

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

presented by
M.Sc. Bioinformatics. Richa Batra
Born in: India
Oral-examination: 29.04.2013

Computational methods to analyze image-based siRNA knockdown screens

Referees:

PD. Dr. Rainer König

PD. Dr. Stefan Wiemann

न चौरहार्यम् न च राज हार्यम् ।
न भ्रातृभाज्यम् न च भारकारी ॥
व्यये कृते वर्धत् एव नित्यम् ।
विद्या धनम् सर्वधन प्रधानम् ॥

सुभाषितानि

It cannot be stolen by thieves, nor can it be taken away by kings. It cannot be divided among brothers and it does not cause a load on your shoulders. If spent, it indeed always keeps growing. The wealth of knowledge is the most superior wealth of all!

Abstract

Computational methods to analyze image-based siRNA knockdown screens

Neuroblastoma is the most common extra-cranial solid tumor of early childhood. Standard therapies are not effective in case of poor prognosis and chemotherapy resistance. To improve drug therapy, it is imperative to discover new targets that play a substantial role in tumorigenesis of neuroblastoma. The mitotic machinery is an attractive target for therapeutic interventions and inhibitors can be developed to target mitotic entry, spindle apparatus, spindle activation checkpoint, and mitotic exit. Thus, we performed a study to find genes that cause mitosis-linked cell death upon inhibition in neuroblastoma cells.

We investigated gene expression studies of neuroblastoma tumors and selected 240 genes relevant for tumorigenesis and cell cycle. With these genes we performed image-based time-lapse screening of gene knockdowns in neuroblastoma cells. We developed a classifier to classify images into cellular phenotypes, using SVM, performing manual evaluation and automatic corrections. This classifier yielded better predictions of cellular phenotypes than the standard classification protocol. We further developed an elaborated analysis pipeline based on the phenotype kinetics from the gene knockdown screening to identify genes with vital role in mitosis to identify therapeutic targets for neuroblastoma. We developed two methods (1) to generate clusters of genes with similar phenotype profiles and (2) to track the sequence of phenotype events, particularly mitosis-linked-cell-death.

We identified six genes (DLGAP5, DSCC1, SMO, SNRPD1, SSBP1, and UBE2C) that cause mitosis-linked-cell-death upon knockdown in both of the neuroblastoma cell lines tested (SH-EP and SK-N-BE(2)-C). Gene expression analysis of neuroblastoma patients show that these genes are up-regulated in aggressive tumors and they show good prediction performance for overall survival. Four of these hits (DLGAP5, DSCC1, SSBP1, UBE2C) are directly involved in cell cycle and one (SMO) indirectly which is involved in cell cycle regulation. Functional association and gene-expression analysis of these hits indicated that monitoring cell cycle dynamics enabled finding promising drug targets for neuroblastoma cells.

In summary, we present a bioinformatics pipeline to determine cancer specific therapeutic targets by first performing a focused gene expression analysis to select genes followed by a gene knockdown screening assay of live cells.

Abstract

Computergestützte Methoden für die Analyse von bild-basierten siRNA-Knockdown-Screens.

Das Neuroblastom ist der häufigste solide extrakranielle Tumor in der frühen Kindheit. Standard-Therapien sind hier unwirksam und gehen einher mit einer schlechten Prognose und Chemotherapie-Resistenz. Zur Verbesserung der medikamentösen Therapie ist es daher unerlässlich neue Ansatzpunkte zu entdecken, die eine wesentliche Rolle in der Tumorgenese von Neuroblastomen spielen. Der Prozess der Mitose bietet verschiedene Ansatzstellen für die Entwicklung therapeutischer Interventionen, indem Hemmstoffe entwickelt werden, welche auf die Einleitung der Mitose, den Spindelapparat, die Aktivierung des Spindel-Kontrollpunkts und den Austritt aus der Mitose abzielen. Aus diesem Grund haben wir eine Studie durchgeführt um Gene zu finden, deren Hemmung zu einem Mitose-assoziierten Zelltod von Neuroblastomzellen führt.

Hierzu untersuchten wir Genexpressionsstudien von Neuroblastom-Tumoren und wählten 240 für Tumorgenese und Zellzyklus relevante Gene aus. Mit diesen Genen führten wir bildbasierte Zeitverlaufsstudien von Gen-Knockdowns in Neuroblastom-Zellen durch. Zur Einteilung der Bilder in zelluläre Phänotypen entwickelten wir einen Klassifizierungsalgorithmus, welcher auf Support-Vektor-Maschinen sowie manuellen Auswertungen basiert und automatische Korrekturen durchführt. Unser Klassifikator erzielte bessere Vorhersagen der zellulären Phänotypen als das Standard-Klassifizierungsprotokoll. Weiterhin entwickelten wir eine detaillierte Analysepipeline auf Basis der Kinetik der Phänotypen aus dem Gen-Knockdown-Screen, um essentielle Gene der Mitose zu identifizieren und um so therapeutische Ansatzpunkte gegen das Neuroblastom zu finden. Wir haben zwei Verfahren entwickelt: (1) um Gruppen von Genen mit ähnlichen Phänotyp Profile zu finden und (2) um die Abfolge der phänotypischen Ereignisse zu verfolgen, insbesondere den Mitose-assoziierten Zelltod.

Mit Hilfe unserer Methoden konnten wir sechs Gene (DLGAP5, DSCC1, SMO, SNRPD1, SSBP1 und UBE2C) identifizieren, die nach einem Knockdown den Mitose-assoziierten Zelltod in beiden getesteten Neuroblastom Zelllinien (SH-EP und SK-N-BE(2)-C) verursachten. Genexpressionsanalysen von Neuroblastom-Patienten zeigen, dass diese Gene in aggressiven Tumoren hochreguliert sind. Zudem erwiesen sich diese Gene als gute Indikatoren für die Gesamtüberlebensdauer. Vier dieser Treffer (DLGAP5, DSCC1, SSBP1, UBE2C) sind direkt im Zellzyklus und einer (SMO) ist indirekt in dessen Regulation involviert. Funktionelle Assoziations- und Genexpressions-Analysen dieser Treffer deuteten darauf hin,

dass die Verfolgung der Zellzyklus-Dynamik das Auffinden vielversprechender Wirkstoffziele für Neuroblastomzellen ermöglicht hat.

Zusammenfassend stellen wir eine bioinformatische Pipeline zur Bestimmung Krebs-spezifischer therapeutischer Wirkziele vor, indem zuerst eine gezielte Gen-expressionsanalyse zur Auswahl von Kandidatengenomen erfolgt und einem anschließenden *in vitro* Gen Knockdown Screen.

CONTENTS

List of Publications	IV
List of Figures	VI
List of Tables	VIII
1 Introduction	3
1.1 Objective of the study	6
1.2 Outline of the thesis	6
2 Background	9
2.1 Biology	9
2.1.1 Cell cycle	9
2.2 Biotechnology	11
2.2.1 Gene expression profiling by microarrays	11
2.2.2 RNA interference	12
2.3 Bioinformatics and Statistics	13

2.3.1	Survival analysis of tumor patients	13
2.3.2	Classification	15
2.3.3	Statistical tests	19
3	Methods	23
3.1	Selecting genes for screening using gene expression analysis	23
3.2	Preparation of cell arrays and imaging	24
3.3	Image Processing	25
3.4	Classification of nuclei into phenotypes	27
3.4.1	Training set	27
3.4.2	Feature normalization	28
3.4.3	Classification model	28
3.4.4	Filter	29
3.4.5	Manual evaluation of classification	29
3.4.6	Automatic error correction	30
3.5	Quantitative analysis of phenotype kinetics	31
3.5.1	Normalization	31
3.5.2	Defining the phenotype signal	31
3.5.3	Estimating the phenotypic score	33
3.6	Analyzing phenotype profiles	33
3.6.1	Clustering of phenotype profiles	33
3.6.2	Tracking of phenotype events	34
3.7	Estimating the periodicity	35
3.8	Expression analysis of the hits	35
3.9	Enrichment Analysis	36

