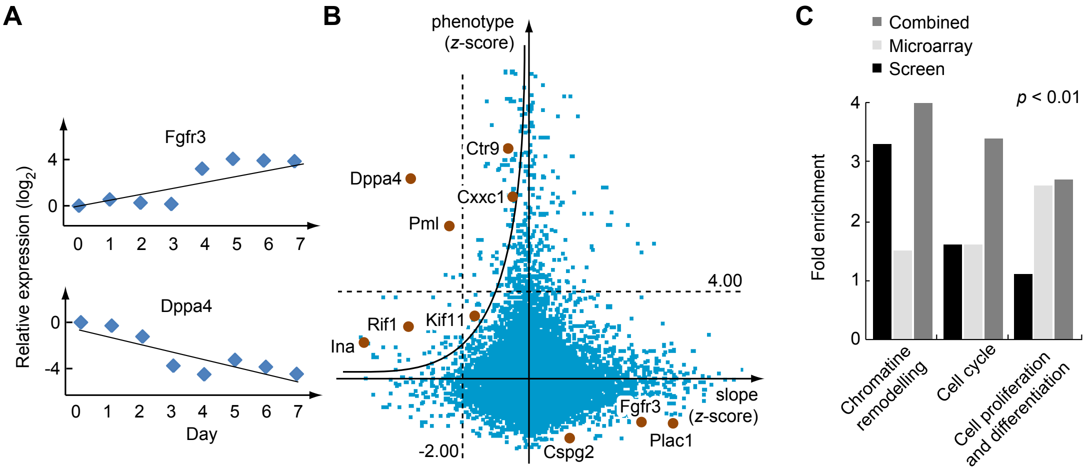


Figure 4.2: Combined analysis of RNAi and time-course microarray data

(A) Changes of expression of two selected genes during the process of differentiation and embryonic body formation. Expression levels are shown in a logarithmic scale, relative to the expression level in day 0. Slopes of the linear regression (solid lines) were used to quantify the change of expression over time. Fibroblast growth factor receptor 3 (Fgfr3) is a membrane protein that interacts with growth hormones and activates signalling cascades responsible for mitogenesis and differentiation. Its expression is switched on during early development (Keegan *et al.*, 1991). Developmental pluripotency associated protein 4 (Dppa4) inhibits ESC differentiation into ectoderm and its expression has to be reduced to continue development (Masaki *et al.*, 2007). (B) Graphic representation of z-scores from the Oct4 modulators screen (y-axis) plotted against z-score normalized changes of expression during ESC differentiation (x-axis). Dashed lines indicate hit selection thresholds applied to results of individual experiments. The solid line curve represents a function used for selecting hits based on data from both experiments (see Equation 4.1). Selected pluripotency maintenance (quadrant I) and developmental genes (quadrant IV) are shown as red dots. (C) Gene set enrichment analysis of hits from the Oct4 modulators screen, genes being significantly downregulated during ESC differentiation and genes selected from the combined analysis of the two experiments. Resuls of enrichment of selected pluripotency- and differentiation-related terms show an increased sensitivity of the combined selection compared to selection based on individual experiments.

To evaluate temporal changes of gene expression, I calculated a linear regression for each time-course expression profile. The slope coefficient of the regression was used as a measure of expression change over time (Figure 4.2A). To make possible the comparison with the phenotypic data, I converted regression coefficients into z-scores, taking mean and standard deviations of complete data set as basis for the normalization. I formulated a new statistic that combines the results of both experiments as a function that multiplies respective z-scores:

$$z_{\text{Combined}} = z_{\text{Expression}} \cdot z_{\text{Oct4}} \tag{4.1}$$

To select genes putatively involved in pluripotency maintenance, I applied the following selection criteria (Figure 4.2B):

$$z_{\text{Oct4}} > 0 \tag{4.2}$$

$$z_{\text{Expression}} < 0 \tag{4.3}$$

$$z_{\text{Combined}} > 4 \tag{4.4}$$

The threshold of 4 was selected arbitrarily, based on visual inspection of the plot in Figure 4.2B.

A list of hits obtained from the meta-analysis of phenotypic and gene expression data shows an increased enrichment of annotation terms related to stem cell maintenance, compared to lists of genes obtained from the individual analysis of the two experiments (Figure 4.2C). In the list obtained by applying criteria 4.3–4.4, I found developmental repressors and transcription factors that were not identified by the primary RNAi screen as well as few genes of unknown function (Table 4.1).

Results of our meta-analysis opened new directions for further research. By altering the threshold criteria and selecting genes having their expression increased during differentiation and whose silencing leads to upregulation of Oct4, I obtained a list of genes containing developmental markers (*e.g.* Fgfr3, Cspg2, Plac1). Further analysis of this data set may lead to discovery of novel genes that could be targets for cellular fate reprogramming.

Table 4.1: Genes nominated as 'hits' by the combined analysis but not detected by the primary RNAi screen.

Ensembl Gene ID	Gene	z_{Oct4}	$z_{\text{Microarray}}$	Annotation
ENSMUSG00000002980	Bcam	3.65	−1.52	Cell adhesion
ENSMUSG00000028655	Mfsd2	3.21	−1.60	
ENSMUSG00000012443	Kif11	2.96	−1.85	Mitotic centrosome separation
ENSMUSG00000041846	1110034C04Rik	2.90	−1.60	
ENSMUSG00000028047	Thbs3	2.74	−1.52	Cell adhesion
ENSMUSG00000019773	Fbxo5	2.58	−2.19	Regulation of mitotic cell cycle
ENSMUSG00000026917	Wdr5	2.50	−2.70	Regulation of transcription, development
ENSMUSG00000018740	Slc25a35	2.41	−3.63	Transmembrane transport
ENSMUSG00000024287	Thoc1	2.38	−1.85	Regulation of apoptosis, RNA splicing
ENSMUSG00000021532	Fastkd3	2.22	−1.94	Apoptosis
ENSMUSG00000036202	Rif1	2.20	−5.23	Stem cell maintenance
ENSMUSG00000022208	Jph4	2.17	−2.28	
ENSMUSG00000021906	Oxnad1	1.99	−2.19	Oxidation reduction
ENSMUSG00000027559	Car3	1.97	−3.04	Response to oxidative stress
ENSMUSG00000024376	Epb4.1l4a	1.90	−3.04	
ENSMUSG00000019590	Cyb561	1.87	−3.96	Electron transport chain
ENSMUSG00000024078	Ttc27	1.86	−2.61	
ENSMUSG00000044224	4930461P20Rik	1.83	−3.37	
ENSMUSG00000021018	Polr2h	1.82	−2.53	Transcription
ENSMUSG00000040370	4930469P12Rik	1.80	−2.70	
ENSMUSG00000019977	Hbs1l	1.73	−2.87	Translation
ENSMUSG00000041020	2900002G04Rik	1.69	−5.82	
ENSMUSG00000034336	Ina	1.59	−8.77	Nervous system development
ENSMUSG00000020705	Ddx42	1.58	−3.63	DEAD box protein
ENSMUSG00000033294	Noc4l	1.56	−2.78	
ENSMUSG00000057531	Dtnbp1	1.53	−2.87	Actin cytoskeleton reorganization
ENSMUSG00000024683	Mrpl16	1.52	−2.70	Translation
ENSMUSG00000021953	Tdh	1.52	−5.90	
ENSMUSG00000051316	Taf7	1.46	−3.04	Regulation of transcription
ENSMUSG00000024827	Gldc	1.37	−5.56	Glycine metabolism
ENSMUSG00000030505	Prmt3	1.21	−4.55	Protein amino acid methylation
ENSMUSG00000044149	Nkrf	1.15	−4.13	NF-κB repression
ENSMUSG00000000730	Dnmt3l	1.12	−7.67	In utero embryonic development
ENSMUSG00000031714	Gab1	1.05	−4.05	Epidermis development
ENSMUSG00000025050	Pcgf6	1.04	−4.38	Negative regulation of transcription
ENSMUSG00000022272	Myo10	0.98	−4.38	Signal transduction
ENSMUSG00000022652	Morc1	0.97	−8.35	Cell differentiation
ENSMUSG00000022425	Enpp2	0.91	−5.31	Lipid catabolic process
ENSMUSG00000000365	Rnf17	0.87	−6.41	Spermatogenesis
ENSMUSG00000050917	Fgf4	0.61	−8.26	Stem cell maintenance

4.3 Experimental results

4.3.1 Self-renewal assays

Further work focused on 16 genes nominated by the primary RNAi screen. To substantiate their direct role in ESC identity, three additional, independent self-renewal assays were performed. First, changes in cell morphology and alkaline phosphatase (AP) staining after esiRNA transfection were evaluated. For the negative control (Luc)-transfected

ESCs, nearly all of the colonies showed an undifferentiated morphology and highly positive AP staining (Figure 4.3A). In contrast, knockdown of most candidates resulted in reduced AP staining and obvious morphological changes in the Oct4-Gip cells, demonstrating the loss of pluripotency along with differentiation (Figure 4.3A). To exclude the possibility that these effects are Oct4-Gip cell-type specific, the same experiments were performed in R1/E ESCs and identical results were obtained.

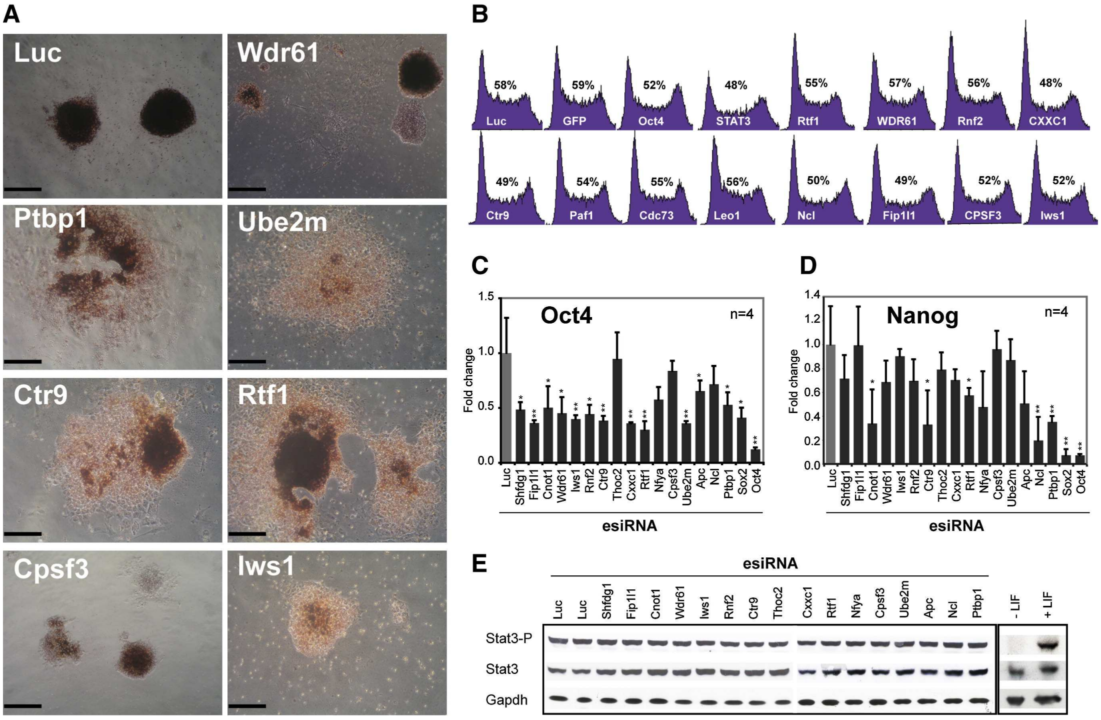


Figure 4.3: Functional Analysis of Validated Hits

(A) AP staining of ESCs 4 days after treatment with indicated esiRNAs. Note the loss of staining and the morphological changes of cells transfected with esiRNAs against Wdr61, Ptbp1, Ube2m, Ctr9, Rtf1, Cpsf3, and Iws1. Scale bars, 200 µm. (B) Cell-cycle analysis of ESCs after gene knockdown. Cell profiles recorded 4 days after transfection with indicated esiRNAs are shown. The percentage of cells in S phase is indicated above each graph. (C) Quantification of endogenous Oct4 transcript levels. qRT-PCR analysis performed 4 days after esiRNA transfection with indicated esiRNAs are shown. (D) Quantification of Nanog transcript levels. qRT-PCR analysis performed 4 days after esiRNA transfection with indicated esiRNAs are shown. (E) Analysis of Stat3 and Stat3-P protein levels. Protein extracts prepared from ESCs 4 days after transfection with indicated esiRNAs, and analyzed by immunoblotting with anti-Stat3, anti-Stat3-P, and anti-Gapdh (loading control) antibodies are presented. Cells grown for 4 days without LIF (-LIF) are shown as a positive control. (C and D) All values are means ± SD from at least triplicate experiments. * indicates significant ($p < 0.05$) and ** highly significant ($p < 0.01$) results based on Student's *t*-test analyses.

Second, changes in cell-cycle profiles upon RNAi were analyzed. Transfection of all tested esiRNAs lead to a reduction of cells in S phase (Figure 4.3B), demonstrating a change of cell-cycle regulation and consistent with an exit of these cells from self-renewal.

Third, endogenous Oct4 and Nanog transcript levels were quantified by qRT-PCR after esiRNA transfections. The transfection of all esiRNAs led to a reduction of Oct4 expression with 12 knockdowns, resulting in a significant ($p < 0.05$) or highly significant ($p < 0.01$) decrease of Oct4 transcript levels (Figure 4.3C), demonstrating that the Oct4-driven GFP expression closely mimicked the endogenous Oct4 expression. Many candidates also reduced Nanog expression (Figure 4.3D), albeit to a lower extent, possibly reflecting that some knockdowns primarily affect Oct4 expression. To test whether knockdown of the candidate genes induced ESC differentiation by interfering with the LIF/Stat3 cascade, Stat3 and phosphorylated-Stat3 (Stat3-P) levels were analyzed. No obvious change of Stat3 and Stat3-P protein levels for the knockdowns were observed, indicating that the loss of ESC identify was not mediated through perturbing the LIF/Stat3 pathway (Figure 4.3E). Collectively, these results suggest that candidate genes directly affect ESC identity through reduction of Oct4 levels.

4.3.2 The Paf1C is essential for ESC identity

Among the validated knockdowns strongly reducing Octq4 levels in ESCs were the genes Rtf1 and Ctr9 (Figure 4.1B), two components of the Pol II-associating factor 1 complex (Paf1C). The Paf1C, minimally composed of Paf1, Ctr9, Cdc73, Rtf1, and Leo1, has been implicated in multiple processes such as transcription initiation and elongation, transcript start site selection, and RNA processing (Costa & Arndt, 2000; Penheiter *et al.*, 2005; Stolinski *et al.*, 1997). In addition, the Paf1C has been linked to histone modifications in different organisms (Adelman *et al.*, 2006; Krogan *et al.*, 2003) through a stimulation of H2B ubiquitination (Ng *et al.*, 2003a), coupling Pol II elongation with SET1 and SET2 activities (Carrozza *et al.*, 2005; Krogan *et al.*, 2003; Ng *et al.*, 2003b). Because of its potential role for the regulation of ESC chromatin and because two independent components of this protein complex were among the top hits, the Paf1C was selected for further analysis.

To test whether the whole Paf1C is required for ESC identity, the expression of the remaining known Paf1C components (Paf1, Leo1, and Cdc73) was knocked down in ESCs with two independent esiRNAs. Obtained differentiation phenotypes and decreased Oct4 expression levels were similar to those observed upon Ctr9 and Rtf1 knockdown (Figures 4.4A and 4.4B and data not shown), demonstrating that the whole Paf1C is required to maintain ESC identity.

To further validate the role of the Paf1C for ESC identity, cross-species RNAi rescue experiments (Kittler *et al.*, 2005b) were performed for three of the Paf1C components (Figures 4.4C-4.4E). Stable expression at physiological levels of the human Ctr9, Rtf1, and Leo1 genes in mouse ESCs rendered these cell lines resistant to the corresponding esiRNAs, but not to the esiRNAs targeting the other Paf1C components. Therefore, the human Ctr9, Rtf1, and Leo1 genes function in mouse ESCs and can substitute their mouse orthologs. More importantly, these results manifest an essential role of the Paf1C for ESC identity.