3.10	Validation experiments	36
3.10.1	Ploidy and Cell cycle analysis	37
3.10.2	Cytogenetic analysis	37
3.10.3	MAD2L1 knockdown	37
4	Results and Discussion	39
4.1	Selecting relevant genes for knockdown screening	39
4.2	Data and Data processing	41
4.3	Automated classification of cellular phenotypes	42
4.3.1	Classification performance based on training set	42
4.3.2	Optimization of feature normalization	43
4.3.3	Manual evaluation of classification performance	45
4.3.4	Automatic error correction	47
4.3.5	Summary	48
4.4	Quality control of the experimental set-up	48
4.5	Estimating cell cycle kinetics	49
4.6	Clustering phenotype profiles	52
4.6.1	Phenocluster: Knockdowns dependent on MYCN and p53	54
4.7	Temporal tracking of phenotype events	55
4.7.1	Predicting upstream kinase regulators	57
4.7.2	Comparing the two neuroblastoma cell lines	59
4.8	Data access via data repository iCHIP	62
5	Conclusion	65
	Appendix	69

CONTENTS	IV
A Appendix: Screened genes	71
B Appendix: Phenotypes in SHEP Cell line	85
C Appendix: Phenotypes in Be2C Cell line	95
Glossary	105
Bibliography	108
Acknowledgments	125
Index	127
Erklärung	129

LIST OF PUBLICATIONS

Some of the work from this thesis has already been published in a journal or presented in a conference. Here is a list of the same.

Conference proceedings

1. N. Harder, **R. Batra**, S. Gogolin, N. Diessl, R. Eils, F. Westermann, R. König, and K. Rohr (2011). Large-Scale Tracking For Cell Migration and Proliferation Analysis, and Experimental Optimization of High-Throughput Screens. *Microscopic Image Analysis with Application in Biology, Heidelberg, Germany*
2. N. Harder, **R. Batra**, S. Gogolin, N. Diessl, R. Eils, F. Westermann, R. König, and K. Rohr (2012). Cell Tracking for Automatic Migration and Proliferation Analysis in High-Throughput Screens. *Bildverarbeitung fr die Medizin 2012: 243-248*

Journal Publications

1. **R. Batra**, N. Harder, S. Gogolin, N. Diessl, R. Eils, K. Rohr, F. Westermann, and R. König (2012). Time-lapse imaging of neuroblastoma cells to determine cell fate upon gene knockdown. *PLoS One*. 7(12):e50988. PMID: 23251412.

2. S. Gogolin, **R. Batra**, N. Harder, V. Ehemann, T. Paffhausen, N. Diessl, S. Gade, I. Nolte, K. Rohr, R. König, and F. Westermann (2012). MYCN-mediated overexpression of mitotic spindle regulatory genes and loss of p53-p21 function jointly support the survival of tetraploid neuroblastoma cells. *Cancer Letters*. PMID: 23186832.

Posters

1. **R. Batra**, M. Oswald, R. Eils, and R. König. HIV-1, Human Protein Interactions: Meta-Analysis. *International Conference on Systems Biology (ICSB), Heidelberg, Germany, 2011*
2. **R. Batra**, N. Harder, N. Diessl, A. Suratane, H. Erfle, K. Rohr, F. Westermann, M. Schwab, R. Eils, and R. König. Data analysis of high throughput time-lapse RNAi screen of neuroblastoma cell lines. *Systems Genomics, Heidelberg, Germany, 2010*
3. **R. Batra**, N. Diessl, A. Suratane, N. Harder, F. Westermann, M. Schwab, H. Erfle, K. Rohr, R. Eils, and R. König. Analyzing genome wide RNAi knockdown screen of neuroblastoma. *Annual meeting of NGFN PLUS and NGFN-TRANSFER, Berlin, Germany, 2009*

LIST OF FIGURES

1.1	Consequences of a gene knockdown on the cell cycle and cell fate.	5
2.1	Cell cycle	10
2.2	Mitosis	11
2.3	RNA interference.	13
2.4	A linear classifier.	16
2.5	A non-linear classifier.	16
2.6	K-fold cross validation process.	17
2.7	Nested cross validation.	18
3.1	Sample images of the four phenotype classes	27
3.2	Cell arrays before and after normalization.	32
3.3	Determination of hits that show cell death during or after mitosis.	34
4.1	The workflow	40
4.2	Sample images of cell lines	42

4.3	Classification performance after 5-fold cross validation with normalized and non-normalized features.	45
4.4	Manually evaluated performance of the classifiers.	45
4.5	Classification performance after automatic correction.	47
4.6	Protocol for nuclei classification.	48
4.7	Experimental quality control.	50
4.8	Time series of interphase cells during five days of screening.	51
4.9	Phenoclusters	53
4.10	Selection of time-lapse images illustrating cell fate observed in the SH-EP cell line for the six hits.	61
4.11	Kaplan Meier plots of the six hits.	63

LIST OF TABLES

2.1	Contingency table.	20
4.1	Functional enrichment of screened genes.	41
4.2	Confusion matrix of the classification results after 5-fold cross validation and non-normalized features	43
4.3	Confusion matrix of the classification results after 5-fold cross validation and normalized features.	44
4.4	Confusion matrix of the classification results after manual evaluation.	46
4.5	Confusion matrix of the classification results after automatic corrections.	47
4.6	Genes which cause mitosis-linked-cell-death phenotype.	56
4.7	Upstream kinase enrichment of candidate genes	57
4.8	Gene expression analysis of the six hits.	62
A.1	Screened gene list	71
B.1	Phenotypes in SH-EP cell lines.	85

C.1 Phenotypes in SK-N-BE(2)-C cell lines.	95
--	----

CHAPTER 1

INTRODUCTION

Neuroblastoma is an embryonic tumor arising in the sympathetic nervous system, mostly in adrenal glands. The genetic causes of neuroblastoma are still unclear however mutations in ALK [Mosse et al., 2008] and PHOX2B [Mosse et al., 2004] have been identified in most familial cases of neuroblastoma. While somatic mutations in BARD1 [Capasso et al., 2009], chromosome band 6p22.3 [Maris et al., 2008], copy number variation at 1q21 [Diskin et al., 2009], are frequently observed in sporadic neuroblastoma. The clinical courses of neuroblastoma are very heterogeneous. Some tumors undergo spontaneous regression without therapy, whereas, high-risk neuroblastoma patients are often resistant to available therapies and undergo a fatal clinical outcome [Deyell and Attiyeh, 2011]. These varied clinical courses depend on the age of the patient, stage of the disease and genetic abnormalities like, MYCN amplification [Brodeur et al., 1984] or aberrations of chromosome 11q [Gaudray et al., 1992].

MYCN serves as a prognostic marker for neuroblastoma [Brodeur, 2003, Westermann et al., 2008] and is a central regulator of the cell cycle [Obaya et al., 1999]. Our group was involved in a study to predict genes regulated by MYCN. In this study [Westermann et al., 2008] a genome-wide search was performed for genes that were directly regulated by MYC/MYCN or indirectly involved in MYCN-induced regulation. The gene expression profiles were clustered and enrichment analysis of MYC targets was performed to predict new targets.

Neuroblastoma exhibits heterogeneous clinical courses. Stage 4 classified tumors have a very poor prognosis (aggressive tumors), in contrast to stage 1 tumors which have a very good prognosis and often show spontaneous regression [Brodeur, 2003]. Risk-classifiers have been developed for distinguishing tumors with varying clinical courses. These classifiers consider the features such as age of the patient, stage of disease and other biological variables [Maris, 2010]. Our group was involved in the development of such a classifier which was based on the differential gene expression. In this study [Oberthuer et al., 2006], gene expression in 251 neuroblastoma patients was analyzed to find differential expression in clinical subgroups. Using the maximally divergent clinical course in 77 patients a 144-gene-based classifier was developed to assist risk estimation for neuroblastoma patients.