4.3.3 Paf1C affects the expression of pluripotency and lineage-marker genes

Next, I analyzed global transcript changes upon Paf1C depletion. For this purpose, ESCs were transfected with esiRNAs targeting Ctr9, a core component of the Paf1C (Adelman *et al.*, 2006), and control esiRNA (Luc) and changes in the trancriptome were analyzed using microarrays. I identified a total of 1139 genes whose expression was perturbed significantly ($p < 0.0001$) after treatment with Ctr9 esiRNA, with 529 and 610 genes downregulated or upregulated, respectively, suggesting that the Paf1C both activates and represses target genes. GO term analysis of these genes indicated that genes implicated in biological processes relevant to embryonic development such as cell morphology and motility, cell-cycle control, cell proliferation and differentiation, oncogenesis, and ectoderm and mesoderm formation were highly enriched. Intriguingly, most known key regulators of pluripotency in ESCs, such as Nanog, Oct4, Tbx3, Esrrb Bmp4, Tcl1, Klf4, and Klf5 were downregulated upon Ctr9 depletion. In contrast, many lineage-control genes were strongly upregulated (Figure 4.4F), suggesting that the Ctr9 knockdown induced extensive differentiation. For example, ectoderm marker Fst increased 2-fold, and mesoderm markers Lef1 and Mest were upregulated 2.6-fold and 1.8-fold, respec-

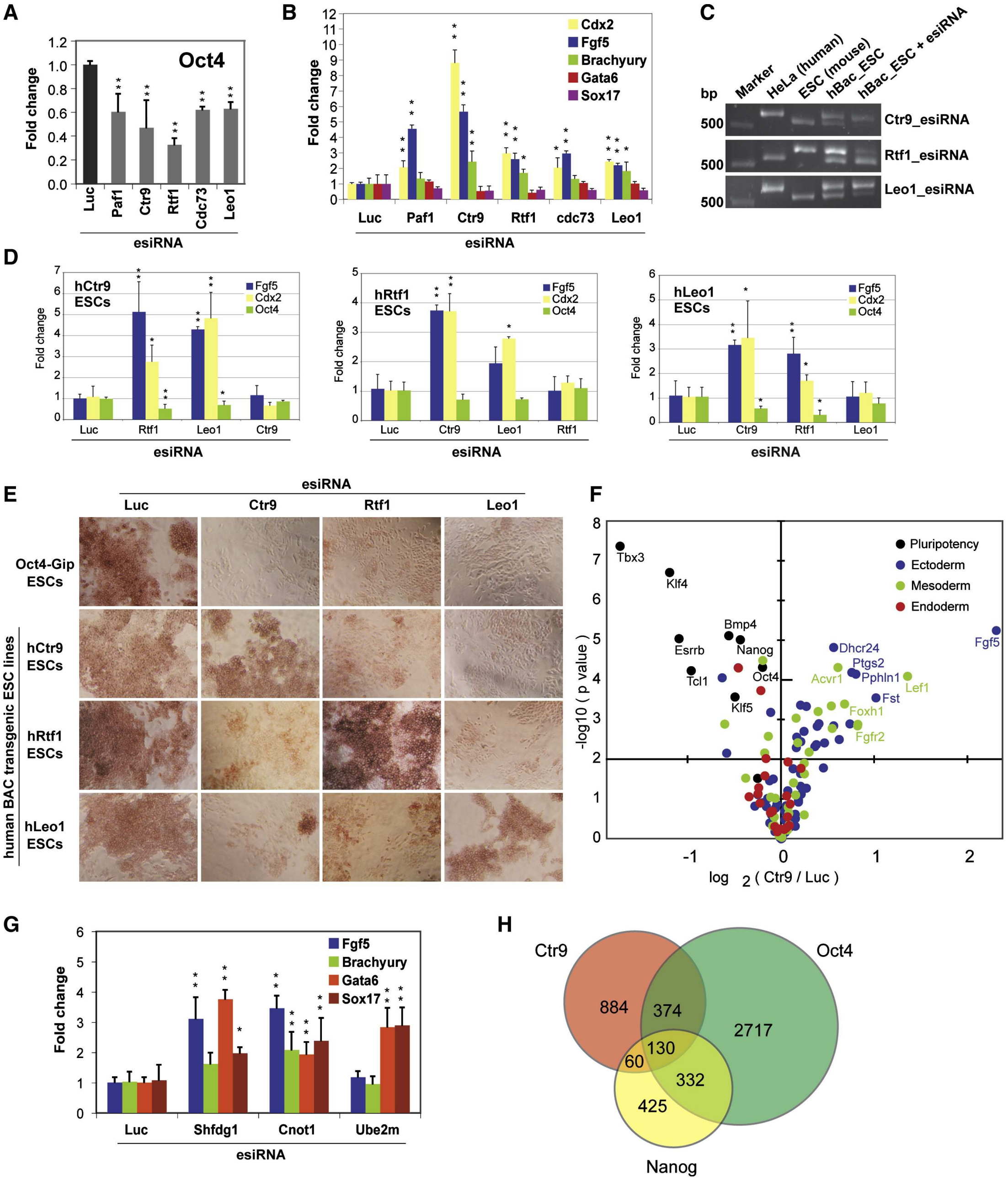


Figure 4.4: Paf1C Affects the Expression of Pluripotency and Lineage-Marker Genes

Figure 4.4: cont.

(A) RNAi of all Paf1C components leads to downregulation of endogenous Oct4 expression. qRT-PCR quantifications of Oct4 expression levels 4 days posttransfection of indicated esiRNAs are presented. (B) RNAi of all Paf1C components induces similar lineage-restricted differentiation. qRT-PCR quantifications of indicated ectoderm (blue), mesoderm (green), and endoderm (light and dark red) markers after transfection with indicated esiRNAs are shown. (C) RNAi in human BAC-transgenic ESC lines specifically depletes the mouse transcripts. PCR fragments digested with a restriction enzyme discriminating between the human and mouse transcripts are shown for human cells (HeLa), WT ESC (ESC), human BAC-transgenic ESCs (hBAC_ESC), and human BAC-transgenic ESCs transfected with indicated esiRNAs (hBAC_ESC + esiRNA), respectively. Note the reduced band intensities of the mouse product after esiRNA transfections in the human BAC-transgenic ESCs. The 500 bp band of the marker is shown on the left. (D) Oct4 and lineage-marker expression analyses in human BAC-transgenic ESCs. qRT-PCR analysis of Oct4 (green), Cdx2 (yellow), and Fgf5 (blue) after transfection of indicated esiRNAs in the indicated human BAC-transgenic ESC lines are shown. Note the rescue of the phenotypes in cell lines transfected with the corresponding esiRNAs. (E) Expression of the human gene in the mouse ESCs rescues the phenotypes. AP stainings of indicated ESC lines with indicated esiRNAs are shown. (F) Expression changes of selected genes after Ctr9 depletion. Pluripotency genes (black), lineage-marker genes, ectoderm (blue), mesoderm (green), and endoderm (red) are plotted as $\log_2(\mathrm{Ctr9/Luc})$ *versus* $-\log_{10}(p)$. (G) Knockdown of Shfdg1, Cnot1, and Ube2m induces the expression of lineage markers, including endoderm. qRT-PCR quantifications of indicated genes with indicated esiRNAs are shown. (H) Venn diagram of the number of genes regulated by Ctr9, Oct4, and Nanog. The diagram shows the overlap of genes affected by Ctr9 (red), Oct4 (green), and Nanog (yellow) knockdowns. (A, B, D, and G) All values are means ±SD from at least triplicate experiments. * indicates significant ($p < 0.05$) and ** highly significant ($p < 0.01$) results based on Student's t-test analyses.

tively, indicating differentiation of the cells toward these lineages. Differentiation of the cells was also confirmed by monitoring Fgf5 protein levels in cells transfected with Ctr9 esiRNA, as measured by immunostaining with an Fgf5 antibody and quantified utilizing an Fgf5 reporter cell line. Interestingly, I did not observe an upregulation of any gene implicated in endoderm development, suggesting lineage-restricted differentiation upon Ctr9 depletion.

To determine whether depletion of other Paf1C components results in similar expression changes, a knockdown of Paf1, Rtf1, Leo1, and Cdc73 in ESCs was performed and the expression changes of selected pluripotency and lineage-marker genes were measured by qRT-PCR. Consistent with the Ctr9 microarray data, the knockdown of all other Paf1C components induced similar expression changes (Figure 4.4B) and hence validated the downregulation of pluripotency genes and upregulation of specific lineage-control genes upon Paf1C depletion. The induced expression of trophectoderm, ectoderm, and mesoderm, but not endoderm markers, suggests that regulation of the Paf1C in the early mouse embryo may contribute to lineage specification. To test whether the observed lineage-restricted differentiation is Paf1C specific, the regulation of lineage-control genes was examined by the knockdown of three other candidate genes identified in the primary screen (Shfdg1, Cnot1, and Ube2m). These three knockdowns induced the expression of lineage markers, including endoderm markers Gata6 and Sox17 (Figure 4.4G), indicating that lineage-restricted differentiation is not generally observed upon knockdown of Oct4 modulators. In fact, the knockdown of Ube2m showed an upregulation of Gata6 and Sox17 only, indicating that the knockdown of Ube2m leads to endoderm-specific differentiation.

Depletion of the Paf1C by RNAi resulted in downregulation of both Oct4 and Nanog, suggesting that the Paf1C may be part of the Oct4-Nanog transcription circuit. Therefore, I compared the gene expression profiles upon Ctr9 knockdown to those obtained for Oct4 and Nanog knockdown (Loh *et al.*, 2006). I compared all genes regulated with a stringent cutoff ($p < 0.0001$) and found a marked overlap between the Ctr9, Oct4, and Nanog knockdown profiles (Figure 4.4H), with Pearson test correlation coefficients of 0.45 (Ctr9 *versus* Oct4), 0.50 (Ctr9 *versus* Nanog), and 0.58 (Oct4 *versus* Nanog). One hundred thirty genes were significantly regulated for all three knockdowns. Importantly, genes that were significantly downregulated included the pluripotency genes Nanog, Esrrb, Tcl1, Bmp4, and Klf4.

4.3.4 Paf1C binds to promoters of pluripotency genes

To further study the function of Paf1C in mouse ESCs, an ES line stably expressing a location and affinity purification (Cheeseman & Desai, 2005) (LAP)-tagged Ctr9 fusion protein was generated by using the bacterial artificial chromosome (BAC)-based TransgeneOmics approach (Kittler *et al.*, 2005b; Poser *et al.*, 2008). Fluorescence microscopy analysis of cell clones expressing Ctr9-LAP showed a punctate, nuclear localization, consistent with a role of Ctr9 in regulating transcription/chromatin. To investigate whether pluripotency and lineage-control genes differentially regulated upon Paf1C depletion are direct targets of the Paf1C, I analyzed the binding of the Ctr9-LAP fusion protein by ChIP-chip and identified 2175 promoter regions that were bound by Ctr9 . GO term analysis indicated that genes bound by Ctr9 were enriched for processes relevant for ESC biology, such as cell cycle, apoptosis, development, and chromatin packaging and remodeling (Figure 4.5A). Notably, many genes that are highly expressed in ESCs were not bound by the Paf1C (e.g., Ubiquitin B, α-tubulin 2, Enolase 1, Hexokinase 1, Calmodulin 2, etc.), suggesting that the Paf1C is not present at promoters of all actively transcribed genes. A closer inspection of genes falling into the developmental classification indicated that promoters of many pluripotency genes, such as Oct4, Nanog, and Sox2 were bound by the Paf1C (Figure 4.5B). The comparison of genes that were downregulated upon Ctr9 RNAi (see Figure 4.4) with genes that are bound by Ctr9 revealed a highly significant overlap ($p = 3.03 \times 10^{-07}$) and included pluripotency genes such as Oct4 and Nanog. Therefore, the Paf1C directly influences the expression of important pluripotency genes.

To further investigate the binding of the Paf1C to pluripotency and lineage commitment genes, chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR) of selected genes. Marked Ctr9-LAP binding was observed 5' proximal to the transcription start sites and coding regions of seven out of eight tested pluripotency genes (Figures 4.5C and 4.5D). This finding suggests that the Paf1C binds close to the transcription initiation site of pluripotency genes in ESC and likely travels alongside Pol II during elongation. In contrast, weak or no signals were detected for most lineage-marker and housekeeping genes. An exception was the gene Gata6, a marker for endoderm development (Hay *et al.*, 2004). Although strong Ctr9 binding at the Gata6 promoter was measured, but no signal in the coding region, indicating that the Paf1C occupies this promoter before the gene is expressed. Collectively, these results suggest that the Paf1C is a specific regulator of transcription for a subset of genes, which in ESCs include many pluripotency genes.

4.3.5 Paf1C is required to maintain the chromatin structure of pluripotency genes in ESCs

A potential mechanism for Paf1C action on the promoters of pluripotency genes may be the modulation of the local chromatin structure. In yeast, the Paf1C has been implicated in multiple aspects of histone methylation via the recruitment of methyltransferase complexes to Pol II (Krogan *et al.*, 2003). To investigate a potential role of the Paf1C for histone methylation in mouse ESCs, the effects of Ctr9 depletion on histone modifications associated with actively transcribed (H3K4 trimethylation or H3K4me3) and repressed (H3K27 trimethylation or H3K27me3) chromatin were measured for promoter regions

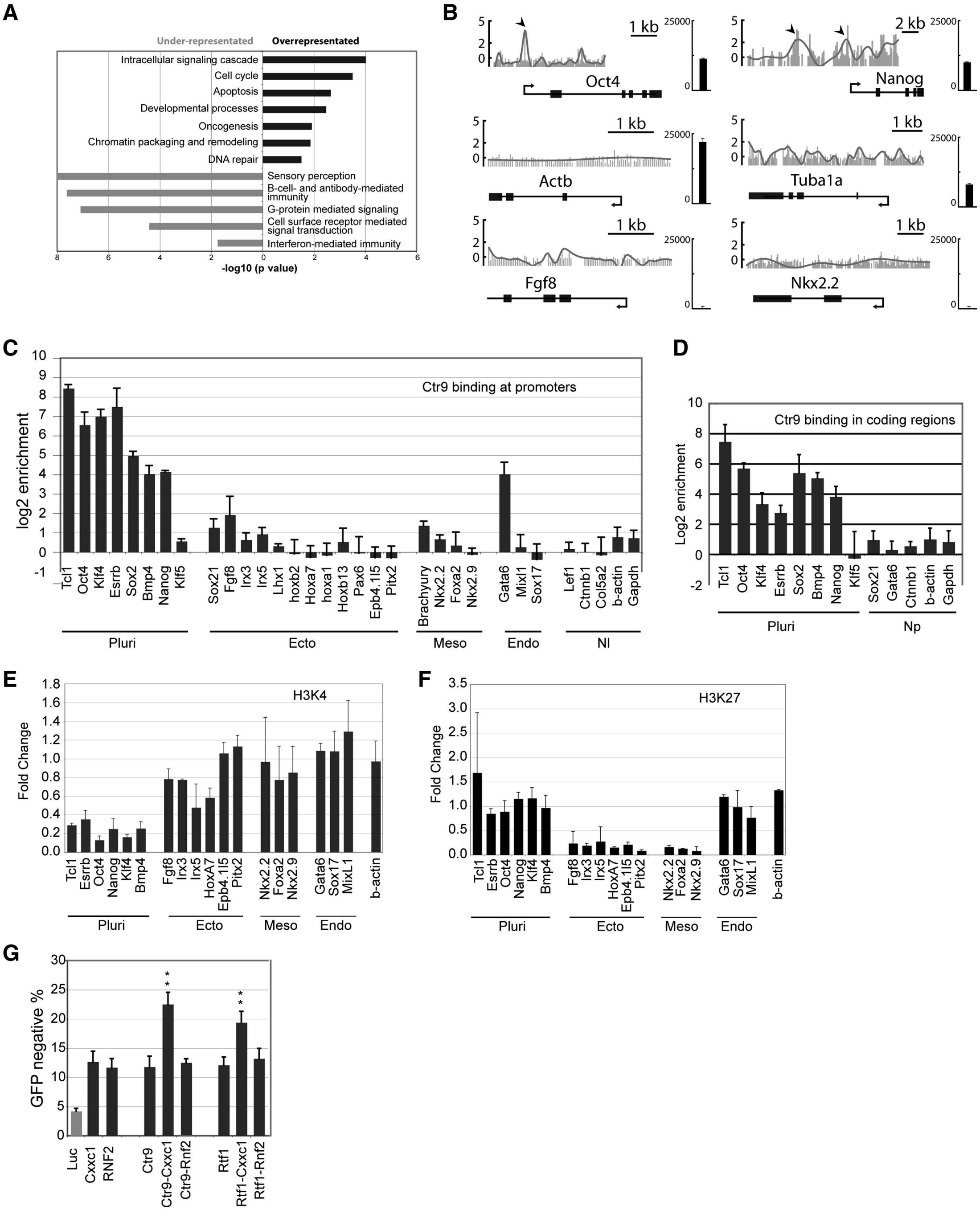


Figure 4.5: Ctr9 Binds to Promoters and Coding Regions of Pluripotency Genes and Is Required to Maintain the Chromatin Structure in ESCs

Figure 4.5: cont.

(A) ChIP-chip analysis of Ctr9 target genes. GO term enrichment analysis of selected overrepresented and underrepresented categories are presented. (B) Ctr9-binding profiles of selected genes, extracted from the ChIP-chip experiments. The binding profiles of the pluripotency genes Oct4 and Nanog, the housekeeping genes β-actin (Actb) and α-tubulin (Tuba1a), and the differentiation genes Fgf8 and Nkx2.2 are shown. Arrows indicate significant enrichment peaks for Ctr9 binding. The genomic structure of the genes is shown below the profiles with an arrow indicating the predicted transcriptional start site. The relative expression of the genes in ESCs is shown to the right of each profile. (C) Ctr9-ChIP-qPCR analysis at the promoter regions of indicated genes grouped for pluripotency (pluri), ectoderm (ecto), mesoderm (meso), endoderm (endo), and nonlineage (nl) control genes. $\log_2$ enrichment represents the abundance of enriched DNA fragments over mock controls. (D) Ctr9-ChIP-qPCR analysis at the coding regions of indicated genes. Np; nonpluripotency genes. (E) Depletion of Ctr9 results in a decrease of H3K4me3 on the promoters of pluripotency genes. ChIP assays for indicated genes are shown. Fold enrichment represents the abundance of enriched DNA fragments over mock controls. (F) Depletion of Ctr9 results in a decrease of H3K27me3 on the promoter of some lineage-control genes. ChIP assays for indicated genes are shown. (G) Synthetic analysis of Paf1C components with Cxxc1 and Rnf2. FACS quantification of GFP-negative cells 4 days after transfection of Oct4-Gip ESCs with indicated esiRNAs are shown. (C, D, E, F, and G) All values are means ±SD from at least triplicate experiments. * indicates significant ($p < 0.05$) and ** highly significant ($p < 0.01$) results based on Student's t-test analyses.

of selected genes. ChIP-qPCR analyses indicated that H3K4me3 levels on promoters of pluripotency genes strongly decreased upon Ctr9 depletion, suggesting that the Paf1C is required for the maintenance of H3K4me3. For the lineage-control genes, the H3K4me3 levels remained essentially unchanged, whereas the H3K27me3 levels for the ectoderm and the mesoderm specification genes decreased markedly (Figures 4.5E and 4.5F). To substantiate this finding, Paf1C double knockdowns were performed together with the trithorax group-like, Set1 complex subunit gene, Cxxc1 (required for H3K4me3), or with the polycomb group Pc1 complex subunit gene, Rnf2/Ring1b (required for H3K27me3), both of which individually affect ESC pluripotency (Fazzio *et al.*, 2008; Lee & Skalnik, 2005, 4.1B). A robust enhancement of the phenotype was observed when Ctr9 or Rtf1 was cosilenced together with Cxxc1, but no effect when the same Paf1C subunits were cosilenced together with Rnf2 (Figure 4.5G). Inspection of Oct4 expression levels and changes in the expression of the differentiation markers Cdx2, Fgf5, and Brachyury indicated that the simultaneous depletion of Paf1C with Cxxc1 enhanced the expression of these differentiation markers. Together, these results indicate that the Paf1C synergizes with the Set1 complex to maintain ESC pluripotency and support a direct role of the Paf1C in maintaining H3K4me3 at promoter regions of pluripotency genes.