Aneuploidy is a common feature of cancers [Bharadwaj and Yu, 2004]. It is the condition where a cell has an abnormal chromosome number. It is caused by abnormal mitosis, where chromosome segregation during anaphase is defective. Such mitosis results in loss or duplication of chromosomes in the daughter cells [Griffiths AJF, 2000]. Cells with such aberrations are usually non-viable. There are many possible components of mitosis when defective which can cause aneuploidy [Gordon et al., 2012]. Inefficient Metaphase-Anaphase (M-A) checkpoint is one of the sources of aneuploidy and is considered an anticancer strategy. M-A checkpoint is the surveillance system to ensure proper attachment of chromosomes to the microtubules and the tension in microtubules. It can inhibit chromosome segregation in anaphase if there are any defects [Bharadwaj and Yu, 2004]. When this checkpoint is affected the miss-segregation rate increases leading to aneuploidy. Inhibitors of the mitotic spindle have been extensively used in chemotherapy [Li and Li, 2006]. However, susceptibility to these drugs is dependent on the tumor type [Kavallaris, 2010]. Though, given the high degree of heterogeneity in response to anti-mitotic drugs in different tumor cells [Gascoigne and Taylor, 2008], identification of target proteins that are substantial for the etiology of neuroblastoma is a challenging task. Hence, the search for genes with therapeutic potential requires a sophisticated approach.

Functional genomics and cancer genetics consistently exploit high-throughput RNA interference knockdown screens to investigate consequences of eliminating specific genes [Willingham et al., 2004, Cole et al., 2011, Holzel et al., 2010].

siRNA assays based on a single readout, such as cell viability, growth rate, or reporter activity (luciferase) are easy to scale up in high throughput. However, they contain limited information as they provide only an endpoint snapshot of a cells reaction [Markowetz, 2010]. In turn, image-based knockdown screens provide multi-parametric readouts and enable tracking more complex phenotypes. However, these assays are laborious on a high-throughput scale. We combined the best of both to infer gene function in a time-dependent manner, as explained in detail in the following. To gain functional information from images, image processing methods were established to segment whole cells and cell nuclei (i.e. to separate them from the image background) and to extract their morphological features [Harder et al., 2009] [Harder et al., 2011]. Techniques have been developed to distinguish and quantify different cell shapes [Bakal et al., 2007], to determine sub cellular localizations [Conrad et al., 2004], to identify mitotic phases [Harder et al., 2006], and to cluster genes based on phenotypic similarity [Fuchs et al., 2010].

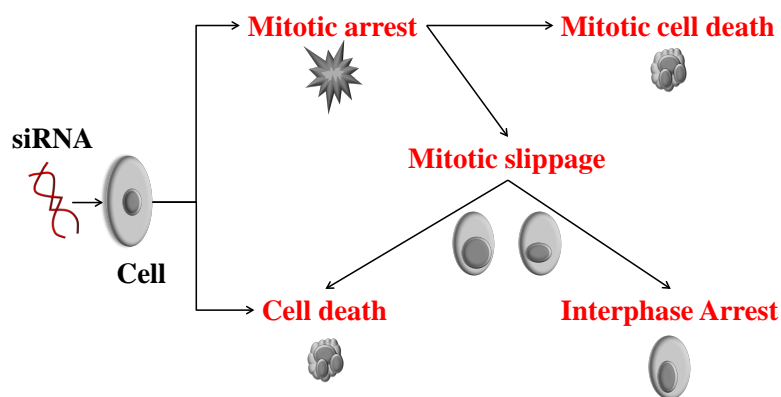


Figure 1.1: Consequences of a gene knockdown on the cell cycle and cell fate: These effects can be observed (directly or indirectly) by imaging cells with silenced genes following a time-lapse screening assay. Cells may directly be affected from the loss-of-function of a gene and die (cell death); they may enter mitosis and die before completion of mitosis (cell death in mitotic arrest), or may undergo mitotic slippage followed by interphase arrest or cell death.

1.1 Objective of the study

In this study, we used a set of genes relevant to neuroblastoma. We selected genes from an established gene-expression-based classifier developed previously in our group [Oberthuer et al., 2006]. Furthermore, we selected genes which are regulated by the prognostic marker MYCN/MYC as found in our previous *in vitro* study [Westermann et al., 2008]. With these genes we performed time-lapse image-based loss-of-function assays to determine cell fate upon gene knockdown. As an example, different outcomes of gene silencing are shown in Figure: 1.1 (page: 5). For instance, perturbation of constitutively expressed anti-apoptotic genes may lead to cell death. As such, targeting mitotic genes can lead to mitotic arrest and this may lead to cell death depending on the mitotic component that was targeted [Manchado et al., 2012, Vakifahmetoglu et al., 2008]. Targeting the mitotic checkpoint can cause aneuploidy resulting in asymmetric segregation of chromosomes during anaphase or tetraploidy. An abnormal division can result in non-viable daughter cells. Some knockdowns can cause mitotic arrest and after prolonged mitotic arrest, a cell can either die or exit mitosis without cell division known as mitotic slippage. Knockdowns resulting in such abnormal mitotic fate are attractive therapeutic candidates. Hence, we focused our analysis on identifying such perturbations.

1.2 Outline of the thesis

This thesis contains 5 chapters.

Chapter 1: Introduction, it conveys the motivation and objective of the work. It gives an overview of the neuroblastoma cancer and mentions how functional genomics has contributed in cancer functional genomics. Chapter 2: Background, details the main techniques and concepts used in this thesis. It describes the cell cycle as our analysis tracked cell fate upon gene knockdown, RNA interference which is the technique that has been used for gene knockdown, Support Vector Machines which has been used for phenotype classification of cells, and the applied statistics methods. Chapter 3: Methods, it describes the methods developed in this study. It explains the gene selection scheme that we employed to select a

set of 240 genes that were screened. It describes the phenotype classification scheme and optimization process that we performed. It presents a method to cluster phenotype profiles and a novel method to determine cell fate upon a gene knockdown, based on the temporal tracking of phenotype emergence. Chapter 4: Results and Discussion, it presents the results of the entire screening process and its analysis. It shows that the gene selection process helped us in deriving a set of cell cycle associated genes. It depicts the results of the improved classification scheme. It discusses that cells in the screen have synchronized cell cycle, thus the population-average response can be computed to represent the knockdown effect. Cell fate upon knockdown of each gene is also tabulated. Validations of the results as performed by our collaborators are also described in this chapter. Chapter 5: Conclusion of the study and possible follow up.

CHAPTER 2

BACKGROUND

2.1 Biology

2.1.1 Cell cycle

The cell cycle describes the process of growth and division of a cell. There are several stages of this process. Interphase and mitosis are two major stages of cell division. A typical eukaryotic cell cycle has duration of approximately 24 hours. Of the 24 hours mitosis is 1 hour and the rest of the time cell is in interphase [Cooper, 2000].

Interphase

During interphase DNA of the cell replicates and the chromosomes duplicate, the cell grows in size and prepares to divide into two cells. In interphase, the chromosomes are de-condensed and distributed in the nucleus and so they appear morphologically uniform [Cooper, 2000]. The interphase is further divided into three phases, G1 (gap 1), S (synthesis), G2 (gap 2). G1 is the primary growth phase and is the longest lasting phase. A cell spends most of its life in G1 phase. Most of the proteins are synthesized during this phase and are used in the later part of the cell cycle. S is the synthesis phase where genetic material

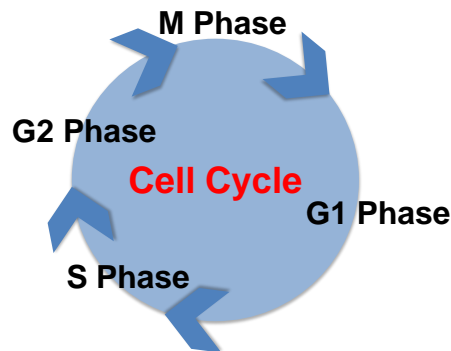


Figure 2.1: Cell cycle. The arrows mark the phases of the cell cycle. G1 phase is the primary growth phase, genome replication occurs during S phase, G2 phase is the preparatory phase for division, M phase is the division phase called mitosis.

of a cell is replicated such that each chromosome pair duplicates. The two pairs of chromosomes stay attached to each other and are called sister chromatids. G2 phase prepares the cell for the M phase. The chromosomes condense and microtubules start assembling. Microtubules are proteins that help in separation of chromosomes during the division of the cell [Raven, 2007].

Mitosis

Mitosis involves separation of the chromosomes and usually ends with cytokinesis i.e. division of the cell [Cooper, 2000]. The chromosomes bind to the microtubules during mitosis. As the mitosis proceeds, microtubules constrict in opposite directions and separate the sister chromatids. Mitosis is followed by cytokinesis. Mitosis also has several phases. During the prophase the nucleus membrane disintegrates and the chromosomes condense, during metaphase the chromosomes attach to the microtubules and align themselves along the equator of the cell, in anaphase microtubules start contracting towards each pole, in telophase the chromosomes reach the poles, the chromosomes condense and the nucleus membrane re-forms.

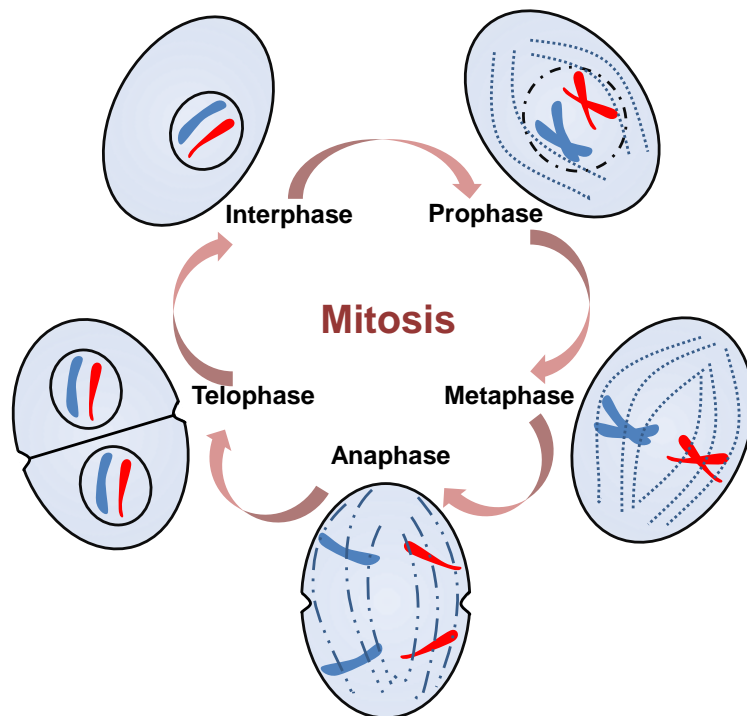


Figure 2.2: Mitosis. The separation of the chromosomes duplicated during S phase is separated into two daughter cells during M phase. It is followed by cytokinesis.

2.2 Biotechnology

2.2.1 Gene expression profiling by microarrays

In a cell, genes are expressed at a specific time according to the function and developmental stage of the cell. To study the circumstances at which a gene is expressed a gene profiling assay called microarrays had been developed. DNA microarrays allow rapid and simultaneous screening of thousands of genes [David L. Nelson, 2005]. Alteration in gene expression owing to certain treatments, diseases or developmental stages can be studied using DNA microarrays. In this assay, segments of DNA called probes are placed on a chip and then probed with mRNAs to identify the genes that are expressed in those cells. A microarray chip is a solid surface of glass, plastic or silicon chip. It contains 1000s of probes attached to it by a covalent bond. It is based on the principle of hybridization of complementary DNA strands through hydrogen bonds formed between the

complementary bases. For the experiment, mRNAs from cells are isolated at a certain time point or after treatment of interest. The mRNAs are converted to complementary DNAs (cDNA) using reverse transcriptase. The nucleotides used in the construction of these cDNAs are fluorescently labeled. These cDNAs are probed on to the microarray chip, such that the cDNAs bind to the spots which have complementary sequences. Under the scanner the fluorescence allows the recognition of those probes which have paired with a cDNA. The intensity of the fluorescence is proportional to the amount of probe and cDNA pairs formed, more the intensity higher the expression of that gene and vice versa.

Like other experiments, microarray data is also subjected to systematic errors. Several normalization or standardization methods have been proposed to analyze microarray results. Quantile normalization is a common and robust microarray normalization method. In this method, the intensities of each experiment are sorted. The ranked intensities are then replaced by the mean of the values of that rank of all experiments, for instance the highest signal is replaced by the mean of all the highest signals and so on. This maintains the difference between the values in different ranges, and the data is not skewed by outliers [Bolstad et al., 2003].

2.2.2 RNA interference

RNA interference is the process of RNA mediated gene regulatory process. There are two types of small RNA molecules involved in RNA interference -micro RNA and small interfering RNA (siRNA). For high throughput gene silencing screens siRNA is used.

Mechanism of RNA mediated gene silencing The siRNA knockdown technique has greatly improved functional genomics studies. The function of a protein-coding gene is analyzed by perturbing its translation. Exogenous siRNA specific to a target mRNA is introduced in the cell there by degrading its target mRNA. Large scale high-throughput screens are commonly performed to study loss-of-function phenotypes in various species.

Dicer, a dsRNA specific RNAase III, binds and cleaves any double stranded siRNA (endogenous or exogenous) into small fragments (21-22 nucleotides) called

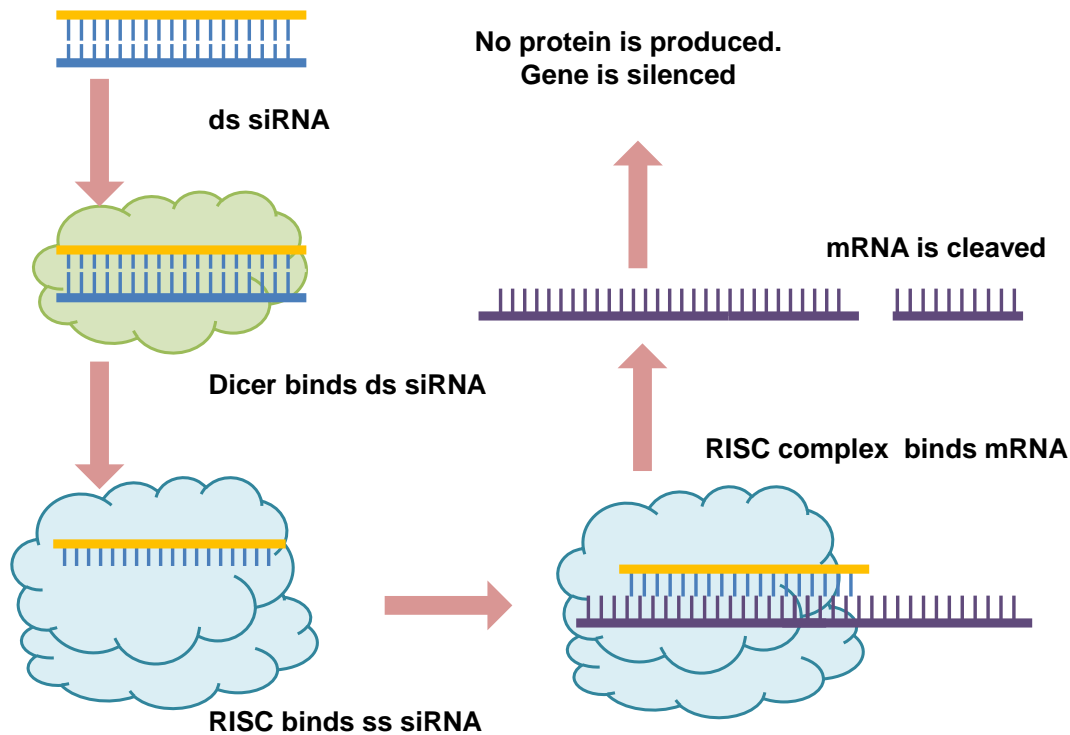


Figure 2.3: RNA interference. The mechanism of RNA mediated gene silencing.

small interfering RNAs (siRNAs). siRNAs can then bind to Slicer an argonaut protein to form the ribonucleotide silencing complex (RISC). RISC enables the unwinding of the siRNA, resulting in a guide strand and a passenger strand. The passenger strand is quickly degraded. RISC uses the guide strand to recognize target mRNA. mRNA-siRNA binding takes place in a sequence-specific manner, finally leading to the degradation of the target mRNA [Echeverri and Perrimon, 2006].