4.3.6 Proteomic analyses of Paf1C in ESCs

The Paf1C is composed of at least five subunits, Paf1, Ctr9, Rtf1, Cdc73, and Leo1. Recently, the human Paf1C complex has been shown to interact with hSki8, a component of the SKI complex, which together with the exosome mediates 3'-5' mRNA degradation (Zhu *et al.*, 2005). This interaction suggests a possible link of the human Paf1C to RNA quality control and extends the proteins that comprise the human Paf1C. Interestingly, the mouse ortholog of hSki8, Wdr61 was among the 16 validated strongest hits in our primary screen (Figure 4.1B), suggesting a potential role of this gene in mouse ESCs as a component of the Paf1C.

To identify potential interaction partners of the Paf1C in ESCs, a proteomic analysis using the Ctr9-LAP and a Leo1-LAP-tagged ESC lines was performed. Both lines express similar levels of the LAP-tagged and the endogenous Paf1C transcripts. After affinity purification with an anti-GFP antibody, prey proteins were eluted and analyzed by mass spectrometry. This analysis revealed most of the known Paf1C components (Table 4.2). In addition, several Pol II-associated and chromatin-modifying proteins were immunoprecipitated, including Pabc1, Ruvbl1, Ruvbl2, Sfpq, Tceb3, Polr2a, Polr2e, CoREST, and Smarce1, implying a complex interacting network of Paf1C with other transcriptional and chromatin regulators. Importantly, Wdr61 was identified as a Paf1C interaction partner by both Ctr9-LAP and Leo1-LAP pulldowns, authenticating the interaction of this protein with the Paf1C in mouse ESCs (Table 4.2). To validate the interaction of Wdr61 with Paf1C, a Wdr61-LAP ESC line was generated and its interaction partners were analyzed via mass spectrometry. The Wdr61-LAP pulldown identified two Paf1C components, Cdc73 and Leo1, and the Pol II-associated proteins Ruvbl1 and Ruvbl2, further validating the interaction between Wdr61, the Paf1C, and several other transcriptional regulators (Table 4.2). Similar knockdown phenotypes and protein-protein interactions strongly suggest that Wdr61 functions together with Paf1C to maintain ESC identity.

Abbreviations: Baits, Bait used for IP; Name, Name of the gene; ACC_NO, accession number in International Protein Index; Mass, molecular weight of predicted protein; Score, probability-based MOWSE score of MASCOT software; Matches, number of total matched peptides via MASCOT; PEP_UNIQ, number of unique peptide sequences

Table 4.2: Proteomic Analyses of the Paf1C in ESCs

Baits	Name	Accession Number	Mass	Score	Matches	PEP_UNIQ	SEQ_COV (%)
Ctr9	Ctr9	IPI00477468	133.420	2177	55	43	44
Ctr9	Paf1	IPI00331654	60.481	1140	28	23	54
Ctr9	Cdc73	IPI00170345	60.539	1116	28	25	52
Ctr9	Leo1	IPI00474486	76.912	817	17	13	23
Ctr9	Wdr61	IPI00112320	33.778	640	21	13	59
Ctr9	Pabpc1	IPI00124287	70.626	414	8	8	17
Ctr9	RuvB-like 1	IPI00133985	50.182	272	6	6	17
Ctr9	COREST	IPI00226581	53.272	201	4	4	12
Ctr9	RuvB-like 2	IPI00123557	50.949	182	4	4	9
Ctr9	SFPQ	IPI00129430	75.394	148	3	3	7
Leo1	Leo1	IPI00103090	75.359	798	25	17	22
Leo1	Ctr9	IPI00120919	133.420	960	26	24	17
Leo1	Cdc73	IPI00170345	60.539	905	27	20	35
Leo1	Paf1	IPI00331654	60.481	788	18	14	26
Leo1	Wdr61	IPI00112320	33.752	472	14	10	27
Leo1	Tceb3	IPI00317167	87.124	378	11	10	13
Leo1	Polr2a	IPI00136207	217.039	153	5	5	2
Leo1	Smarce1	IPI00119892	46.610	116	2	2	5
Leo1	RuvB-like 2	IPI00123557	51.081	111	2	2	4
Leo1	Polr2e	IPI00337955	24.555	107	3	3	14
Wdr61	Wdr61	IPI00019269	33.557	806	19	13	51
Wdr61	Cdc73	IPI00170345	60.539	208	5	5	13
Wdr61	Leo1	IPI00474486	75.596	116	2	2	3
Wdr61	SF3B4	IPI00154082	44.327	196	3	3	10
Wdr61	RuvB-like 1	IPI00133985	50.182	175	6	6	18
Wdr61	RuvB-like 2	IPI00123557	51.081	101	3	3	8

identified via MASCOT; SEQ_COV, percentage of predicted protein sequence covered by matched peptides via MASCOT.

4.3.7 Paf1C is downregulated during embryoid body formation

To analyze possible expression changes of the Paf1C during early development, the expression of Paf1C subunits were measured during ESC differentiation in embryoid bodies (EBs). For all Paf1C subunits, downregulated expression was observed during the formation of EBs (Figure 4.6A), indicating that the Paf1C is downregulated during early embryonic development. To test whether the reduction in Paf1C subunit expression levels is just part of a general decrease in the expression of Pol II-associated protein complexes during differentiation, the expression levels of subunits of the Pol II-associated mediator complex were measured in EBs (Myers & Kornberg, 2000). In contrast to the Paf1C, transcript levels of the investigated mediator complex subunits did not change significantly during EB formation (Figure 4.6A). Hence, the downregulation of the Paf1C may be important for proper embryonic development. Support for a role of the Paf1C

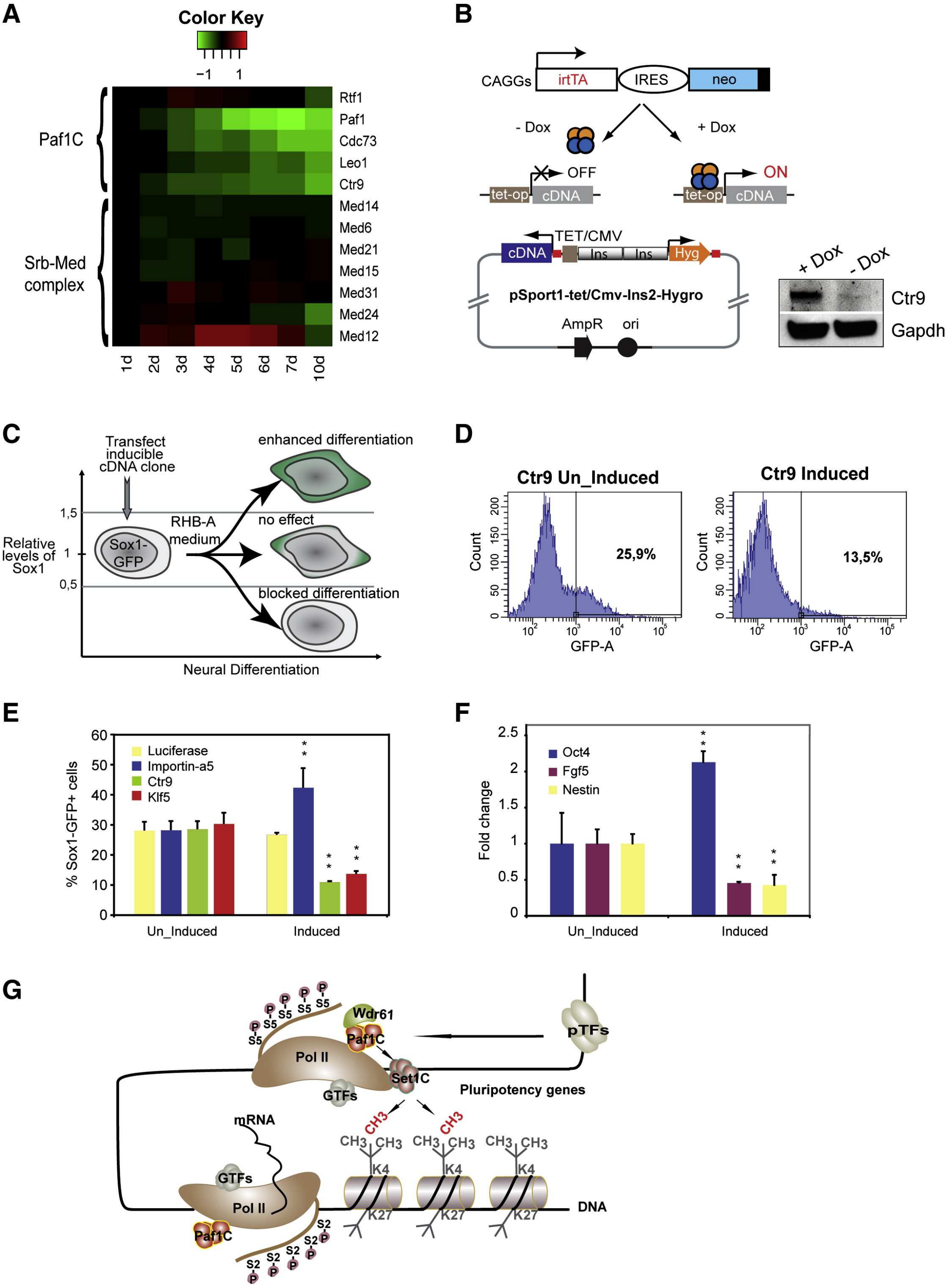


Figure 4.6: Paf1C Is Downregulated during Embryoid Body Formation, and Overexpression Blocks ESC Differentiation.

Figure 4.6: cont.

(A) Expression changes of indicated genes from the Paf1C and the mediator complex (Srb-Med) during EB formation are shown. The heat map shows data ($\log_2$ values) extracted from gene expression arrays of RNA hybridized from day 1 to day 10 (1d–10d). (B) Shown are constructs used to transfect the Sox1-GFP ESCs. Relevant elements and a scheme of the function of the inducibility of the expression constructs by doxycycline (-Dox, +Dox) is presented. A western blot analysis shows the induced expression of Ctr9 after doxycycline treatment. CAGGs, chicken β-actin promoter coupled to a CMV early enhancer; irtTA, reverse-tetracycline repressor; IRES, internal ribosome entry site; neo, neomycin resistance gene; Ins, chicken β-globin insulator; tet-op, tetracycline operator-binding sites; Hyg, hygromycine resistance gene; AmpR, ampicillin resistance gene; ori, ColE1 origion of replication. (C) Scheme of the differentiation assay. Arrows show possible outcomes after induced expression of cDNA constructs. The variable amount of GFP expression in the cells is indicated by different intensities of green. (D) FACS profiles and the percentage of GFP-positive Sox1-GFP cells without (Ctr9 Un_Induced) and with (Ctr9 Induced) Ctr9 overexpression are presented. (E) Comparison of different genes blocking, or enhancing, differentiation in the Sox-GFP assay. The percentage of GFP-positive cells after transfection of cDNA constructs of indicated genes, with and without doxycyline treatment, are presented. (F) Expression changes of indicated genes upon overexpression of Ctr9 (induced), or without Ctr9 overexpression (Un_induced), are shown. (G) Model for Paf1C function in ESCs. The model depicts the concerted action of pluripotency transcription factors (pTFs), the Paf1C, and the Set1C to maintain active transcription (H3K4me3) of pluripotency genes. GTFs, general transcription factors; S5-P, phosphorylated serine 5 of the CTD tail of Pol II; S2-P, phosphorylated serine 2 of the CTD tail of Pol II. Values in (E) and (F) are means ±SD from experiments done in triplicate. ** indicates highly significant ($p < 0.01$) results based on Student's t-test analyses.

during development comes from observations in Drosophila and Zebrafish, in which a variety of embryonic developmental defects have been described for Paf1C component depletions (Akanuma *et al.*, 2007; Tenney *et al.*, 2006), and from Cdc73 knockout mice, which develop normally up to E3.5 but then die either at hatching or implantation (Wang *et al.*, 2008a).

4.3.8 Paf1C overexpression blocks ESC differentiation

Many pluripotency genes block differentiation when overexpressed (Chambers *et al.*, 2003; Turksen, 2001). To test whether overexpression of a Paf1C component can block ESC differentiation, a Sox1-GFP reporter cell line (Aubert *et al.*, 2003) was transfected with tetracycline-inducible Ctr9 and control constructs (Figures 4.6B and 4.6C). Sox1 is not expressed in undifferentiated ESCs but is strongly induced during neuronal differentiation (Aubert *et al.*, 2003) and during cultivation in a neuronal growth and differentiation medium (Diogo *et al.*, 2008). Because a prominent activation of ectodermal genes was observed upon Paf1C RNAi, potentially overexpression of a Paf1C component will block differentiation, as reported by the acquisition of Sox1-GFP expression. Indeed, a marked reduction of GFP-positive cells upon transfection of the Ctr9 construct was only detected when tetracycline was present in the culture medium (Figures 4.6D and 4.6E). A block of differentiation was further supported by analyses of the differentiation markers Fgf5 and Nestin and the pluripotency factor Oct4 in these cells (Figure 4.6F). These data demonstrate that an excess of a Paf1C component can block the differentiation of ESCs.

4.4 Discussion

A detailed molecular understanding on how genetic factors influence the balance between pluripotency and differentiation of mammalian cells is crucial to develop ESCs for their potential therapeutic use. This understanding is becoming increasingly important because it is now possible to generate ESC-like cells from somatic cells via direct reprogramming (reviewed in Jaenisch & Young, 2008), avoiding many ethical issues. In the future, it seems feasible to bank personalized iPS cells that may be used to generate differentiated cells to replace damaged tissue in the body when needed. Obviously, this scenario requires a detailed and systematic understanding of ESC/iPS self-renewal and differentiation to develop protocols for safe tissue replacement therapies.

To initiate a global survey of genes required to maintain mouse ESC identity, a genome-scale RNAi screen was performed. This screen identified 296 candidate genes with numerous functions that may influence Oct4 expression levels, suggesting a complex interplay of several biological processes required to maintain pluripotency. This data set should present a useful resource to characterize factors that influence ESC self-renewal and ultimately may also be useful to improve the reprogramming protocols of somatic cells. Oct4 is one of the factors used to reprogram somatic cells into iPS cells (reviewed in Jaenisch & Young, 2008). Thus, identifying the genes that regulate the expression of Oct4 should be instrumental for understanding the reprogramming process. Understanding the endogenous regulation of Oct4 may reveal alternative ways to activate Oct4 expression in somatic cells. As such, the genes that alter the expression of Oct4 upon RNAi present a good starting point to unravel the network of Oct4 regulation.

Most of the strongest hits identified in our screen are transcription factors and/or chromatin modifiers that regulate transcriptional processes. Modulation of transcription activity requires the interaction of transcription factors, the Pol II basal transcription machinery, and factors regulating chromatin structure. The enrichment of transcription factors and chromatin modifiers on the one hand highlights the crucial role of transcription regulation on ESC fate decision and, on the other hand, offers an opportunity to unravel the interactions between transcription factors, chromatin modifiers, and their target DNA for the maintenance of ESC identity. Here we focused on the Paf1C was chosen because we had two independent, highly significant, starting hits, and also we were excited by the potential discovery of a novel regulatory mechanism.

The Paf1C was originally identified in yeast as a specific Pol II-associated protein complex biochemically distinct from the Srb/Med-containing Pol II holoenzyme (Shi *et al.*, 1997). Further genetic and biochemical studies showed that the Paf1C is implicated in transcript start site selection (Stolinski *et al.*, 1997), initiation and elongation (Costa & Arndt, 2000; Mueller & Jaehning, 2002), poly(A) site utilization (Penheiter *et al.*, 2005), and histone methylation (Krogan *et al.*, 2003), suggesting a complex role of the Paf1C for gene regulation in yeast. In metazoans, the Paf1C has additionally been implicated in Notch and Wnt signaling, indicating that this protein complex participates in biological processes such as oncogenesis and embryogenesis (Akanuma *et al.*, 2007; Mosimann *et al.*, 2006; Tenney *et al.*, 2006).

This chapter demonstrates an important role of the Paf1C for the maintenance of mouse ESC identity. Knockdown of all known complex subunits led to a decreased expression of Oct4 and other pluripotency markers, accompanied by a loss of ESC self-renewal and subsequent differentiation. These phenotypic consequences could be rescued by physiological expression of the human Paf1C components. Global gene expression analysis after Paf1C depletion revealed a change in the expression of specific genes, with most known pluripotency genes being downregulated. This observation was substantiated by ChIP studies, where the Paf1C was found to directly bind to the promoters of many pluripotency genes. Hence, the Paf1C is important for the sustained expression of pluripotency genes in mouse ESCs. The large overlap of regulated genes upon Oct4, Nanog, and Ctr9 knockdown argues that the Paf1C is an integral part of the Oct4-Nanog regulatory circuit. Because the Paf1C interacts with the Pol II machinery,it is reasonable to hypothesize that the Paf1C integrates signals from the pluripotency transcription factors to establish a specialized Pol II complex that maintains active transcription of these genes (Figure 4.6G). The observed downregulation of Paf1C components during

EB formation and the block of differentiation upon overexpression further supports this notion.

The recent identification of 'bivalent' promoters carrying both H3K4me3 and H3K27me3 in ESCs identified a subset of genes crucial to lineage commitment decisions. To this subset, we now add another important subset of genes regulated by the Paf1C, which are crucial to self-renewal. Furthermore, our study identifies genes whose downregulation may be essential to initiate differentiation into specific lineages in the developing embryo.

Because the role of the Paf1C is multifaceted, its role in maintaining pluripotency may also be complex. Here it is shown that the Paf1C is required for the maintenance of H3K4me3 on pluripotency genes. The selective binding of the Paf1C to the promoters of pluripotency genes combined with the synergistic effect of Paf1C knockdown with a key subunit of the Set1 H3K4 methyltransferase complex indicates a direct role of the Paf1C in the modulation of chromatin structure at pluripotency genes in ESCs. In yeast, the Paf1C is required for recruitment of H3K4 methyltransferase activity to Pol II (Krogan *et al.*, 2003; Ng *et al.*, 2003b). Our data suggest a similar role of the Paf1C at pluripotency genes in ESCs. According to this model (Figure 4.6G), depletion of the Paf1C would result in a decrease of the Set1 complex at promoter regions of pluripotency genes, which consequently would result in the loss of H3K4 trimethylation. As a result, expression levels of pluripotency genes would diminish, initiating the differentiation process.

The proteomic analyses of the Paf1C components support an intricate role of this protein complex in ESCs. Although Ctr9 and Leo1 pulldowns identified most known Paf1C components, they also revealed specific interactions with other Pol II-associated factors. Interestingly, some of these specific interactors have been implicated in ESC pluripotency. For instance, Tceb3, a Leo1-interacting protein, has been shown to regulate transcription of a subset of genes linked to cell-cycle progression in mouse ESCs (Yamazaki *et al.*, 2003).

Lineage-restricted differentiation upon Paf1C depletion was unexpected. Instead of uniformly exiting from self-renewal into all four alternatives (trophectoderm, endoderm, ectoderm, mesoderm), Paf1C-depleted ESCs appear impaired to differentiate toward endoderm. Furthermore, we note that the promoter region of the endoderm control gene Gata6 was occupied by the Paf1C before the gene was expressed. Consequently, we speculate that key endoderm control genes require the Paf1C for the induction of their expression and that Paf1C plays a specific role in early endodermal commitment.

In practical terms, the lineage-restricted ESC differentiation upon Paf1C and Ube2m knockdowns may be useful for the directed generation of specific cell types from ESCs. To date, strategies to accomplish this goal mainly rely on overexpression of lineage commitment genes and the use of certain growth-factor-enriched media (Turksen, 2001). Our study suggests that enhanced ESC differentiation protocols can be developed by combining these protocols with the depletion of certain genes by RNAi. These experimental modifications should improve the efficiency for generating specific cell types for experimental or therapeutic transplantation.

In summary, our work provides a global view of ESC pluripotency, revealing that an integrated analysis of transcription, chromatin structure, signaling, and possibly other biological processes is needed to fully understand ESC pluripotency. By analyzing one of the protein complexes in more detail, we have begun to uncover specific connections between these processes, in which interplay of transcription factors, Pol II-associated factors, and chromatin regulation is needed for maintaining ESC identity. A picture emerges in which specific transcription factor combinations work in concert with specialized Pol II complexes assembled on particularly marked chromatin structures to control the expression of pluripotency genes.

4.5 Methods

4.5.1 Mouse esiRNA Library

The templates for the esiRNA synthesis were generated using the cDNA library Mouse Unigene Set, RZPD 2.1 (Imagenes, `http://www.imagenes-bio.de/products/sets_libraries/non_redundant_sets`). EsiRNAs were synthesized as described previously (Kittler *et al.*, 2005a) and normalized to 100 ng/µl in 384-well plates for the genome-scale screen.

4.5.2 Cell Culture and High-Throughput Screen

ESCs (E14TG2a, R1/E, Oct4-Gip Ying *et al.* (2002), and BAC-transgenic ESC lines) were cultured on gelatin-coated plates in Glasgow Minimum Essential Medium (Sigma) supplemented with 10 % FBS (Pan biotech), 2.2 mM L-glutamine, 1 mM sodium pyruvate, 50 µM 2-mercaptoethanol, 1× NEAA (Invitrogen), and LIF (generated in house) as previously described (Bernstein *et al.*, 2006). ESCs were trypsinized and split every 2 days, and the medium was changed daily. For the genome-scale screen, reverse transfections were performed using mixtures of 20 ng esiRNA and 0.075 µl Lipofectamine 2000 (Invitrogen) in 10 µl optimum medium (Invitrogen). ESCs were plated in 384-well plates with a density of 900 cells per well in 60 µl ESC medium. On each plate, eight negative controls (Luciferase esiRNA) and eight positive controls (four GFP esiRNAs, four Sox2 esiRNAs) were placed to monitor the transfection efficiency. GFP fluorescence and cell numbers were measured 96 h posttransfection using a FACS Calibur (BD Biosciences) equipped with an HTS loader for high-throughput readout.

4.5.3 Alkaline Phosphatase Staining

ESCs (2 × 103) were reverse transfected with 50 ng esiRNAs and 0.2 µl Lipofectamine 2000 in 96-well plates. Four days posttransfection, ESCs were fixed in 4 % paraformaldehyde (Sigma) for 5 min at room temperature. After two times rinsing with PBS, ESCs were stained using the Alkaline Phosphatase Red Microwell Substrate (Sigma).

4.5.4 Western Blotting

Oct4-Gip cells (8 × 104) were reverse transfected with 800 ng esiRNAs and 2 µl Lipofectamine 2000 in 6-well plates. Four days posttransfection, ESCs were harvested, and 10 µg of protein extracts were separated using NuPAGE 4 %–12 % Bis-Tris protein gels (Invitrogen) and blotted to nitrocellulose membrane (Millipore). The membranes were probed with the primary antibodies against Stat3 (H-190; sc-7179, Santa Cruz), phospho-Stat3 (Tyr705; 44-380G, Invitrogen), and Gapdh (NB300-221, Novus Biologicals).

4.5.5 RT-qPCR

Total RNA was isolated by using the RNeasy Mini kit (QIAGEN), and 1 µg RNA was reverse transcribed with SuperScript III Reverse transcriptase (Invitrogen) utilizing an oligo(dT)18 primer. qPCRs were performed with the SYBR Green qPCR kit (Abgene) on an MX P3000 qPCR machine (Stratagene). Measured transcript levels were normalized to Gapdh. Samples were run in triplicate.

4.5.6 Gene Expression Analyses

Oct4-Gip cells (8 × 104) were reverse transfected with 800 ng esiRNAs and 2 µl Lipofectamine 2000 in 6-well plates. Four days posttransfection, RNA was prepared using the RNeasy Mini kit (QIAGEN) and labeled with the One-Cycle Target Labeling and Control Reagent Package (Affymetrix), as described in the manufacturer's instructions. Extracts from four biological replicated were hybridized to Mouse Genome 430 2.0 arrays (Affymetrix), and the data were analyzed by hierarchical clustering using Cluster 3.0 software. Affymetrix data are accessible through the GEO series under accession number GSE12078.

4.5.7 Cell-Cycle Profiling

Oct4-Gip cells were transfected as for the gene expression analyses. After 96 h, cells were trypsinized, washed with PBS, and fixed overnight with ice-cold 70 % ethanol. After washing with PBS, cells were stained with PI solution (25 µg/ml PI, 20 µg/mll RNaseA, 0.02 % Triton X-100) for 30 min in the dark, and the cell-cycle profiles were acquired by FACS.

4.5.8 Cross-Species RNAi Rescue

Cross-species RNAi rescue experiments were performed as previously described (Kittler *et al.*, 2005b). Briefly, Oct4-Gip and human BAC-transgenic ESC lines were generated and transfected with esiRNAs, which only target the mouse transcripts. Cells were transfected as for the gene expression analyses. Four days posttransfection, cells were stained with alkaline phosphstase, and RNAs from esiRNA-transfected ESCs were prepared by using the RNeasy Mini kit (QIAGEN) and reverse transcribed with SuperScript III Reverse transcriptase (Invitrogen). qPCRs were performed as described before to quantify the expression of Cdx2, Fgf5, and Oct4. Samples were run in triplicate, and measured transcript levels were normalized to Gapdh. To evaluate the specificity and efficiency of knockdowns, cDNA fragments from HeLa cells, WT mouse ESCs, human BAC-transgenic ESCs, and human BAC-transgenic ESCs transfected with esiRNAs were PCR amplified with primers that perfectly match the human and mouse transcripts. To discriminate the human and mouse transcripts, PCR products for Ctr9, Rtf1, and Leo1 were digested with XhoI, XbaI, and PstI, respectively, and separated on a 2 % agarose gel.

4.5.9 Chromatin Immunoprecipitation and Protein Identification by Mass Spectrometry

ChIP assays of mouse ESCs were essentially performed as previously described by Bernstein *et al.* (2006) using antibodies against H3K4 (Ab8580, Abcam) and H3K27 (07-449, Upstate). LAP tag-based ChIP assays were performed with BAC-transgenic ESC lines as previously described using a polyclonal goat anti-GFP antibody (Poser *et al.*, 2008). BAC-transgenic ESC lines Ctr9-LAP, Leo1-LAP, Wdr61-LAP, and Rtf1-LAP were generated by tagging the BACs RP11-77G5, RP11-56B16, RP11-1132M10, and RP11-16O9, respectively. Enrichments of target genes were quantified by qPCR. Genome-wide location analysis was carried out using mouse promoter array 1.0R array (Affymetrix). The data are accessible through the GEO series accession number GSE14654. GFP-tagged protein complexes were isolated by immunoaffinity chromatography as previously described (Poser *et al.*, 2008) and analyzed by mass spectrometry.

4.5.10 Overexpression Studies in Sox1-GFP ESCs

The pSport1-tet/Cmv-Ins-Hygro expression constructs contain full-length cDNAs (Luciferase, Importin-a5, Ctr9, and Klf5) under the control of a tetracycline-inducible CMV-minimal promoter. The Sox1-GFP (46C) ESC line (Ying *et al.*, 2003), stably expressing a codon-optimized tetracycline activator irtTA (Anastassiadis *et al.*, 2002) from the CAGGs promoter, was grown in a chemically defined medium (ESGRO complete, Millipore) with BMP4 and LIF. The cells were then trypsinized and plated in RHB-A medium (Stem Cell Sciences) on gelatin-coated 6-well plates in duplicates and retrotransfected at day 0 with the tetracycline-inducible plasmids (1 µg each) plus the normalization plasmid ptet-Gaussia Luciferase (200 ng) in the presence or absence of doxycyline (1 µg/ml). RHB-A medium was changed daily. At day one, the supernatant was collected, and the Gaussia luciferase activity was measured to normalize for transfection efficiencies. At day three, cells were collected, fixed in 2 % paraformaldehyde, and analyzed by FACS. At the same time, RNA was extracted for expression analyses by qRTPCR.

Chapter 5

PhenoFam–gene set enrichment analysis through protein structural information

With the current technological advances in high-throughput biology, the necessity to develop tools that help to analyse the massive amount of data being generated is evident. A powerful method of inspecting large-scale data sets is gene set enrichment analysis (GSEA) and investigation of protein structural features can guide determining the function of individual genes. However, a convenient tool that combines these two features to aid in high-throughput data analysis has not been developed yet. To fill this niche, we developed the user-friendly, web-based application, PhenoFam.

PhenoFam performs gene set enrichment analysis by employing structural and functional information on families of protein domains as annotation terms. My tool is designed to analyse complete sets of results from quantitative high-throughput studies (gene expression microarrays, functional RNAi screens, *etc.*) without prior prefiltering or hits-selection steps. PhenoFam utilizes Ensembl databases to link a list of user-provided identifiers with protein features from the InterPro database, and assesses whether results associated with individual domains differ significantly from the overall population. To demonstrate the utility of PhenoFam, we analysed a genome-wide RNA interference screen and discovered a novel function of plexins containing the cytoplasmic RasGAP domain. Furthermore, a PhenoFam analysis of breast cancer gene expression profiles revealed a link between breast carcinoma and altered expression of PX domain containing proteins.

PhenoFam provides a user-friendly, easily accessible web interface to perform GSEA based on high-throughput data sets and structural-functional protein information, and therefore aids in functional annotation of genes.

5.1 Background

Analysis of large sets of results derived from high-throughput experiments is a challenging but promising field of study. Enrichment analysis is a very powerful strategy helping researchers in identifying biological processes or pathways related to their studies. Most of the currently available tools, *i.e.* Onto-Express (Khatri *et al.*, 2002), DAVID (Dennis *et al.*, 2003), FatiGO + (Al-Shahrour *et al.*, 2007), ConceptGene (Sartor *et al.*, 2009) and others reviewed in (Huang *et al.*, 2009), search for enrichment of Gene Ontology (GO) terms (Ashburner *et al.*, 2000), KEGG pathways (Kanehisa *et al.*, 2010) or other functional properties in a pre-selected subset of genes by contrasting it with the background set, usually a whole genome. This approach strongly relies on a chosen hit selection algorithm and user-defined thresholds. Moreover, the experimental results (*i.e.* level of expression or phenotype strength) are not considered. There are few applications overcoming these limitations by performing gene set enrichment analysis (GSEA) (Mootha *et al.*, 2003). They search for gene annotations enriched on the top or the bottom of a complete list of genes ranked by their experimental values. This allows even mild effects to contribute to the overall enrichment score. However, to my knowledge, annotations used by available GSEA tools have so far primarily been used in combination with GO terms, pathways or transcription factors, and only few of these applications are web-based, *e.g.* GSEA (Subramanian *et al.*, 2005), FatiScan (Al-Shahrour *et al.*, 2007), GeneTrail (Keller *et al.*, 2008).

In recent years, access to high-resolution protein structural information has increased considerably. Many new structures reveal the presence of domains known from other proteins, and the domain composition of a protein can help forming a hypothesis about its biological function, *e.g.* a homeodomain fold indicates a transcription factor activity involved in cellular differentiation (Gehring *et al.*, 1994). Moreover, Hahne *et al.* demonstrated, that the domain composition of proteins could be used for predicting their pathway membership (Hahne *et al.*, 2008). There are many databases classifying and providing information about protein families, domains, regions and functionally relevant sites. InterPro (Hunter *et al.*, 2009) constitutes a repository that integrates a number of the most well established sources of data: PROSITE (Hulo *et al.*, 2008), HAMAP (Lima *et al.*, 2009), Pfam (Finn *et al.*, 2010), PRINTS (Attwood *et al.*, 2003), ProDom (Corpet *et al.*, 2000), SMART (Letunic *et al.*, 2002), TIGRFAMs (Haft *et al.*, 2003), PIRSF (Wu *et al.*, 2003), SUPERFAMILY (Gough *et al.*, 2001), Gene3D (Pearl *et al.*, 2005) and PANTHER (Mi *et al.*, 2005). I have developed a GSEA web application that can be used for analysing data from large-scale experiments (phenotypes, gene expression, *etc.*). My tool combines the experimental results with annotations from the databases integrated in InterPro (called ‘member databases’), thereby allowing a streamlined structure/function annotation of proteins. Utilization of information about protein domain families in GSEA is a novel approach that can be used in parallel to other enrichment analysis applications.

5.2 Implementation

5.2.1 Data management

PhenoFam is a Java web application running on a Tomcat 5.5 server. It uses a MySQL database to store mappings between various protein, gene or probe names and identifiers related to member databases of InterPro (Figure 5.1). This database is an easily updatable compilation of the current releases of the Ensembl database (Hubbard *et al.*, 2009). Client-server communication is mainly handled by AJAX technologies. User-uploaded data sets and calculation results are stored as session objects on the server side for at least 30 minutes after closing the browser window.

5.2.2 Identifiers association

One of the key features of my application is that it accepts as input a wide range of identifiers used in all genomes integrated in the Ensembl database (Hubbard *et al.*, 2009). Identifiers provided by the user are translated into respective Ensembl (gene, or transcript) identifiers and, using mappings from the InterPro database, linked to none, one

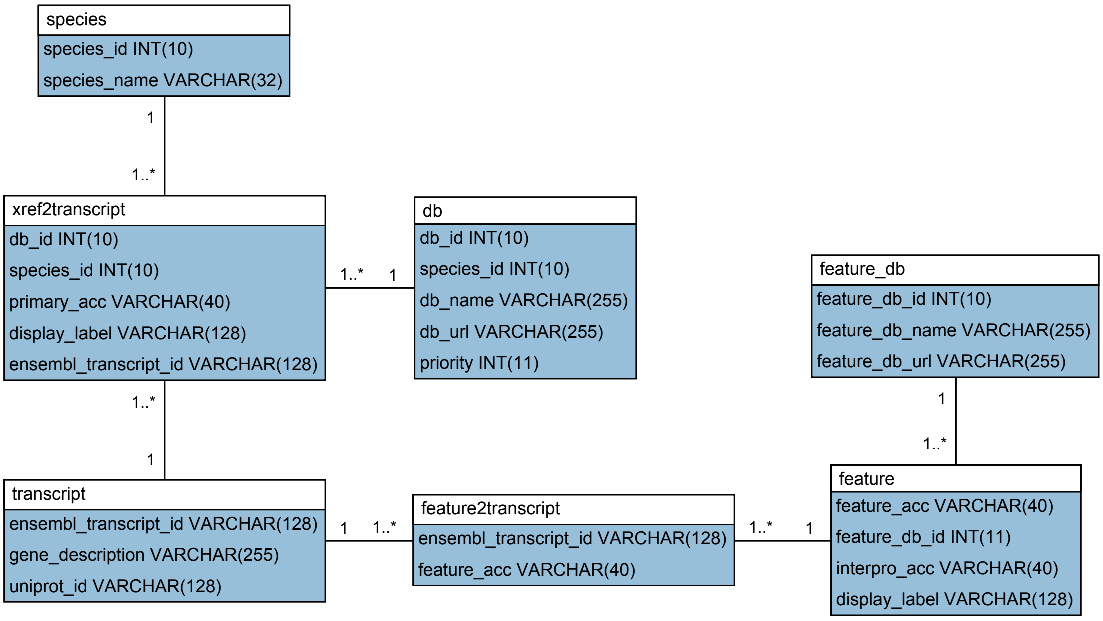


Figure 5.1: PhenoFam database scheme.