2.3 Bioinformatics and Statistics

2.3.1 Survival analysis of tumor patients

Survival analysis deals with time and event. In cancer studies, survival analysis is used to study the occurrence of a significant event after the prognosis or treat-

ment. The event of consideration could be regression, relapse, or death. The probability of the survival of an individual from the time of origin (e.g. diagnosis of cancer) at a specific time t is called survival probability $S(t)$. The probability that an individual has an event at time t is called hazard probability $h(t)$.

Kaplan-Meier Survival estimates

The survival probability can be predicted using the Kaplan-Meier method which is based on the observed survival times of patients [Clark, 2003]. The Kaplan-Meier equation is given as

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right) \quad (2.1)$$

where $S(t_j)$ is the probability of the survival of the patients at time j , $S(t_{j-1})$ is the probability of survival at time $j - 1$, d_j is the number of events at t_j , n_j is the number of patients alive at t_j .

As $S(t)$ is computed based on a previous event the predicted probability is a step function i.e. it changes from event to event i.e. $S(t)$ is constant between events. A plot of KM survival probability against time is called KM survival curve. The KM survival curve shows a summary survival over time for all investigated patients [Kishore, 2010].

Log-rank test

The log-rank test is a common non-parametric test to compare survival distributions of two or more groups. These groups can be prognostic groups or treatment groups. In this method, the expected number of events at a given time is computed at each event time, given the time of the previous event in each group. The total expected number of events are then summed up for each group. The log-rank test compares the expected and observed number of events [Clark, 2003]. The test statistic is given by

$$L = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \quad (2.2)$$

where E_1 , E_2 is the expected number of events in group 1, group 2, O_1 , O_2 is the observed number of events in group 1 and group 2 respectively.

The calculated value can then be compared to a critical value as per the chi-square table at degrees of freedom $g - 1$ where g is the number of groups compared. This yields the significance level (p-value) of the difference between the two groups [Kishore, 2010].

2.3.2 Classification

Classification is the supervised process of dividing a set of objects in groups based on their properties. There are several algorithms used for classification. Supervised algorithms derive rules for classification based on the input training samples with predefined classes. While unsupervised algorithms may derive classification rules from the input data distribution in feature space with no prior knowledge of expected classes. Supervised algorithms perform better when expected classes are well defined [V. Kovalev and Rohr, 2006] as is in case of mitotic phase identification.

Support vector machines

SVM is the most suitable and popular choice for nuclei classification because of higher prediction accuracy [Conrad et al., 2004][Neumann et al., 2006][Fuchs et al., 2010][Walter et al., 2010]. It needs a training set i.e. a set of samples with known classes. The basic task of SVMs is to derive the rules which separate the classes in the training set. These rules can be then applied to the cases where the class is not known. It is primarily a binary classifier, which aims at maximizing the separation between margin of two classes. It can be applied to multi-class problems using multiple binary classifiers.

Linear SVM

For a linearly separable case suppose x_i is a set of points and belonging to one of the classes in $y_i : y_i \in -1, 1$. Using the training set the algorithm tries to place a hyperplane between the points where $y_i = -1$ and points where $y_i = 1$. A hyperplane is a generalization of a plane in any number of dimensions. The

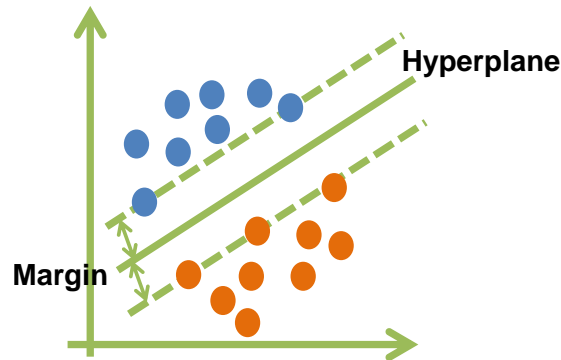


Figure 2.4: A linear classifier. The linear decision boundary divides the two groups.

separating hyperplane separates the space into two half spaces. There can be more than one hyperplane that can classify the data. The hyperplane which maximizes the margin between the classes is chosen, as it reduces low certainty decisions. Once a separation between the two classes is done, a new point can be classified based on which side of the hyperplane it lays [Thomaz and Gillies, 2011].

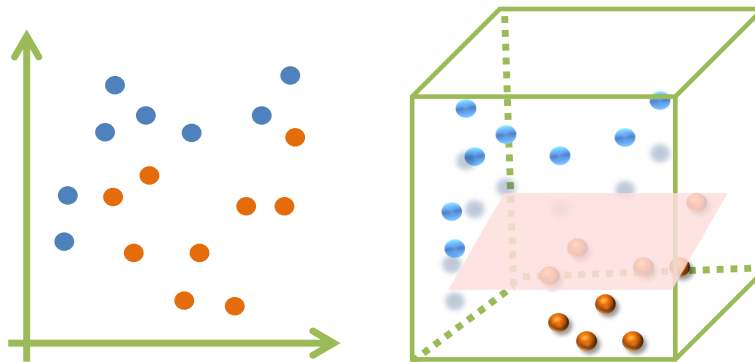


Figure 2.5: A non-linear classifier. A kernel trick is used to compute decision boundary in high-dimensional space.

Non-linear SVM

SVMs can also be extended for non-linearly separable cases by transformation of data points in a higher dimensional space. With this the data can be separated by a linear hyperplane though the transformation is non-linear. Mapping a data point in a higher dimensional space is computationally extensive and so a kernel trick is used. Kernel computes the dot-product in the high dimensional space without transformation of the data points [Hur and Weston, 2011].

Process of classification using SVMs

The entire process of classification using support vector machines involves the following steps.

1. Annotation of a training set: a set of data points are identified that belong to each class.
2. Model generation: the hyperplane is generated using the training set.
3. Validation: a test set is used to evaluate the performance of the model. A test set is a set of annotated samples apart from the training set [wei Hsu et al., 2010, Hur and Weston, 2011, Thomaz and Gillies, 2011].

Cross validation

Cross validation is a method to evaluate the prediction accuracy of a classifier when the annotated data is limited. The annotated samples are divided into a training set, from which the classifier is generated and a test set, on which the classifier is tested. This process is called holdout method and it helps in identifying the performance of the classifier on the unseen data i.e. the data which was not used in building the classifier.

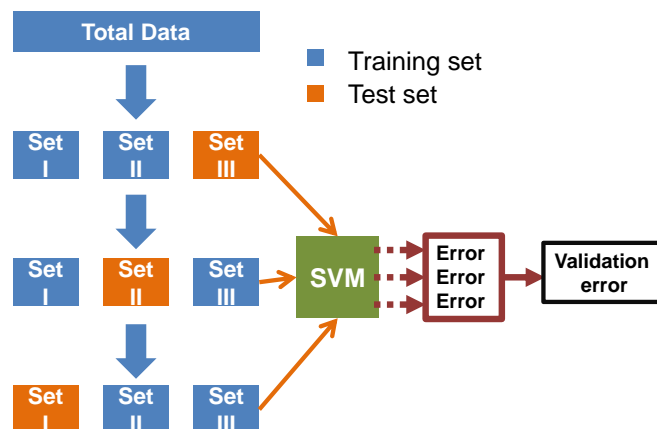


Figure 2.6: K-fold cross validation process.