Tables `species` and `db` are dictionaries carrying information about identifiers available for the users-uploaded data. Table `xref2transcript` maps those external identifiers to ensembl transcript IDs (table `transcript`). Protein features (table `feature`) from InterPro member databases (table `feature_db`) are linked to Ensembl transcripts with the table `feature2transcript`.

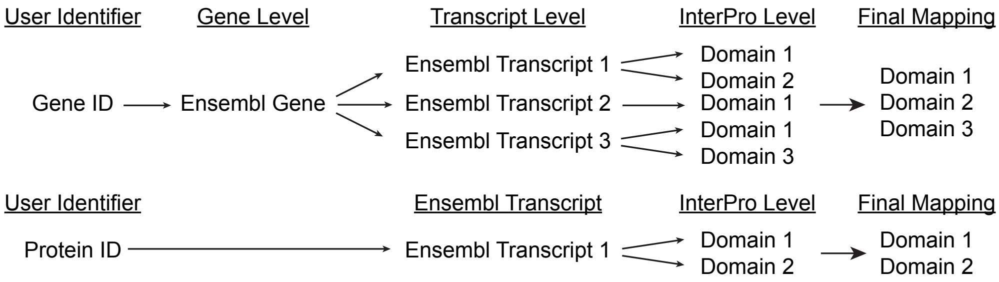


Figure 5.2: Identifier mapping procedure.

Gene-related identifiers (*e.g.* Gene Names) are mapped to Ensembl Gene IDs and further to all protein-coding Ensembl Transcript IDs. Each of the transcripts is associated with protein features from the InterPro database. Redundant identifiers are removed in the final mapping. Protein- or transcript-related identifiers (*e.g.* UniProt IDs) are directly linked to Ensembl Transcript IDs and then to protein features.

or several protein domains or features from different InterPro database members (Figure 5.2). Reversing the mapping, each protein domain is linked with at least one user identifier and at least one experimental value.

It must be noted that all identifier mappings are based on contents of the Ensembl database, which establishes the links based on sequence similarity of entities stored in remote databases to sequences stored in Ensembl. This approach provides the highest quality of associations. However, care must be taken if gene-related identifiers are used. Due to alternative splicing, different gene products may be composed of different protein domains or even encode different proteins (*i.e.* shift in the reading frame). In such cases, a value associated with the user-provided identifier is mapped to all possible protein features that can be associated with the gene (Figure 5.2).

5.2.3 Gene set enrichment analysis

To test if a set of values associated with a given domain is significantly higher or lower than the remaining set of values, I use the Mann–Whitney U-test. The U-test is the most powerful nonparametric alternative to the Student's t-test. Its main advantage is that it makes no assumptions about the underlying distributions and is more robust in case of outliers. The U-test is also implemented in other popular GSEA tools, *i.e.* GeneTrail (Keller *et al.*, 2008) or PANTHER (Thomas *et al.*, 2003, 2006). Other applications, such as GOdist (Ben-Shaul *et al.*, 2005) or GSEA (Subramanian *et al.*, 2005), implemented the Kolmogorov-Smirnov (KS) test, another non-parametric procedure that checks whether two samples (values associated with a given domain and the other values) may be assumed to come from the same distribution. However, the KS test is also sensitive to differences in the general shapes of the distributions, which limits its use for my PhenoFam application. Parametric analysis, which was proposed by Kim *et al.* and implemented in PAGE (Kim & Volsky, 2005), is also not suitable for GSEA of protein domains because many domains are associated with small number of proteins (< 10). In those cases, the normality criteria required for parametric tests might not be satisfied.

Adjustment for multiple testing is done using the false discovery rate (FDR) control procedure designed by Benjamini and Hochberg (Benjamini & Hochberg, 1995) and resulting q-values are obtained by applying Storey's algorithm (Storey, 2002; Storey & Tibshirani, 2003). Additionally, I calculate a Herrnstein's ρ statistic (Hernstein *et al.*, 1976), which is an unbiased measure of the overlap between distributions of values in the two compared sets. It can reach values between 0 and 1, where 0.5 indicates a complete overlap of the two distributions and both extreme values show a complete separation. This statistic shows how much a median of domain-associated values differs from a median of the other values, and together with the p-value can help identifying domains of interest. I recommend using it for sorting results that passed the significance-threshold criteria.

Due to the fact that InterPro is a collection of partially redundant databases, the enrichment analysis and the adjustment for multiple testing procedure are performed for each database independently. Otherwise, treating InterPro as a uniform set of annotations would lead to a significant underestimation of the results.

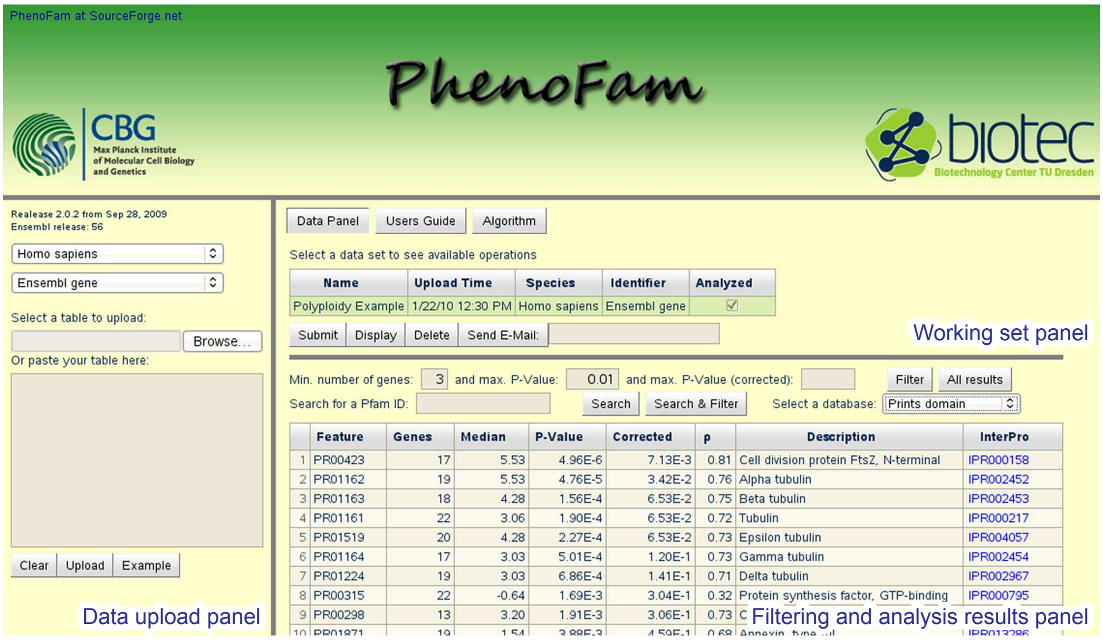
5.2.4 User interface

To implement the user interface and to ensure compatibility with all major browsers, I used the Google Web Toolkit (GWT) framework. I designed a simple and user-friendly data management system for storing uploaded data sets and the analysis results. It allows users to investigate and compare multiple data sets at the same time.

My GSEA algorithm reports the following information: a member database identifier, the domain description, a number of user identifiers associated with the domain, a median of the values, a p-value reported by the Mann–Whitney U-test, a FDR corrected p-value, a ρ statistic and the InterPro identifier. The results associated with one of the selected InterPro member databases are displayed in a pageable table (Figure 5.3) that can be sorted and filtered. I also provide a possibility to search for specific domains. For each selected domain, I also show a table of associated values together with original identifiers, UniProt accessions and descriptions.

5.3 Results

PhenoFam allows many data sets as the starting point, such as results of microarray studies, systematic RNA interference (RNAi) screens, ChIP-Chip/ChIP-Seq experiments or comparative mass-spectrometry (*i.e.* SILAC) results. To test the utility of PhenoFam, I analysed a data-set derived from a genome-scale cell cycle progression RNAi screen



Realease 2.0.2 from Sep 28, 2009
Ensembl release: 56

Homo sapiens
Ensembl gene

Select a table to upload:

Or paste your table here:

Select a data set to see available operations

Name	Upload Time	Species	Identifier	Analyzed
Polyploidy Example	1/22/10 12:30 PM	Homo sapiens	Ensembl gene	☑

Min. number of genes: 3 and max. P-Value: 0.01 and max. P-Value (corrected):
Search for a Pfam ID:
Select a database: Prints domain

	Feature	Genes	Median	P-Value	Corrected	p	Description	InterPro
1	PR00423	17	5.53	4.96E-6	7.13E-3	0.81	Cell division protein FtsZ, N-terminal	IPR000158
2	PR01162	19	5.53	4.76E-5	3.42E-2	0.76	Alpha tubulin	IPR002452
3	PR01163	18	4.28	1.56E-4	6.53E-2	0.75	Beta tubulin	IPR002453
4	PR01161	22	3.06	1.90E-4	6.53E-2	0.72	Tubulin	IPR000217
5	PR01519	20	4.28	2.27E-4	6.53E-2	0.73	Epsilon tubulin	IPR004057
6	PR01164	17	3.03	5.01E-4	1.20E-1	0.73	Gamma tubulin	IPR002454
7	PR01224	19	3.03	6.86E-4	1.41E-1	0.71	Delta tubulin	IPR002967
8	PR00315	22	-0.64	1.69E-3	3.04E-1	0.32	Protein synthesis factor, GTP-binding	IPR000795
9	PR00298	13	3.20	1.91E-3	3.06E-1	0.73	C	
10	PR01871	19	1.54	3.88E-3	4.59E-1	0.68	Annexin, type	IPR013286

Figure 5.3: PhenoFam web interface.

The main user interface display is divided into three panels. The 'Data upload panel' allows uploading data sets for the GSEA analysis either by pasting the data or by selecting a text file. All uploaded data sets are displayed in the 'Working set panel', where the user can submit data for the analysis, view the results in the browser or send them by e-mail. The sortable table with results is displayed in the 'Filtering and analysis results panel'. The top section of the panel contains a form that provides searching and filtering functionality. The displayed table contains a list of significantly enriched PRINTS domains.

(Kittler *et al.*, 2007a). In this screen, a genome-wide study of genes was carried out providing z-scores for cell cycle progression phenotypes (*i.e.* cells in G1, S, G2/M phases and polyploidy) for each knockdown.

A PhenoFam analysis of the complete RNAi data-set revealed that plexins containing a cytoplasmic RasGAP domain were enriched ($p < 0.005$) for polyploidy phenotypes (Figure 5.4A, Table 5.1). Knockdown of most transcripts encoding these genes resulted in an increase of polyploidy cells. Although in the published RNAi screen (Kittler *et al.*, 2007a) only genes with the strongest polyploidy phenotypes of z-score > 6 were selected for further investigation, the PhenoFam analysis suggests that plexins not passing this criteria might also have a function in cytokinesis.

Moreover, based on this result I predicted that knockdown of the gene PLXNB3, which belongs to the same family, but had not been tested in the screen, would also

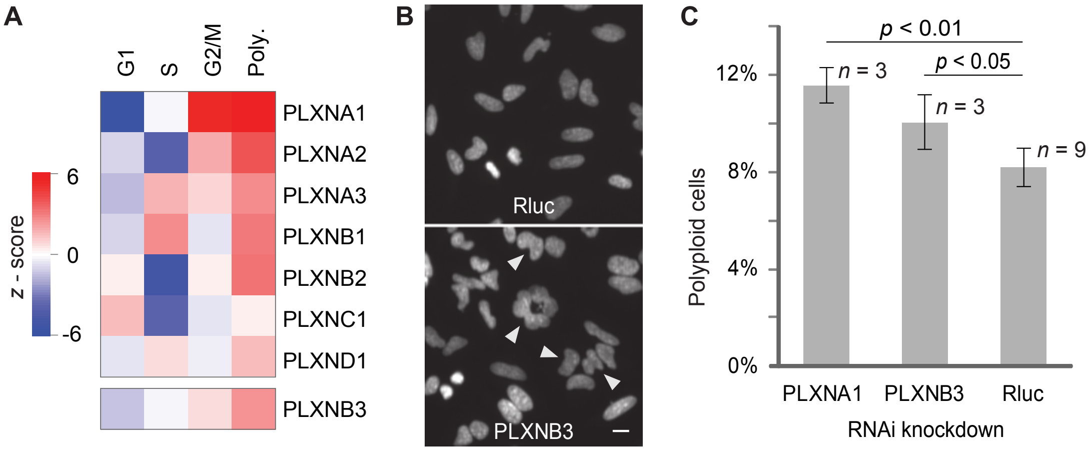


Figure 5.4: Plexins are enriched for polyploidy RNAi phenotypes.

(A) Cell cycle phenotypic profiles after knockdown of plexins. The top heatmap shows data extracted from the genome-wide RNAi screen (Kittler *et al.*, 2007a), and the bottom heatmap shows the cell cycle profile after knock-down of PLXNB3. The PLXNB3 profile was obtained from the automated analysis of microscopy images, which quantifies proportions of cells in different phases of the cell cycle. z-scores were calculated by normalization to the mean and standard deviation of respective values obtained from the analysis of negative control images. (B) Fluorescence microscopy images of HeLa cells 48 h after transfection of esiRNA (endoribonuclease-prepared siRNA) against Rluc (negative control) and PLXNB3. The images show DAPI-stained nuclei, and arrows indicate cells with polyploidy phenotype. The scale bar represents 10 µm. Investigation of both images shows that the knock-down of PLXNB3 results in a polyploidy phenotype compared to the negative control condition. (C) Quantification of the image analysis of polyploidy phenotypes among cells treated by different silencing triggers ($\approx$ 5000 cells per replicate). Error bars represent one SD. Student's t-test confirmed that the ratio of polyploid cells is significantly increased after knock-down of indicated plexins, compared to the control condition (Rluc).

Table 5.1: Normalized values of polyploidy RNAi phenotypes of plexins, from the primary cell cycle progression screen

Ensembl ID	Gene	Polyploidy (z-score)
ENSG00000114554	PLXNA1	13.15
ENSG00000076356	PLXNA2	4.05
ENSG00000130827	PLXNA3	2.72
ENSG00000164050	PLXNB1	3.14
ENSG00000196576	PLXNB2	3.33
ENSG00000004399	PLXND1	1.45
ENSG00000136040	PLXNC1	0.32

increase the degree of polyploidy. Indeed, an increased number of polyploid cells were measured after PLXNB3 knockdown (Figure 5.4), indicating that depletion of this gene, like other plexins with cytoplasmic RasGAP domains, influences proper cytokinesis. This example demonstrates that PhenoFam can be a valuable support for selecting hits from the RNAi screens.

To show that PhenoFam is also suitable for analysis of other large-scale data-sets, I examined publicly available gene expression data that compares transcriptomes of human breast carcinoma and healthy tissue (Cheng *et al.*, 2008). GSEA of this data-set using GeneTrail (Keller *et al.*, 2008) showed that genes whose expression is altered in breast cancer are significantly enriched with the 'signal transduction' and 'cell differentiation' gene ontologies, highlighting the importance of these biological processes during cellular transformation (data not shown). However, the analysis with GeneTrail did not provide information of enrichment of certain protein domains. In contrast, analysis of the same data-set with PhenoFam showed that among differentially expressed genes, Ras-family proteins and phox (PX) domain-containing proteins were enriched ($p < 0.001$, data not shown).

Ras GTPases are known to play a role in breast cancer development (Li & Sparano, 2003) and, therefore, it is not surprising that this group of proteins was enriched in this set. Proteins containing a PX domain are involved in cell signalling, vesicular trafficking, protein sorting and lipid modification, and are primarily found in sorting nexins (Worby & Dixon, 2002). Previous studies suggest that various sorting nexins are involved in leukemia (Fuchs *et al.*, 2001), colon tumorigenesis (Nguyen *et al.*, 2006) and, in general, contribute to cell cycle progression in mammalian cells (Fuster *et al.*, 2010). However, their role in breast cancer has not been described so far. This PhenoFam analysis suggests that proteins with PX domains are frequently misregulated in breast cancer. Hence, I propose that these proteins should be investigated for a possible role in breast cancer development.

Chapter 6

Discussion

It is clear that a systematic approach to all aspects of biology will play an increasingly important role. Complete reconstruction of regulatory networks in cells, tissues and organisms will be possible only through integration of various ‘omics’ data. Thus, further development of the data analysis methods, which utilize partial information provided by different experiments, is indespensable. However, the experience shows that development of a single and universal methodology, most probably, will not be feasible. Each biological question has to be treated individually and the analysis algorithms, as well as used data sets, have to be chosen accordingly. With an increasing number and volume of data sets used in analysis pipelines, the role of data management systems will increase, making the use of databases a critical part of any experimental workflow.

6.1 High-throughput data integration

DSViewer, as described in Chapter 2, is a flexible and user-friendly database of systematic RNAi screens. This web-accessible application organizes results of the primary genome-wide experiments, as well as data coming from secondary assays and final lists of hits annotated with confirmed phenotype labels. It proved to be especially useful for selecting hits and comparing outcomes of different screening projects.

DSViewer was implemented in J2EE technology, as a platform-independed web application that uses a freely-available MySQL relational database for data management. Relational databases are currently the most efficient systems for storing and querying information. They organize data in structured ways, in a form of fixed set of tables, where each table stores a fixed set of attributes describing physical objects or concepts. The main challenge of the database design was to elaborate a rigid scheme of tables that would be flexible enough to be capable of holding experimental results generated by different assays and algorithms and, in consequence, containing different sets of attributes (*i.e.* phenotypic read-outs and annotations) per entry.

This problem was solved by partial separation of attributes from data sets. By treating each assay-reported value and each annotation as properties of respective silencing triggers, it became possible to store all information in a single table. This approach allowed creating a centralised repository of RNAi libraries and systematic screenings. Moreover, the database scheme of DSViewer also allows storage of gene expression data, as well as any other gene-centered information. Thanks to the user interface of DSViewer, which allows data integration and flexible property filtering, a scientist can take advantage of data generated by various experiments and query for genes having very specific phenotypic and expression profiles.