K-fold cross validation is the most popular cross validation process (Figure: 2.7, page: 18). In this process, the annotated set is divided into k sets. The holdout

method is repeated k times and each time $k - 1$ sets are used as training and 1 set is used for testing. The advantage of this method is that each sample gets to be in the test set at least once. As the number of k increases, the variance in the estimate decreases. The average accuracy of the k rounds is the estimated accuracy of the classifier.

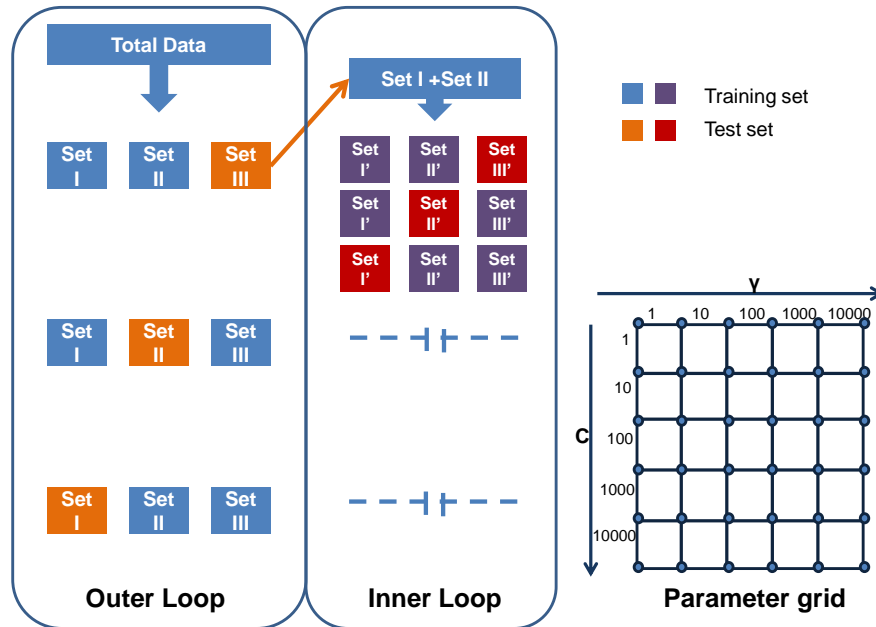


Figure 2.7: Nested cross validation. A three-fold inner loop for parameter optimization with a three-fold outer loop for estimating classifier's performance.

Parameter optimization

There are two parameters, C (the cost of misclassification) and γ (corresponding to the flexibility of decision boundary) that need optimization. The optimization aims to find the values of C and γ with low prediction errors. A grid search is the most effective and popular way to perform parameter optimization. In this process a series of values of both these parameters are used to build a classifier, one pair at a time. The data is divided into k sets such that one set is validation set and the rest are used for building the classifier with the given pair of parameters. This process is repeated k times for a pair of parameters. The pair which gives the lowest error is selected. The entire process is shown in the Figure 2.7 (page 18).

2.3.3 Statistical tests

Normal distribution

The normal distribution is a family of curves with a typical bell shaped curve. A population X is said to be normally distributed if for every pair of samples $a \leq b$, the chance that

$$a < (Xm)/s < b \quad (2.3)$$

is

$$P(a < (Xm)/s < b) = area \quad (2.4)$$

where *area* is area under the curve a and b , m is the mean of the distribution and s is the standard deviation. In special cases a normal curve has *mean* = 0 and $\sigma^2 = 1$, this curve is called standard normal curve [Le, 2003][Stark, 1997].

Students T-test

It is a parametric inferential statistical test used to compare the mean of two groups. It assumes that the population is normally distributed and variances are equal of the two groups.

Wilcoxon test

It is a non-parametric inferential test, used for comparing two samples drawn from two independent populations. It differs from the T-test such that the distribution of the population need not be normal, but the distributions of the populations need to be same. It compares the median of the population, for normal distribution mean and median are same [Le, 2003].

For instance, if we want to compare the number of books girls own versus number of books boys own in a class. For one sided test, suppose the null hypothesis is that the median number of books girls own is smaller than the median number of books boys own. Alternative hypothesis is the median number of books girls own is greater than the median number of books boys own. Wilcox test is performed by ranking such that the students who have least number of books rank 1 and so on. Sum the ranks for the group (girls) which may have lower ranks as per

null hypothesis. At 0.05 significance level, if sum of the ranks is greater than the critical value, the null hypothesis is rejected.

Hypergeometric distribution

The hypergeometric distribution is the distribution of "good" objects in a simple random sample of size n from a population of N objects of which G are "good". In other words, it provides the probability of finding exactly i good objects when randomly drawn from a population of N objects of which G are good. Hypergeometric distribution assumes sampling is without replacement [Stark, 1997]. The probability of i successful selections is given by

$$P(x = i) = \frac{[i'] [ni]}{[t']} \quad (2.5)$$

where i' is the number of ways of ni successes, n' is the number of ways for $n - i$ failures, t' is the total number of ways to select.

Fishers exact test

Fisher's exact test is used when there are two categorical variables. It is used to compare the proportion of these variables in their respective populations. The hypergeometric distribution is used to calculate the probability of getting the observed data [NIST, 2003].

Table 2.1: Contingency table.

	bag1	bag2	
marble	m1	m2	M
ball	b1	b2	B
	n1	n2	N

Consider there are two bags of marbles and balls. Let p_1 and p_2 be the proportion of marbles in the two bags. Let n_1 and n_2 be the size of sample from bag 1 and bag 2 respectively. Let N be the total number of marbles in both samples. The null hypothesis to be tested is that $p_1 = p_2$, based on simple random sampling from each bag. The two nominal variables are bags (bag 1 and bag 2) and objects (marble and ball). If the two bags have same distribution of objects then the number of marbles selected from both bags should be same. The observations

can be written in the form of the matrix as Table: 2.1 (page: 20). Let M and B be row sums, $n1$ and $n2$ be column sums. The conditional probability of getting the actual matrix given the row and column sums is given by

$$P_{cutoff} = \frac{(M!B!)(n1!n2!)}{(N!)(m1!m2!b1!b2!)} \quad (2.6)$$

We selected genes involved in the malignant progression of neuroblastomas based on gene expression analysis. Subsequently, these genes were subjected to time-lapse image-based knockdown screens in the SH-EP cell line from neuroblastoma. By automated image processing and machine learning through Support Vector Machines (SVMs), a quantitative description of phenotypic classes and cell nuclei were obtained from raw bitmaps. Thereafter, perturbation consequence was inferred from the analysis of the phenotypic dynamics focusing on cell death, death in mitosis and death after mitosis. The analysis was repeated using a second neuroblastoma cell line (SK-N-BE(2)-C). This resulted in a small set of genes which was verified using gene expression data from neuroblastoma patients and literature. We predicted potential kinases regulating the candidate genes using a repository on kinase-substrate interactions and verified in literature. One of the genes was validated via cytometric and cytogenetic experiments.

3.1 Selecting genes for screening using gene expression analysis

In a previous study by Oberthuer et al. [Oberthuer et al., 2006], a neuroblastoma-specific microarray chip was designed which covered a high percentage of tran-

scripts that are differentially expressed in the major clinically distinct subgroups of neuroblastoma tumors. Using this customized 11K oligonucleotide microarray, 251 neuroblastoma specimens were analyzed and a 144-gene predictor signature was assembled to predict the course of the disease.

In a follow-up study by Westermann et al. [Westermann et al., 2008], the same neuroblastoma-specific microarray was used to identify MYCN/MYC target genes using a neuroblastoma cell line SH-EP^{MYCN}. SH-EP^{MYCN} is a neuroblastoma cell line that stably expresses an inducible MYCN transgene, thus allowing conditional expression of MYCN. Gene expression profiles of a time series after MYCN induction were obtained with the customized 11K microarray. The profiles were clustered using self-organizing maps (SOM) which resulted in 504 clusters (best matching units, BMUs) of genes with similar gene expression profiles. Clusters (BMU: 140, 168, 195, 280, 308, 336, defined as subgroups I and II in [Westermann et al., 2008]) were detected which were enriched in the E-Box motif (binding motif of MYCN/MYC, p-value ≤ 0.05 using a Fisher's Exact test, adjusted for multiple testing of all BMUs using the method of Benjamini-Hochberg [Benjamini and Hochberg, 1995]), indicating potential targets of the MYC transcription factor family [Westermann et al., 2008].

We selected 127 genes from these clusters. In addition, we selected 80 genes from the BMUs which were enriched in genes from the 144-gene predictor signature. For this, we computed the percentages of the predictive-signature-genes that matched to the identified clusters. The top three clusters (BMU: 504, 476, 475) with the highest odd ratios (0.49, 0.41, and 0.3) were selected. Further, 33 genes which were associated to neuroblastoma tumor progression were selected from literature. Finally a set of 240 genes was assembled and used for the knockdown screen (a list of all genes is given in Appendix A).

3.2 Preparation of cell arrays and imaging

Two neuroblastoma cell lines, SH-EP and SK-N-BE(2)-C, were used in the screen. These cell lines were transfected with a construct of the gene coding for histone H2B tagged with Green Fluorescent Protein (GFP) as described previously [Kanda et al., 1998]. Briefly, a chimeric gene with a cDNA construct of H2B

gene tagged with GFP was sub-cloned into a mammalian expression vector. This vector was used to transfect the cell lines. Thus, the product of this gene H2B-GFP protein was incorporated into the nucleosomes which allowed imaging of mitotic chromosomes and interphase chromosomes. Further, cover glass culture chambers called LabTeks were automatically spotted and dried as previously described [Erfe et al., 2007]. Sample preparation for spotting, mixing of transfection reagents and siRNAs was done using an automated liquid handler. Automated spotting of this transfection solution onto LabTeks was performed with a contact printer. After drying the LabTeks for at least 12 hours, SH-EP/H2B-GFP and SK-N-BE(2)-C/H2B-GFP (60, 000 cells/LabTek) were seeded on the LabTeks and incubated in a stage top chamber by LCI, with 1.5 ml growth medium at 37 C, 95% humidity, and 5% CO₂. Eight LabTeks with 275 spots were used to cover several mock (no siRNA) spots, 2 siRNAs (Ambion) per gene and four replicates per siRNA. Images were acquired (16 hour post seeding) for five days at an acquisition rate of 35-40 minutes using an automated wide-field fluorescence microscope (Olympus X81 'inverted' ScanR System) with 10x magnification.

3.3 Image Processing

This section describes the work done by our collaborators. Dr. Nathalie Harder in the lab of PD. Dr. Karl Rohr.

Nuclei segmentation was performed using a region adaptive thresholding scheme, which allowed detection of cells with varying contrast. Clusters of cells were resolved by Euclidean distance transformation of the segmentation output followed by watershed transformation to split them into single cells. Dense clusters of cells growing on top of each other could not be resolved using this approach and were treated as cluster objects in the subsequent analysis.

To bring all images of different spots and cell arrays to a comparable gray value range, gray value normalization was performed before feature extraction. To this end, the average distribution (histogram) of the foreground pixels of the complete data set was computed and three features of this histogram were extracted (i.e. location of the maximum peak and its width to the left as well as to the right). For gray value normalization each individual image histogram was mapped to this

average histogram and the gray values of the respective images were transformed and scaled accordingly.

A set of 349 image features was computed for each nucleus, describing the texture and morphology as described previously [Harder et al., 2008]. Object-related features include basic characteristics like the size of the object (number of pixels), as well as the objects mean gray value and standard deviation of the gray values. Haralick texture features are based on co-occurrence matrices. Co-occurrence matrices are 2D histograms providing the frequency of pairs of co-occurring gray values with a given spatial relation (i.e. angle and distance). For all co-occurrence matrices (representing different angles and distances), 13 second order statistics were computed such as, correlation and contrast describing the texture of the object. Granularity features depend on the gray value differences of neighboring pixels. Wavelet features are based on a decomposition of the image into different frequency channels and provide information on image contrasts and texture. Gray scale invariant features compute the variation in gray values around a pixel. Zernike moments are used to describe the information content of an image by considering an image as a two-dimensional density distribution function. Moment sets of different orders and with different basis functions can be used to describe the information contained in an image region.

Single cell tracking was done based on the approach described in [Harder et al., 2011]. In essence, first cell-cell correspondences were determined using spatial distance and feature similarity, and second, mitosis events (cell splitting) were detected and the respective trajectories were merged [Harder et al., 2011, Harder et al., 2009]. To determine cell-cell associations, a distance measure was used, combining feature similarity and spatial distance after normalization of both terms [Harder et al., 2011]. The distance measure was computed for objects with a Euclidean distance d_{max} , of centroids in the 2D image space (i.e. maximum cell velocity), followed by local optimization. The mitosis likelihood function is based on the size and mean intensity of the mother and daughter nuclei [Harder et al., 2009]. An additional constraint was added to this mitosis likelihood function, disregarding objects with low mean intensities (compared to the mean intensity of all cells in the particular image) to avoid false positive detections.

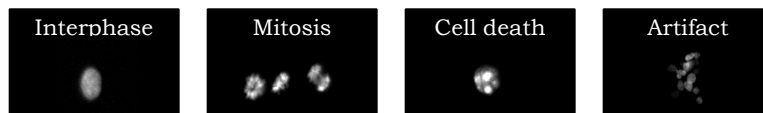


Figure 3.1: Sample images of the four phenotype classes. Interphase cells are round or elliptical with smooth boundaries. The class of mitosis includes cells in the sub-phases of the mitotic process, i.e. pro-metaphase, metaphase, and anaphase. The class cell death represents dying cells observed by disintegrated nuclei. The class artifact represents cell aggregations that could not be further segmented and over-segmented cells.

3.4 Classification of nuclei into phenotypes

Using supervised machine learning each nucleus was classified into one of the following phenotype classes: interphase, mitosis, cell death, and artifact (Figure: 3.1, page: 27). For training of the classifier a set of typical training samples from each class was collected, where each sample was defined by a vector of descriptive image features (e.g., Haralick texture, Zernike moments, Wavelet features, shape descriptors) and a class label (interphase, mitosis, cell death, artifact). The class label was provided by the annotation from an expert. A classification model (classifier) was generated from the training data to distinguish the classes defined in the training set. After training, the classifier was applied to assign class labels to nucleus images for which the classes were not yet known. Each step in this process is explained in the following.

3.4.1 Training set

The training set was manually annotated by an expert. For the SH-EP cell line, a set of 174 interphase samples, 94 mitosis samples, 204 cell death samples, and 118 artifact samples, was manually annotated for training and validating the classifier. For SK-N-BE(2)-C cells, we selected a set of 230 interphase samples, 80 mitosis samples, 120 cell death samples, 100 cluster samples, and 45 artifact samples. Since SK-N-BE(2)-C showed a much higher tendency of clustering, we separated the clustered objects from the artifact class and defined a new class called cluster. These annotation samples were taken from the images of all of the eight LabTeks to account for the variation among the cell arrays of the entire

screen. The imbalanced training set was stratified for the classifier by weighting each sample of class c by

$$w_c = \frac{n_l}{n_c} \quad (3.1)$$

where n_l is the number of samples in the largest class and n_c is the number of samples in class c .

3.4.2 Feature normalization

Feature normalization was done to bring each feature to the same numerical range. To determine the optimal feature scaling scheme, four classifiers were used, a classifier with non-normalized features and three classifiers with normalized features. The feature normalization strategies were based on z-score normalization.

$$Z = \frac{x - \mu}{\sigma} \quad (3.2)$$

where Z is the normalized value of x , μ and σ are mean and standard deviation of the feature population, respectively. Transformation parameters μ and σ are usually computed based on the training set [wei Hsu et al., 2010], as it is assumed that the training set represents the entire data. Our data is compiled from several experiments. It may happen that the training set is not ideally representing the entire data, in spite of the conscious efforts to cover the entire dataset in the training set. Thus, we adapted two more strategies to compute the transformation parameters. In all we had three normalization schemes where in the transformation parameters were derived from (1) the training set, (2) a systematically sampled dataset from different time steps of all experiments simulating the complete data set, and (3) each single image.