6.2 Novel hit selection algorithms

Results of high-throughput RNAi experiments are a subject to high noise, and to keep a low false-positive rate, hit lists extracted from single data sets consist of the genes associated only with the strongest effects. Many of those genes are already well characterized 'hub proteins' that are involved in various, previously studied, cellular processes (Wang *et al.*, 2007). During the follow-up experiments, researchers focus on single or few uncharacterised genes, trying to uncover their molecular function in investigated biological process. In many cases, the other screening results remain unused. However, even the most elaborate assays are not able to measure neither complete cellular states, nor report information about dynamics of all cellular processes. Thus, a combination of data sets coming from different experiments can not only increase the power of hit selection procedures, but also make full use of previously generated results.

Chapter 3 presents a knowledge-directed hit selection methodology based on hierarchical clustering of two genome-wide RNAi screenings, a cell-cycle screen and a cell viability screen. The combined analysis of the two experiments identified many known and previously uncharacterised mitosis-related genes. A detailed experimental verification and characterisation of one of them, lead to a discovery that C13orf3 is a component of the Ska complex required for mammalian cell division.

Applied methodologies of non-parametric normalisation of phenotypes and variable weighting, which accounts for influence of information introduced by different experiments, can be extended to any set of high-throughput results, regardless of the underlying distribution of values. The hierarchical clustering, combined with cluster selection based on gene ontology enrichment analysis, proved to be a powerful tool for selecting genes for follow-up experiments performed in my laboratory.

Chapter 4 describes a novel algorithm for hits selection procedure based on a combined analysis of systematic RNAi screening results and time-course gene expression data. The aim of the RNAi screen was to identify genes involved in pluripotency maintenance of mouse embryonic stem cells. However, a differentiation phenotype upon silencing of a particular gene does not necessary imply its role in pluripotency maintenance. By introducing time-course microarray data, which measures changes of gene expression during embryonic body formation, it was possible to filter the phenotypic data for genes being downregulated upon differentiation, which suggests their relevance for the pluripotent and not for differentiated cells.

Recent study combining an RNAi screen performed in *C. elegans* with gene expression data (Mabon *et al.*, 2009), as well as results showed in this thesis, indicate that introducing additional experimental data in hit selection procedures increases specificity and reduces false positive rates, thereby refining lists of genes used for the follow-up studies. In perspective, integration of biological data obtained from different high-throughput experiments shall contribute to comprehensive picture of cellular biology.

6.3 Enrichment analysis

As described in Chapter 3 and 4, enrichment analysis is a powerful technique for descriptive analysis of sets of genes. It is usually used for examining results of high-throughput experiments. For example, a list of hits from RNAi screen investigating a certain biological process should be enriched with genes being characterized with annotation terms related do that process. Enrichment analysis result can also be used as an evidence supporting a biological hypothesis. If a protein is assumed to be a transcriptional regulator of genes involved in given cellular processes, then a list of genes whose expression is modified by altering abundance of the protein (*e.g.* by RNAi or over-expression) should also be enriched with genes related to these processes.

Although current functional annotations of genomes are still incomplete (*e.g.* in case of human genome, over 20 % of protein-coding genes are missing GO annotations), with continued annotation efforts, enrichment analysis tools will become more and more accurate. Moreover, assuming low error rate of already curated GO annotations, which, according to Jones *et al.* (2007), was between 13 % and 18 %, it is reasonable to conclude that currently calculated enrichment significance levels might be, in many cases, underestimated.

Enrichment analysis does not have to be limited to investigating overrepresentation of GO terms or pathway annotations among a set of genes. By performing analysis that uses *cis*-regulatory motifs or structural features of proteins as annotation terms, a researcher can obtain results complementing GO-based enrichment analysis and therefore achieve a higher coverage of the whole knowledge space. Moreover, in case of less annotated genomes, such as *X. tropicalis* or *D. rerio*, widely used vertebrate models for developmental biology, the GO annotations are currently mapped to only 3 % and 15 % of the protein coding genes, respectively. In cases like these, it becomes necessary to make use of information derived from sequence and structure homology.

Chapter 5 introduces PhenoFam, a novel application performing gene set enrichment analysis by employing structural and functional information on families of protein domains as annotation terms. Using a specific example, it was shown that the application can be used as an additional hit selection tool for functional screens. Typical hit selection procedures (*i.e.* z-score or quantile-based normalization) apply thresholds that can be passed only by genes showing the strongest phenotypes, which often leads to a high false-negatives rate. In case of the implemented GSEA method, a domain may appear to be significantly enriched despite moderate phenotypes of the associated genes. From the potential relationship between the domain and the investigated biological process, genes with moderate phenotypic scores are considered in the list of hits selected from the screen, thereby reducing the false-negative rate.

Notably, it was demonstrated that PhenoFam can help forming novel hypothesis based on gene expression data. Accordingly, PhenoFam should be useful in analysing results of other high-throughput experiments, such as ChIP-Chip/ChIP-Seq and comparative mass-spectrometry. Complementing other enrichment analysis tools, PhenoFam can assist in annotating genes of unknown function and in discovering new functions of already characterised genes. In addition, since the possibility of targeting protein-protein interactions with specific drugs raises expectations of huge impact in the therapeutics field (Fuentes *et al.*, 2009), protein domains that PhenoFam finds enriched in results of disease-related experiments might be of special interest for potential drug discovery efforts.

6.4 Outlook

To enable automatic and unambiguous interpretation of experimental results stored in web-accessible repositories, the data sets have to be accompanied by complete and machine-understandable descriptions of the experiments and data acquisition techniques. A standard for presenting and exchanging such information already exist for microarray experiments and is widely known as Minimum Information About a Microarray Experiment (MIAME) (Brazma *et al.*, 2001). Similar standards for cellular asssays, or RNAi experiments in particular, are in development (Brazma *et al.*, 2006) and future implementations of DSViewer, and similar software, should definitely implement them.

With current trends in application development, more and more databases and analysis tools become accessible on-line and newly developed pipelines may take an advantage of this fact. By creating distributed applications having different elements of the analysis algorithm spread all over the world, where each of them is developed and maintained by a group of experts in a particular field, we can utilize global technical and intellectual resources. Platforms that support creation of data analysis pipelines based on distributed resources are being currently developed, one good example is KNIME, the 'information miner' developed at the *University of Konstanz* (Berthold *et al.*, 2007). However, the main prerequisite for building global analysis networks is a proper implementation of individual web services, which have to provide a programmatic interface to their resources, such as Simple Object Access Protocol (SOAP).

6.5 Conclusions

The software applications described in this thesis were tailored to become a comprehensive, versatile and user-friendly tools for data management and analysis. Introduced data mining techniques proved to be useful to unveil meaningful biological information and demonstrated the necessity of heterogenous data integration.

Eventually, these investigations attempt to provide the research community with a markedly improved repertoire of computational tools and methods that facilitate the systematic analysis of accumulated information obtained from high-throughput studies into novel biological insights.

References

ADELMAN, K., WEI, W., ARDEHALI, M.B., WERNER, J., ZHU, B., REINBERG, D. & LIS, J.T. (2006). Drosophila Paf1 modulates chromatin structure at actively transcribed genes. *Mol. Cell. Biol.*, **26**, 250–260. 75, 76

AKANUMA, T., KOSHIDA, S., KAWAMURA, A., KISHIMOTO, Y. & TAKADA, S. (2007). Paf1 complex homologues are required for Notch-regulated transcription during somite segmentation. *EMBO Rep.*, **8**, 858–863. 86, 88

AL-SHAHROUR, F., MINGUEZ, P., TARRAGA, J., MEDINA, I., ALLOZA, E., MONTANER, D. & DOPAZO, J. (2007). FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic acids research*, **35**, W91–6. 96

ALEXA, A., RAHNENFÜHRER, J. & LENGAUER, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607. 20

ANASTASSIADIS, K., KIM, J., DAIGLE, N., SPRENGEL, R., SCHÖLER, H.R. & STEWART, A.F. (2002). A predictable ligand regulated expression strategy for stably integrated transgenes in mammalian cells in culture. *Gene*, **298**, 159–172. 93

ANTONOV, A.V. & MEWES, H.W. (2006). Complex functionality of gene groups identified from high-throughput data. *J. Mol. Biol.*, **363**, 289–296. 20

ANTONOV, A.V., SCHMIDT, T., WANG, Y. & MEWES, H.W. (2008). ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–351. 20

ARBEITMAN, M.N., FURLONG, E.E.M., IMAM, F., JOHNSON, E., NULL, B.H., BAKER, B.S., KRASNOW, M.A., SCOTT, M.P., DAVIS, R.W. & WHITE, K.P. (2002). Gene expression during the life cycle of Drosophila melanogaster. *Science (New York, N.Y.)*, **297**, 2270–5. 2

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29. 9, 96

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. & Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic acids research*, **31**, 400–2. 97

Aubert, J., Stavridis, M.P., Tweedie, S., O'Reilly, M., Vierlinger, K., Li, M., Ghazal, P., Pratt, T., Mason, J.O., Roy, D. & Smith, A. (2003). Screening for mammalian neural genes via fluorescence-activated cell sorter purification of neural precursors from Sox1-gfp knock-in mice. *Proc. Natl. Acad. Sci. U.S.A.*, **100 Suppl 1**, 11836–11841. 86

Bauer, S., Grossmann, S., Vingron, M. & Robinson, P.N. (2008). Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651. 20

Bauer, S., Gagneur, J. & Robinson, P.N. (2010). GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic acids research*. 20

Ben-Shaul, Y., Bergman, H. & Soreq, H. (2005). Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics (Oxford, England)*, **21**, 1129–37. 99

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate – a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289–300. 21, 99

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188. 21

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S.L. & Lander, E.S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326. 66, 91, 93

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K. & Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*, Springer. 109

Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Springer. 25

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C.J., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.K., Ooi, H.S., Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S.E., Fu, Y., Green, E.D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Cooper, G.M., Asimenos, G., Dewey, C.N., Hou, M., Nikolaev, S., Montoya-Burgos, J.I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N.R., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Sidow, A., Trinklein, N.D., Zhang, Z.D., Barrera, L., Stuart, R., King, D.C., Ameur, A., Enroth, S., Bieda, M.C., Kim, J., Bhinge, A.A., Jiang, N., Liu, J., Yao, F., Vega, V.B., Lee, C.W.H., Ng, P., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J.C., Couttet, P., Bruce, A.W., Dovey, O.M., Ellis, P.D., Langford, C.F., Nix, D.A., Euskirchen, G., Hartman, S., Urban, A.E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T.H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Aldred, S.F., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Zhang, X., Xu, M., Haidar, J.N.S., Yu, Y., Iyer, V.R., Green, R.D., Wadelius, C., Farnham, P.J., Ren, B., Harte, R.A., Hinrichs, A.S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A.S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R.M., Karolchik, D., Armengol, L., Bird, C.P., de Bakker, P.I.W., Kern, A.D., Lopez-Bigas, N., Martin, J.D., Stranger, B.E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I.B., Huppert, J., Zody, M.C., Abecasis, G.R., Estivill, X., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V.B., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. & de Jong, P.J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816. 3

Blais, A. & Dynlacht, B.D. (2005). Constructing transcriptional regulatory networks. *Genes & development*, **19**, 1499–511. 3

Bogue, M.A. & Grubb, S.C. (2004). The Mouse Phenome Project. *Genetica*, **122**, 71–74. 29

Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, **19**, 185–93. 17

Boutros, M., Brás, L.P. & Huber, W. (2006). Analysis of cell-based RNAi screens. *Genome biology*, **7**, R66. 10, 11

Boxem, M., Maliga, Z., Klitgord, N., Li, N., Lemmens, I., Mana, M., de Lichtervelde, L., Mul, J.D., van de Peut, D., Devos, M., Simonis, N., Yildirim, M.A., Cokol, M., Kao, H.L., de Smet, A.S., Wang, H., Schlaitz, A.L., Hao, T., Milstein, S., Fan, C., Tipsword, M., Drew, K., Galli, M., Rhrissorrakrai, K., Drechsel, D., Koller, D., Roth, F.P., Iakoucheva, L.M., Dunker, A.K., Bonneau, R., Gunsalus, K.C., Hill, D.E., Piano, F., Tavernier, J., van den Heuvel, S., Hyman, A.A. & Vidal, M. (2008). A protein domain-based interactome network for C. elegans early embryogenesis. *Cell*, **134**, 534–545. 55

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. & Sherlock, G. (2004). GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics (Oxford, England)*, **20**, 3710–5. 21

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. & Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, **29**, 365–71. 109

Brazma, A., Krestyaninova, M. & Sarkans, U. (2006). Standards for systems biology. *Nature reviews. Genetics*, **7**, 593–605. 109

Brill, L.M., Salomon, A.R., Ficarro, S.B., Mukherji, M., Stettler-Gill, M. & Peters, E.C. (2004). Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Anal. Chem.*, **76**, 2763–2772. 59

Buchholz, F., Kittler, R., Slabicki, M. & Theis, M. (2006). Enzymatically prepared RNAi libraries. *Nature Methods*, **3**, 696–700. 5, 36

Cantin, G.T., Yi, W., Lu, B., Park, S.K., Xu, T., Lee, J.D. & Yates, J.R. (2008). Combining protein-based IMAC, peptide-based IMAC, and MudPIT for efficient phosphoproteomic analysis. *J. Proteome Res.*, **7**, 1346–1351. 59

Carey, M.F., Peterson, C.L. & Smale, S.T. (2009). Chromatin immunoprecipitation (ChIP). *CSH protocols*, **2009**, pdb.prot5279. 3

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J.M. & Pascual-Montano, A. (2007). Genecodis: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*, **8**, R3. 20

Carpenter, L. & Zernicka-Goetz, M. (2004). Directing pluripotent cell differentiation using 'diced RNA' in transient transfection. *Genesis*, **40**, 157–163. 67

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P. & Workman, J.L. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, **123**, 581–592. 75

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J. & Williams, A.J. (2004). Unbiased Mapping of Transcription Factor Binding Sites along Human Chromosomes 21 and 22 Points to Widespread Regulation of Noncoding RNAs. *Cell*, **116**, 499–509. 18

Chambers, I. & Smith, A. (2004). Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene*, **23**, 7150–7160. 66

CHAMBERS, I., COLBY, D., ROBERTSON, M., NICHOLS, J., LEE, S., TWEEDIE, S. & SMITH, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, **113**, 643–655. 86

CHEESEMAN, I.M. & DESAI, A. (2005). A combined approach for the localization and tandem affinity purification of protein complexes from metazoans. *Sci. STKE*, **2005**, pl1. 46, 59, 79

CHEN, M., DU, Q., ZHANG, H.Y., WANG, X. & LIANG, Z. (2007). High-throughput screening using siRNA (RNAi) libraries. *Expert review of molecular diagnostics*, **7**, 281–91. 5

CHENG, A.S.L., CULHANE, A.C., CHAN, M.W.Y., VENKATARAMU, C.R., EHRICH, M., NASIR, A., RODRIGUEZ, B.A.T., LIU, J., YAN, P.S., QUACKENBUSH, J., NEPHEW, K.P., YEATMAN, T.J. & HUANG, T.H.M. (2008). Epithelial progeny of estrogen-exposed breast progenitor cells display a cancer-like methylome. *Cancer research*, **68**, 1786–96. 103

CHUNG, N., ZHANG, X.D., KREAMER, A., LOCCO, L., KUAN, P.F., BARTZ, S., LINSLEY, P.S., FERRER, M. & STRULOVICI, B. (2008). Median absolute deviation to improve hit selection for genome-scale RNAi screens. *Journal of biomolecular screening : the official journal of the Society for Biomolecular Screening*, **13**, 149–58. 13

COCHRANE, G.R. & GALPERIN, M.Y. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic acids research*, **38**, D1–4. 8

COLE, C., BARBER, J.D. & BARTON, G.J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.*, **36**, 197–201. 64

COLLAS, P. (2010). The Current State of Chromatin Immunoprecipitation. *Molecular biotechnology*. 3

COMA, I., CLARK, L., DIEZ, E., HARPER, G., HERRANZ, J., HOFMANN, G., LENNON, M., RICHMOND, N., VALMASEDA, M. & MACARRON, R. (2009). Process validation and screen reproducibility in high-throughput screening. *Journal of biomolecular screening : the official journal of the Society for Biomolecular Screening*, **14**, 66–76. 11

CORPET, F., SERVANT, F., GOUZY, J. & KAHN, D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic acids research*, **28**, 267–9. 97

COSTA, P.J. & ARNDT, K.M. (2000). Synthetic lethal interactions suggest a role for the Saccharomyces cerevisiae Rtf1 protein in transcription elongation. *Genetics*, **156**, 535–547. 75, 88

CROCKER, L. & ALGINA, J. (2006). *Introduction to classical and modern test theory*. Thomson Wadsworth. 42

DENNIS, G., SHERMAN, B.T., HOSACK, D.A., YANG, J., GAO, W., LANE, H.C. & LEMPICKI, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology*, **4**, P3. 96

DERISI, J.L., IYER, V.R. & BROWN, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science (New York, N.Y.)*, **278**, 680–6. 2

DING, L., PASZKOWSKI-ROGACZ, M., NITZSCHE, A., SLABICKI, M.M., HENINGER, A.K., DE VRIES, I., KITTLER, R., JUNQUEIRA, M., SHEVCHENKO, A., SCHULZ, H., HUBNER, N., DOSS, M.X., SACHINIDIS, A., HESCHELER, J., IACONE, R., ANASTASSIADIS, K., STEWART, A.F., PISABARRO, M.T., CALDARELLI, A., POSER, I., THEIS, M. & BUCHHOLZ, F. (2009). A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell stem cell*, **4**, 403–15. 36

DIOGO, M.M., HENRIQUE, D. & CABRAL, J.M. (2008). Optimization and integration of expansion and neural commitment of mouse embryonic stem cells. *Biotechnol. Appl. Biochem.*, **49**, 105–112. 86

DO, J.H. & CHOI, D.K. (2006). Normalization of microarray data: single-labeled and dual-labeled arrays. *Molecules and cells*, **22**, 254–61. 18

DRYSDALE, R. (2008). FlyBase : a database for the Drosophila research community. *Methods in molecular biology (Clifton, N.J.)*, **420**, 45–59. 29

DURINCK, S., SPELLMAN, P.T., BIRNEY, E. & HUBER, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, **4**, 1184–91. 30

FAZZIO, T.G., HUFF, J.T. & PANNING, B. (2008). An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity. *Cell*, **134**, 162–174. 39, 67, 82

FINN, R.D., MISTRY, J., TATE, J., COGGILL, P., HEGER, A., POLLINGTON, J.E., GAVIN, O.L., GUNASEKARAN, P., CERIC, G., FORSLUND, K., HOLM, L., SONNHAMMER, E.L.L., EDDY, S.R. & BATEMAN, A. (2010). The Pfam protein families database. *Nucleic acids research*, **38**, D211–22. 97

FIRE, A., XU, S., MONTGOMERY, M.K., KOSTAS, S.A., DRIVER, S.E. & MELLO, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**, 806–11. 3

Flockhart, I., Booker, M., Kiger, A., Boutros, M., Armknecht, S., Ramadan, N., Richardson, K., Xu, A., Perrimon, N. & Mathey-Prevot, B. (2006). FlyRNAi: the Drosophila RNAi screening center database. *Nucleic acids research*, **34**, D489–94. 29

Foltz, G., Yoon, J.G., Lee, H., Ryken, T.C., Sibenaller, Z., Ehrich, M., Hood, L. & Madan, A. (2009). DNA methyltransferase-mediated transcriptional silencing in malignant glioma: a combined whole-genome microarray and promoter array analysis. *Oncogene*, **28**, 2667–77. 7

Fouse, S.D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R. & Fan, G. (2008). Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell*, **2**, 160–169. 70

Fridmann-Sirkis, Y., Kent, H.M., Lewis, M.J., Evans, P.R. & Pelham, H.R. (2006). Structural analysis of the interaction between the SNARE Tlg1 and Vps51. *Traffic*, **7**, 182–190. 57, 64

Fuchs, U., Rehkamp, G., Haas, O.A., Slany, R., Kōnig, M., Bojesen, S., Bohle, R.M., Damm-Welk, C., Ludwig, W.D., Harbott, J. & Borkhardt, A. (2001). The human formin-binding protein 17 (FBP17) interacts with sorting nexin, SNX2, and is an MLL-fusion partner in acute myelogeneous leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 8756–61. 103

Fuentes, G., Oyarzabal, J. & Rojas, A.M. (2009). Databases of protein-protein interactions and their use in drug discovery. *Current opinion in drug discovery & development*, **12**, 358–66. 109

Fuster, J.J., González, J.M., Edo, M.D., Viana, R., Boya, P., Cervera, J., Verges, M., Rivera, J. & Andrés, V. (2010). Tumor suppressor p27Kip1 undergoes endolysosomal degradation through its interaction with sorting nexin 6. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. 104

Gagarin, A., Makarenkov, V. & Zentilli, P. (2006). Using clustering techniques to improve hit selection in high-throughput screening. *Journal of biomolecular screening : the official journal of the Society for Biomolecular Screening*, **11**, 903–14. 16

Galvez, T., Teruel, M.N., Heo, W.D., Jones, J.T., Kim, M.L., Liou, J., Myers, J.W. & Meyer, T. (2007). siRNA screen of the human signaling proteome identifies the PtdIns(3,4,5)P3-mTOR signaling pathway as a primary regulator of transferrin uptake. *Genome Biol.*, **8**, R142. 39

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. & Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636. 57

Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G. & Wüthrich, K. (1994). Homeodomain-DNA recognition. *Cell*, **78**, 211–223. 96

Glatter, T., Wepf, A., Aebersold, R. & Gstaiger, M. (2009). An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol. Syst. Biol.*, **5**, 237. 55

Godzik, A. (2003). Fold recognition methods. *Methods Biochem Anal*, **44**, 525–546. 47

Goto, H., Tomono, Y., Ajiro, K., Kosako, H., Fujita, M., Sakurai, M., Okawa, K., Iwamatsu, A., Okigaki, T., Takahashi, T. & Inagaki, M. (1999). Identification of a novel phosphorylation site on histone H3 coupled with mitotic chromosome condensation. *J. Biol. Chem.*, **274**, 25543–25549. 49

Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of molecular biology*, **313**, 903–19. 97

Grützmann, R., Boriss, H., Ammerpohl, O., Lüttges, J., Kalthoff, H., Schackert, H.K., Klöppel, G., Saeger, H.D. & Pilarsky, C. (2005). Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, **24**, 5079–88. 7

Gunsalus, K.C., Yueh, W.C., MacMenamin, P. & Piano, F. (2004). RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic acids research*, **32**, D406–10. 29

Haft, D.H., Selengut, J.D. & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic acids research*, **31**, 371–3. 97

Hahne, F., Mehrle, A., Arlt, D., Poustka, A., Wiemann, S. & Beissbarth, T. (2008). Extending pathways based on gene lists using InterPro domain signatures. *BMC bioinformatics*, **9**, 3. 97

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P. & Kasprzyk, A. (2009). BioMart Central Portal–unified access to biological data. *Nucleic acids research*, **37**, W23–7. 10

Hanisch, A., Silljé, H.H. & Nigg, E.A. (2006). Timely anaphase onset requires a novel spindle and kinetochore complex comprising Ska1 and Ska2. *EMBO J.*, **25**, 5504–5515. 40, 46, 47, 50

Hay, D.C., Sutherland, L., Clark, J. & Burdon, T. (2004). Oct-4 knockdown induces similar patterns of endoderm and trophoblast differentiation markers in human and mouse embryonic stem cells. *Stem Cells*, **22**, 225–235. 80

He, S., Zhang, D., Cheng, F., Gong, F. & Guo, Y. (2009). Applications of RNA interference in cancer therapeutics as a powerful tool for suppressing gene expression. *Molecular biology reports*, **36**, 2153–63. 6

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J.E., Snesrud, E., Lee, N. & Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. *BioTechniques*, **29**, 548–50, 552–4, 556 passim. 2

Hernstein, R.J., Loveland, D.H. & Cable, C. (1976). Natural concepts in pigeons. *J Exp Psychol Anim Behav Process*, **2**, 285–302. 99

Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M. & Snyder, M. (2002). GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 2924–9. 3

Horn, T., Arziman, Z., Berger, J. & Boutros, M. (2007). GenomeRNAi: a database for cell-based RNAi phenotypes. *Nucleic acids research*, **35**, D492–7. 29

Huang, d.a.W., Sherman, B.T. & Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13. 19, 96

Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. & Flicek, P. (2009). Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–697. 8, 32, 97

Hubbell, E., Liu, W.M. & Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics (Oxford, England)*, **18**, 1585–92. 16

Huber, W., von Heydebreck, A., Sueltmann, H., Poustka, A. & Vingron, M. (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical applications in genetics and molecular biology*, **2**, Article3. 16

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S. & O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691. 57

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, P.S., Corinne a nd Langendijk-Genevaux & Sigrist, C.J.A. (2008). The 20 years of PROSITE. *Nucleic acids research*, **36**, D245–9. 97

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daug herty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugra ud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistr y, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J.A., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. & Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic acids research*, **37**, D211–5. 8, 97

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. & Speed, T.P. (2003a). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, **31**, e15. 16

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. & Speed, T.P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, **4**, 249–64. 16

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y. & Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. *Nature*, **442**, 533–538. 66

Jaenisch, R. & Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, **132**, 567–582. 87

Johnson, R. (2005). J2EE development frameworks. *Computer*, **38**, 107–110. 30

JOHNSON, W.E., LI, W., MEYER, C.A., GOTTARDO, R., CARROLL, J.S., BROWN, M. & LIU, X.S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12457–62. 18

JONES, C.E., BROWN, A.L. & BAUMANN, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. *BMC bioinformatics*, **8**, 170. 108

KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M. & HIRAKAWA, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, **38**, D355–60. 96

KAROLCHIK, D., HINRICHS, A.S. & KENT, W.J. (2009). The UCSC Genome Browser. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, **Chapter 1**, Unit1.4. 19

KASPRZYK, A., KEEFE, D., SMEDLEY, D., LONDON, D., SPOONER, W., MELSOPP, C., HAMMOND, M., ROCCA-SERRA, P., COX, T. & BIRNEY, E. (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome research*, **14**, 160–9. 10

KEEGAN, K., JOHNSON, D.E., WILLIAMS, L.T. & HAYMAN, M.J. (1991). Isolation of an additional member of the fibroblast growth factor receptor family, FGFR-3. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 1095–9. 71

KELLER, A., BACKES, C., AL-AWADHI, M., GERASCH, A., KÜNTZER, J., KOHLBACHER, O., KAUFMANN, M. & LENHOF, H.P. (2008). GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. *BMC bioinformatics*, **9**, 552. 20, 96, 99, 103

KELLER, G. (2005). Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes Dev.*, **19**, 1129–1155. 66

KHATRI, P., DRAGHICI, S., OSTERMEIER, G.C. & KRAWETZ, S.A. (2002). Profiling gene expression using onto-express. *Genomics*, **79**, 266–70. 96

KIM, S.Y. & VOLSKY, D. (2005). PAGE: parametric analysis of gene set enrichment. *BMC bioinformatics*, **6**, 144. 99

KITTLER, R., PUTZ, G., PELLETIER, L., POSER, I., HENINGER, A.K., DRECHSEL, D., FISCHER, S., KONSTANTINOVA, I., HABERMANN, B., GRABNER, H., YASPO, M.L., HIMMELBAUER, H., KORN, B., NEUGEBAUER, K., PISABARRO, M.T. & BUCHHOLZ, F. (2004). An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. *Nature*, **432**, 1036–1040. 39, 67

Kittler, R., Heninger, A.K., Franke, K., Habermann, B. & Buchholz, F. (2005a). Production of endoribonuclease-prepared short interfering RNAs for gene silencing in mammalian cells. *Nat. Methods*, **2**, 779–784. 59, 67, 90

Kittler, R., Pelletier, L., Ma, C., Poser, I., Fischer, S., Hyman, A.A. & Buchholz, F. (2005b). RNA interference rescue by bacterial artificial chromosome transgenesis in mammalian tissue culture cells. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2396–2401. 46, 76, 79, 92

Kittler, R., Pelletier, L., Heninger, A.K., Slabicki, M., Theis, M., Miroslaw, L., Poser, I., Lawo, S., Grabner, H.., Kozak, K., Wagner, J., Surendranath, V., Richter, C., Bowen, W., Jackson, A.L., Habermann, B., Hyman, A.A. & Buchholz, F. (2007a). Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nat. Cell Biol.*, **9**, 1401–1412. 101, 102

Kittler, R., Pelletier, L., Heninger, A.K., Slabicki, M., Theis, M., Miroslaw, L., Poser, I., Lawo, S., Grabner, H., Kozak, K., Wagner, J., Surendranath, V., Richter, C., Bowen, W., Jackson, A.L., Habermann, B., Hyman, A.A. & Buchholz, F. (2007b). Genome-scale RNAi profiling of cell division in human tissue culture cells. *Nat. Cell Biol.*, **9**, 1401–1412. 36, 39, 40, 41, 42, 43, 67

Kittler, R., Surendranath, V., Heninger, A.K., Slabicki, M., Theis, M., Putz, G., Franke, K., Caldarelli, A., Grabner, H., Kozak, K., Wagner, J., Rees, E., Korn, B., Frenzel, C., Sachse, C., Sönnichsen, B., Guo, J., Schelter, J., Burchard, J., Linsley, P.S., Jackson, A.L., Habermann, B. & Buchholz, F. (2007c). Genome-wide resources of endoribonuclease-prepared short interfering RNAs for specific loss-of-function studies. *Nat. Methods*, **4**, 337–344. 39, 59, 67

Kittler, R., Pelletier, L. & Buchholz, F. (2008). Systems biology of mammalian cell division. *Cell Cycle*, **7**, 2123–2128. 39

Krogan, N.J., Dover, J., Wood, A., Schneider, J., Heidt, J., Boateng, M.A., Dean, K., Ryan, O.W., Golshani, A., Johnston, M., Greenblatt, J.F. & Shilatifard, A. (2003). The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Mol. Cell*, **11**, 721–729. 75, 80, 88, 89

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., LAM, M.H., BUTLAND, G., ALTAF-UL, A.M., KANAYA, S., SHILATIFARD, A., O'SHEA, E., WEISSMAN, J.S., INGLES, C.J., HUGHES, T.R., PARKINSON, J., GERSTEIN, M., WODAK, S.J., EMILI, A. & GREENBLATT, J.F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643. 57

LARSEN, N.A., AL-BASSAM, J., WEI, R.R. & HARRISON, S.C. (2007). Structural analysis of Bub3 interactions in the mitotic spindle checkpoint. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 1201–1206. 47, 64

LEE, J.H. & SKALNIK, D.G. (2005). CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J. Biol. Chem.*, **280**, 41725–41731. 82

LEHMANN, E. & ROMANO, J.P. (2005). *Testing statistical hypotheses*. Springer; 3rd edition. 21

LETUNIC, I., GOODSTADT, L., DICKENS, N.J., DOERKS, T., SCHULTZ, J., MOTT, R., CICCARELLI, F., COPL EY, R.R., PONTING, C.P. & BORK, P. (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic acids research*, **30**, 242–4. 97

LI, T. & SPARANO, J.A. (2003). Inhibiting Ras signaling in the therapy of breast cancer. *Clinical breast cancer*, **3**, 405–16; discussion 417–20. 103

LIMA, T., AUCHINCLOSS, A.H., COUDERT, E., KELLER, G., MICHOUD, K., RIVOIRE, C., BULLIARD, V., DE CASTRO, E., LACHAIZE, C., BARATIN, D., PHAN, I., BOUGUELERET, L. & BAIROCH, A. (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic acids research*, **37**, D471–8. 97

LOCKHART, D.J. & WINZELER, E.A. (2000). Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–36. 2

LOH, Y.H., WU, Q., CHEW, J.L., VEGA, V.B., ZHANG, W., CHEN, X., BOURQUE, G., GEORGE, J., LEONG, B., LIU, J., WONG, K.Y., SUNG, K.W., LEE, C.W., ZHAO, X.D., CHIU, K.P., LIPOVICH, L., KUZNETSOV, V.A., ROBSON, P., STANTON, L.W., WEI, C.L., RUAN, Y., LIM, B. & NG, H.H. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440. 70, 79

LU, R., MARKOWETZ, F., UNWIN, R.D., LEEK, J.T., AIROLDI, E.M., MACARTHUR, B.D., LACHMANN, A., ROZOV, R., MA/'AYAN, A., BOYER, L.A., TROYANSKAYA, O.G., WHETTON, A.D. & LEMISCHKA, I.R. (2009). Systems-level dynamic analyses of fate change in murine embryonic stem cells. *Nature*, **462**, 358–62. 2

Lyskov, S. & Gray, J.J. (2008). The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.*, **36**, W233–238. 64

Mabon, M.E., Mao, X., Jiao, Y., Scott, B.A. & Crowder, C.M. (2009). Systematic identification of gene activities promoting hypoxic death. *Genetics*, **181**, 483–96. 107

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L.M.L. Cam & J. Neyman, eds., *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, 281–297, University of California Press, Berkeley. 25

Mailhes, J.B., Hilliard, C., Fuseler, J.W. & London, S.N. (2003). Okadaic acid, an inhibitor of protein phosphatase 1 and 2A, induces premature separation of sister chromatids during meiosis I and aneuploidy in mouse oocytes in vitro. *Chromosome Res.*, **11**, 619–631. 55

Masaki, H., Nishida, T., Kitajima, S., Asahina, K. & Teraoka, H. (2007). Developmental pluripotency-associated 4 (DPPA4) localized in active chromatin inhibits mouse embryonic stem cell differentiation into a primitive ectoderm lineage. *The Journal of biological chemistry*, **282**, 33034–42. 71

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A.n., Doremieux, O., Campbell, M.J., Kitano, H. & Thomas, P.D. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic acids research*, **33**, D284–8. 97

Mi, H., Guo, N., Kejariwal, A. & Thomas, P.D. (2007). PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic acids research*, **35**, D247–52. 9

Michael, S., Travé, G., Ramu, C., Chica, C. & Gibson, T.J. (2008). Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. *Bioinformatics*, **24**, 453–457. 57

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E. & Ecker, J.R. (2005). Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**, 1–15. 4

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. & Groop, L.C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273. 20, 96

MOSIMANN, C., HAUSMANN, G. & BASLER, K. (2006). Parafibromin/Hyrax activates Wnt/Wg target gene transcription by direct association with beta-catenin/Armadillo. *Cell*, **125**, 327–341. 88

MUELLER, C.L. & JAEHNING, J.A. (2002). Ctr9, Rtf1, and Leo1 are components of the Paf1/RNA polymerase II complex. *Mol. Cell. Biol.*, **22**, 1971–1980. 88

MUSACCHIO, A. & SALMON, E.D. (2007). The spindle-assembly checkpoint in space and time. *Nat. Rev. Mol. Cell Biol.*, **8**, 379–393. 38

MUYRERS, J.P., ZHANG, Y. & STEWART, A.F. (2001). Techniques: Recombinogenic engineering–new options for cloning and manipulating DNA. *Trends Biochem. Sci.*, **26**, 325–331. 39

MYERS, L.C. & KORNBERG, R.D. (2000). Mediator of transcriptional regulation. *Annu. Rev. Biochem.*, **69**, 729–749. 84

NAKAJIMA, M., KUMADA, K., HATAKEYAMA, K., NODA, T., PETERS, J.M. & HIROTA, T. (2007). The complete removal of cohesin from chromosome arms depends on separase. *J. Cell. Sci.*, **120**, 4188–4196. 53

NAM, D., KIM, S.B., KIM, S.K., YANG, S., KIM, S.Y. & CHU, I.S. (2006). ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics*, **22**, 2249–2253. 20

NASMYTH, K. (2002). Segregating sister genomes: the molecular biology of chromosome separation. *Science*, **297**, 559–565. 38

NÈGRE, N., LAVROV, S., HENNETIN, J., BELLIS, M. & CAVALLI, G. (2006). Mapping the Distribution of Chromatin Proteins by ChIP on Chip. *Methods in Enzymology*, **410**, 316–341. 4

NG, H.H., DOLE, S. & STRUHL, K. (2003a). The Rtf1 component of the Paf1 transcriptional elongation complex is required for ubiquitination of histone H2B. *J. Biol. Chem.*, **278**, 33625–33628. 75

NG, H.H., ROBERT, F., YOUNG, R.A. & STRUHL, K. (2003b). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol. Cell*, **11**, 709–719. 75, 89

NGUYEN, L.N., HOLDREN, M.S., NGUYEN, A.P., FURUYA, M.H., BIANCHINI, M., LEVY, E., MORDOH, J., LIU, A., GUNCAY, G.D., CAMPBELL, J.S. & PARKS, W.T. (2006). Sorting nexin 1 down-regulation promotes colon tumorigenesis. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **12**, 6952–9. 103

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Schöler, H. & Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*, **95**, 379–391. 67

Nigg, E.A. (2001). Mitotic kinases as regulators of cell division and its checkpoints. *Nat. Rev. Mol. Cell Biol.*, **2**, 21–32. 38

Niwa, H., Miyazaki, J. & Smith, A.G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.*, **24**, 372–376. 67

Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J.M. & Pascual-Montano, A. (2009). GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.*, **37**, W317–322. 20

Nousiainen, M., Silljé, H.H., Sauer, G., Nigg, E.A. & Körner, R. (2006). Phosphoproteome analysis of the human mitotic spindle. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 5391–5396. 59

O'Rourke, K. (2007). An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*, **100**, 579–82. 7

O'Shea, K.S. (2004). Self-renewal vs. differentiation of mouse embryonic stem cells. *Biol. Reprod.*, **71**, 1755–1765. 66

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J. & Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes & development*, **16**, 948–58. 5

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R.L., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J. & Orengo, C. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic acids research*, **33**, D247–51. 97

Penheiter, K.L., Washburn, T.M., Porter, S.E., Hoffman, M.G. & Jaehning, J.A. (2005). A posttranscriptional role for the yeast Paf1-RNA polymerase II complex is revealed by identification of primary targets. *Mol. Cell*, **20**, 213–223. 75, 88

Pepperkok, R. & Ellenberg, J. (2006). High-throughput fluorescence microscopy for systems biology. *Nat. Rev. Mol. Cell Biol.*, **7**, 690–696. 39

PESCE, M. & SCHÖLER, H.R. (2001). Oct-4: gatekeeper in the beginnings of mammalian development. *Stem Cells*, **19**, 271–278. 67

PETERS, J.M. (2006). The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nat. Rev. Mol. Cell Biol.*, **7**, 644–656. 38

PIETERSEN, A.M. & VAN LOHUIZEN, M. (2008). Stem cell regulation by polycomb repressors: postponing commitment. *Curr. Opin. Cell Biol.*, **20**, 201–207. 66

POSER, I., SAROV, M., HUTCHINS, J.R., HÉRICHÉ, J.K., TOYODA, Y., POZNIAKOVSKY, A., WEIGL, D., NITZSCHE, A., HEGEMANN, B., BIRD, A.W., PELLETIER, L., KITTLER, R., HUA, S., NAUMANN, R., AUGSBURG, M., SYKORA, M.M., HOFEMEISTER, H., ZHANG, Y., NASMYTH, K., WHITE, K.P., DIETZEL, S., MECHTLER, K., DURBIN, R., STEWART, A.F., PETERS, J.M., BUCHHOLZ, F. & HYMAN, A.A. (2008). BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods*, **5**, 409–415. 39, 40, 46, 57, 59, 63, 79, 93

PUNTERVOLL, P., LINDING, R., GEMÜND, C., CHABANIS-DAVIDSON, S., MATTINGSDAL, M., CAMERON, S., MARTIN, D.M., AUSIELLO, G., BRANNETTI, B., COSTANTINI, A., FERRÈ, 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., KÜSTER, B., HELMER-CITTERICH, M., HUNTER, W.N., AASLAND, R. & GIBSON, T.J. (2003). ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630. 64

RAMASAMY, A., MONDRY, A., HOLMES, C.C. & ALTMAN, D.G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine*, **5**, e184. 7

REN, B., ROBERT, F., WYRICK, J.J., APARICIO, O., JENNINGS, E.G., SIMON, I., ZEITLINGER, J., SCHREIBER, J., HANNETT, N., KANIN, E., VOLKERT, T.L., WILSON, C.J., BELL, S.P. & YOUNG, R.A. (2000). Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y.)*, **290**, 2306–9. 3

RHODES, D.R., YU, J., SHANKER, K., DESHPANDE, N., VARAMBALLY, R., GHOSH, D., BARRETTE, T., PANDEY, A. & CHINNAIYAN, A.M. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9309–14. 7

RICE, T.K., SCHORK, N.J. & RAO, D.C. (2008). *Genetic Dissection of Complex Traits*, vol. 60 of *Advances in Genetics*. Elsevier. 21

Riedel, C.G., Katis, V.L., Katou, Y., Mori, S., Itoh, T., Helmhart, W., Gálová, M., Petronczki, M., Gregan, J., Cetin, B., Mudrak, I., Ogris, E., Mechtler, K., Pelletier, L., Buchholz, F., Shirahige, K. & Nasmyth, K. (2006). Protein phosphatase 2A protects centromeric sister chromatid cohesion during meiosis I. *Nature*, **441**, 53–61. 53

Rines, D.R., Gomez-Ferreria, M.A., Zhou, Y., DeJesus, P., Grob, S., Batalov, S., Labow, M., Huesken, D., Mickanin, C., Hall, J., Reinhardt, M., Natt, F., Lange, J., Sharp, D.J., Chanda, S.K. & Caldwell, J.S. (2008). Whole genome functional analysis identifies novel components required for mitotic spindle integrity in human cells. *Genome Biol.*, **9**, R44. 46

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65. 27

Rush, J., Moritz, A., Lee, K.A., Guo, A., Goss, V.L., Spek, E.J., Zhang, H., Zha, X.M., Polakiewicz, R.D. & Comb, M.J. (2005). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.*, **23**, 94–101. 59

Sachse, C. & Echeverri, C.J. (2004). Oncology studies using siRNA libraries: the dawn of RNAi-based genomics. *Oncogene*, **23**, 8384–8391. 39

Sartor, M.A., Mahavisno, V., Keshamouni, V.G., Cavalcoli, J., Wright, Z., Karnovsky, A., Kuick, R.o., Jagadish, H., Mirel, B., Weymouth, T., Athey, B. & Omenn, G.S. (2009). ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics (Oxford, England).* 96

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, **270**, 467–70. 2

Shi, X., Chang, M., Wolf, A.J., Chang, C.H., Frazer-Abel, A.A., Wade, P.A., Burton, Z.F. & Jaehning, J.A. (1997). Cdc73p and Paf1p are found in a novel RNA polymerase II-containing complex distinct from the Srbp-containing holoenzyme. *Mol. Cell. Biol.*, **17**, 1160–1169. 88

Silva, G.L., Junta, C.M., Mello, S.S., Garcia, P.S., Rassi, D.M., Sakamoto-Hojo, E.T., Donadi, E.A. & Passos, G.A.S. (2007). Profiling meta-analysis reveals primarily gene coexpression concordance between systemic lupus erythematosus and rheumatoid arthritis. *Annals of the New York Academy of Sciences*, **1110**, 33–46. 7

Silva, J. & Smith, A. (2008). Capturing pluripotency. *Cell*, **132**, 532–536. 66

SIMS, D., BURSTEINAS, B., GAO, Q., ZVELEBIL, M. & BAUM, B. (2006). FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets. *Nucleic acids research*, **34**, D479–83. 29

SIPPL, M.J. & FLÖCKNER, H. (1996). Threading thrills and threats. *Structure*, **4**, 15–19. 47

SLONIM, D.K. & YANAI, I. (2009). Getting started in gene expression microarray analysis. *PLoS computational biology*, **5**, e1000543. 3, 17

SMEDLEY, D., HAIDER, S., BALLESTER, B., HOLLAND, R., LONDON, D., THORISSON, G. & KASPRZYK, A. (2009). BioMart–biological queries made easy. *BMC genomics*, **10**, 22. 10

STOLINSKI, L.A., EISENMANN, D.M. & ARNDT, K.M. (1997). Identification of RTF1, a novel gene important for TATA site selection by TATA box-binding protein in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **17**, 4490–4500. 75, 88

STOREY, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 479–498. 99

STOREY, J.D. & TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–5. 99

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V.K., MUKHERJEE, S., EBERT, B.L., GILLETTE, M.A., PAULOVICH, A.D., POMEROY, S.L., GOLUB, T.R., LANDER, E.S. & MESIROV, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545–15550. 20, 96, 99

SUI, Y. & WU, Z. (2007). Alternative statistical parameter for high-throughput screening assay quality assessment. *Journal of biomolecular screening : the official journal of the Society for Biomolecular Screening*, **12**, 229–34. 11

SUI, Y., ZHAO, X., SPEED, T.P. & WU, Z. (2009). Background Adjustment for DNA Microarrays Using a Database of Microarray Experiments. *Journal of Computational Biology*, **16**, 1501–1515. 16

SULLIVAN, M. & MORGAN, D.O. (2007). Finishing mitosis, one step at a time. *Nat. Rev. Mol. Cell Biol.*, **8**, 894–903. 38, 50, 53

TAKAHASHI, K., MURAKAMI, M. & YAMANAKA, S. (2005). Role of the phosphoinositide 3-kinase pathway in mouse embryonic stem (ES) cells. *Biochem. Soc. Trans.*, **33**, 1522–1525. 66

Tenney, K., Gerber, M., Ilvarsonn, A., Schneider, J., Gause, M., Dorsett, D., Eissenberg, J.C. & Shilatifard, A. (2006). Drosophila Rtf1 functions in histone methylation, gene expression, and Notch signaling. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 11970–11974. 86, 88

Theis, M., Slabicki, M., Junqueira, M., Paszkowski-Rogacz, M., Sontheimer, J., Kittler, R., Heninger, A.K., Glatter, T., Kruusmaa, K., Poser, I., Hyman, A.A., Pisabarro, M.T., Gstaiger, M., Aebersold, R., Shevchenko, A. & Buchholz, F. (2009). Comparative profiling identifies C13orf3 as a component of the Ska complex required for mammalian cell division. *The EMBO journal*, **28**, 1453–65. 36

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research*, **13**, 2129–41. 9, 99

Thomas, P.D., Kejariwal, A., Guo, N., Mi, H., Campbell, M.J., Muruganujan, A. & Lazareva-Ulitsky, B. (2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Research*, **34**, W645–W650. 99

Thornton, B.R., Ng, T.M., Matyskiela, M.E., Carroll, C.W., Morgan, D.O. & Toczyski, D.P. (2006). An architectural map of the anaphase-promoting complex. *Genes Dev.*, **20**, 449–460. 38

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.a., Noble, W.S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. & Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, **23**, 137–44. 19

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. & Altman, R.B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1461. 18

Tukey, J.W. (1977). *Exploratory Data Analysis.*. Addison-Wesley Publishing Company, Reading, Mass. et al. 12

Turksen, K. (2001). *Embryonic Stem Cells: Methods and Protocols, First Edition.* Humana Press, Totowa, NJ. 86, 90

Van Ryzin, G.G. (1995). Cluster Analysis as a Basis for Purposive Sampling of Projects in Case Study Evaluations. *American Journal of Evaluation*, **16**, 109–119. 26

Varetti, G. & Musacchio, A. (2008). The spindle assembly checkpoint. *Curr. Biol.*, **18**, R591–595. 38

Vassilev, L.T., Tovar, C., Chen, S., Knezevic, D., Zhao, X., Sun, H., Heimbrook, D.C. & Chen, L. (2006). Selective small-molecule inhibitor reveals critical mitotic functions of human CDK1. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 10660–10665. 50

Vidal, M., Braun, P., Chen, E., Boeke, J.D. & Harlow, E. (1996). Genetic characterization of a mammalian protein-protein interaction domain by using a yeast reverse two-hybrid system. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 10321–10326. 63

Villarroel, L., Marshall, G. & Barón, A.E. (2009). Cluster analysis using multivariate mixed effects models. *Statistics in medicine*, **28**, 2552–65. 26

Waizenegger, I.C., Hauf, S., Meinke, A. & Peters, J.M. (2000). Two distinct pathways remove mammalian cohesin from chromosome arms in prophase and from centromeres in anaphase. *Cell*, **103**, 399–410. 38, 53

Wang, E., Lenferink, A. & O'Connor-McCourt, M. (2007). Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cellular and molecular life sciences : CMLS*, **64**, 1752–62. 106

Wang, P., Bowl, M.R., Bender, S., Peng, J., Farber, L., Chen, J., Ali, A., Zhang, Z., Alberts, A.S., Thakker, R.V., Shilatifard, A., Williams, B.O. & Teh, B.T. (2008a). Parafibromin, a component of the human PAF complex, regulates growth factors and is required for embryonic development and survival in adult mice. *Mol. Cell. Biol.*, **28**, 2930–2940. 86

Wang, X., Yang, Y., Duan, Q., Jiang, N., Huang, Y., Darzynkiewicz, Z. & Dai, W. (2008b). sSgo1, a major splice variant of Sgo1, functions in centriole cohesion where it is regulated by Plk1. *Dev. Cell*, **14**, 331–341. 53

Wang, Y., Shyy, J.Y. & Chien, S. (2008c). Fluorescence proteins, live-cell imaging, and mechanobiology: seeing is believing. *Annu Rev Biomed Eng*, **10**, 1–38. 39

Watanabe, Y. (2005). Shugoshin: guardian spirit at the centromere. *Curr. Opin. Cell Biol.*, **17**, 590–595. 38

Weisberg, H. (1991). *Central Tendency and Variability*. SAGE Publications, Inc. 12

Williams, V.S.L., Jones, L.V. & Tukey, J.W. (1999). Controlling Error in Multiple Comparisons, with Examples from State-to-State Differences in Educational Achievement. *Journal of Educational and Behavioral Statistics*, **24**, 42–69. 21

Worby, C.A. & Dixon, J.E. (2002). Sorting out the cellular functions of sorting nexins. *Nature reviews. Molecular cell biology*, **3**, 919–31. 103

Wu, C.H., Yeh, L.S.L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P.a., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J. & Barker, W.C. (2003). The Protein Information Resource. *Nucleic acids research*, **31**, 345–7. 97

Wu, C.H., Sahoo, D., Arvanitis, C., Bradon, N., Dill, D.L. & Felsher, D.W. (2008). Combined analysis of murine and human microarrays and ChIP analysis reveals genes associated with the ability of MYC to maintain tumorigenesis. *PLoS genetics*, **4**. 7

Yamagishi, Y., Sakuno, T., Shimura, M. & Watanabe, Y. (2008). Heterochromatin links to centromeric protection by recruiting shugoshin. *Nature*, **455**, 251–255. 53

Yamazaki, K., Aso, T., Ohnishi, Y., Ohno, M., Tamura, K., Shuin, T., Kitajima, S. & Nakabeppu, Y. (2003). Mammalian elongin A is not essential for cell viability but is required for proper cell cycle progression with limited alteration of gene expression. *J. Biol. Chem.*, **278**, 13585–13589. 89

Yang, D., Buchholz, F., Huang, Z., Goga, A., Chen, C.Y., Brodsky, F.M. & Bishop, J.M. (2002a). Short RNA duplexes produced by hydrolysis with Escherichia coli RNase III mediate effective RNA interference in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9942–7. 5

Yang, D., Buchholz, F., Huang, Z., Goga, A., Chen, C.Y., Brodsky, F.M. & Bishop, J.M. (2002b). Short RNA duplexes produced by hydrolysis with Escherichia coli RNase III mediate effective RNA interference in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 9942–9947. 39

Ying, Q.L., Nichols, J., Evans, E.P. & Smith, A.G. (2002). Changing potency by spontaneous fusion. *Nature*, **416**, 545–548. 69, 91

Ying, Q.L., Stavridis, M., Griffiths, D., Li, M. & Smith, A. (2003). Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. *Nat. Biotechnol.*, **21**, 183–186. 93

Yoder, S.J. & Enkemann, S.A. (2009). ChIP-on-Chip Analysis methods for Affymetrix Tiling Arrays. *Methods in molecular biology (Clifton, N.J.)*, **523**, 367–81. 4, 18

Young, R.A. (2000). Biomedical discovery with DNA arrays. *Cell*, **102**, 9–15. 2

Zanella, F., Lorens, J.B. & Link, W. (2010). High content screening: seeing is believing. *Trends in biotechnology*. 6

Zhang, J., Chung, T. & Oldenburg, K. (1999). A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *Journal of biomolecular screening : the official journal of the Society for Biomolecular Screening*, **4**, 67–73. 10, 11

Zhang, S., Cao, J., Kong, Y.M. & Scheuermann, R.H. (2010). GO-Bayes: Gene Ontology-based over-representation analysis using a Bayesian approach. *Bioinformatics.* 20

Zhang, X., Kuan, P., Ferrer, M., Shu, X., Liu, Y., Gates, A., Kunapuli, P., Stec, E., Xu, M., Marine, S., Holder, D., S trulovici, B., Heyse, J. & Espeseth, A. (2008). Hit selection with false discovery rate control in genome-scale RNAi screens. *Nucleic acids research*, **36**, 4667–4679. 16

Zhang, X.D., Yang, X.C., Chung, N., Gates, A., Stec, E., Kunapuli, P., Holder, D.J., Ferrer, M. & Espeseth, A.S. (2006). Robust statistical methods for hit selection in RNA interference high-throughput screening experiments. *Pharmacogenomics*, **7**, 299–309. 11, 13, 15

Zhang, Y., Muyrers, J.P., Testa, G. & Stewart, A.F. (2000). DNA cloning by homologous recombination in Escherichia coli. *Nat. Biotechnol.*, **18**, 1314–1317. 59

Zhu, B., Mandal, S.S., Pham, A.D., Zheng, Y., Erdjument-Bromage, H., Batra, S.K., Tempst, P. & Reinberg, D. (2005). The human PAF complex coordinates transcription with events downstream of RNA synthesis. *Genes Dev.*, **19**, 1668–1673. 83