3.4.3 Classification model

For classification we used Support Vector Machines (SVMs) with a radial basis function (RBF) kernel. We applied a one-against-one approach for multiclass classification (i.e. binary classification between all pairs, followed by voting) as implemented in the R-package `e1071` [Meyer, 2012]. The model parameters C (cost function) and γ (kernel width) were optimized by a grid search $C = 2^1, 2^2, \dots, 2^{10}$,

$\gamma=2^{-16}, 2^{-15}..2^{-6}$ employing a 10-fold cross validation on the training data (inner loop). To choose C and γ each pair of the parameters C and γ was tested and the pair with the lowest validation error (the average number of misclassified samples) was chosen and used for training an SVM on the complete training dataset.

To estimate the performance of the classifiers, the SVMs were trained and validated by a 5-fold cross validation (outer loop). The annotated data was split into five subsets; four subsets were selected as training data and the remaining subset as test data. The whole process was repeated 5 times (outer loop) yielding performance estimations of the classifiers. For classifying new samples, new SVMs were trained with all samples from the training data.

3.4.4 Filter

Cells which could not be assigned to any phenotype with high confidence were removed based on the likelihood for their respective class label as determined by the classifier. The confidence values were obtained using the R-package e1071. A probability model was used which computes a posteriori probabilities for the multi-class problem by a quadratic optimization [T.-F. Wu, 2004]. This provides the likelihood of each class label for a sample. For ambiguous samples the likelihood values for multiple classes were similar without a clear maximum, and consequently, the classifier output was less reliable. Therefore we defined a reliability score r ; which was computed for each sample by

$$r = l_1 - l_2 \quad (3.3)$$

where l_1 and l_2 are the two highest likelihood values (predicted by the SVMs). All samples with a reliability score of $r \leq 0.2$ were discarded from the further analysis.

3.4.5 Manual evaluation of classification

For evaluating the performance of the classifier on real data (including samples which were hard to distinguish), a set s of 800 nuclei was randomly selected which included samples from each class. Set s was classified using the above

model and filter. Independently, this set was manually annotated. Single cell tracking as described in [Harder et al., 2011] was used to extract the trajectory tr of each of the selected nuclei of s . tr of a nucleus consisted of three snapshots before and after the target snapshot (i.e. the snapshot which is a part of s) and this time series was used for supporting the manual annotation of the nuclei into phenotype classes. The two labels of the samples (manual annotation, classifier) were compared. These errors were studied to formulate the correction rules as described below.

3.4.6 Automatic error correction

Classification correction was performed based on a finite state model (FSM) as described previously [Harder et al., 2009] which is described briefly in the following. Each cell was tracked over the whole time series as previously explained [Harder et al., 2011, Harder et al., 2009]. Classification results were overlaid on these trajectories resulting in a sequence comprising phenotype classes of a nucleus over time. A correction scheme was developed for better separating the class mitosis from interphase and cell death. This automatic correction scheme was aimed at (1) avoiding false negative prediction of mitosis, (2) avoiding false positive prediction of mitosis, and (3) avoiding false positive prediction of cell death. For (1), all splitting events were identified, and then the mother nucleus as well as the immediate daughter nuclei were labeled as mitosis. For (2), all nuclei classified by the classifier as mitosis were validated by inspecting any of the four conditions: (a) if it was involved in a splitting event (mother or daughter), (b) if there was a splitting event preceding or following the nucleus, (c) if the succeeding object was a cluster (a mitotic splitting event would not be detectable in a cluster), or (d) if it was followed by cell death. If none of the conditions were true, the nucleus was corrected to interphase. For (3), all the successors of the nucleus were scanned until the end of the trajectory. A nucleus was considered to be in cell death if the immediate successor of the nucleus and at least 50% of the following trajectory had the label cell death, if not, the sample was corrected to interphase.

3.5 Quantitative analysis of phenotype kinetics

After classifying each nucleus, we performed a quantitative analysis to obtain time-lapse profiles for each phenotype class and knockdown. The pipeline included the following steps:

3.5.1 Normalization

We used B-Score normalization for normalization within the LabTeks and between LabTeks, accounting for spatial error corrections of each cell array per time-lapse and per phenotype class. B-Score normalization subtracts the row mean and column mean to account for the row and column variability, followed by correction for plate deviations by subtracting the plate mean and dividing by the plate median absolute deviation [Brideau et al., 2003], i.e.

$$B_{score} = \frac{r_{RC} - (\mu_{pl} + \mu_R + \mu_C)}{MAD_{pl}} \quad (3.4)$$

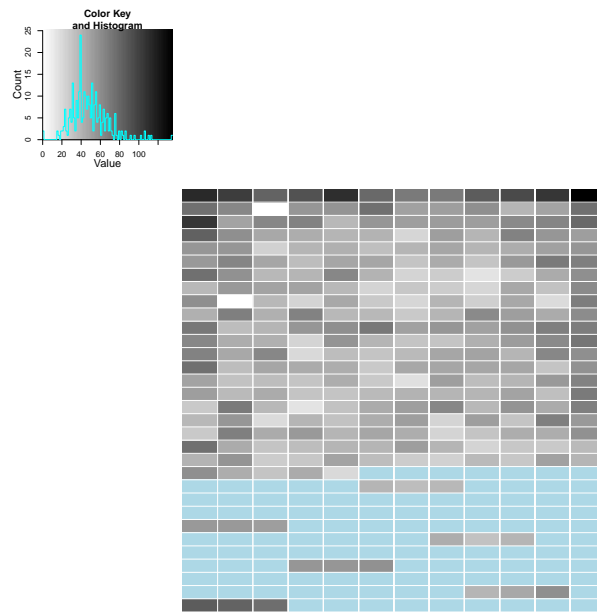
where, B_{Score} is the normalized value, r_{RC} is the raw value of plate pl at row R and column C , μ_{pl} is mean of the plate pl , μ_R is mean of row R of plate pl , μ_C is mean of column C of plate pl , MAD_{pl} is median absolute deviation.

$$MAD = median(|x_i - \mu_m|) \quad (3.5)$$

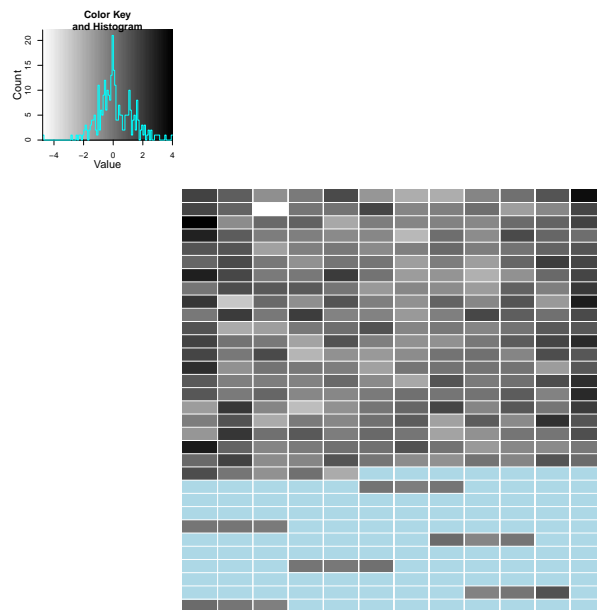
where x_i is the vector of values, μ_m is the median of x_i . Note that, the median absolute deviation is more robust than the standard deviation as the median is less sensitive to outliers [Birmingham et al., 2009]. B-score normalization also accounted for edge effects which were evident in the cell arrays before normalization (Figure: 3.2, page: 32).

3.5.2 Defining the phenotype signal

To smooth fluctuations, each phenotype class was quantified in time-frames with 24 hours of imaging data. Each time-frame had a shift of 8 hours from the previous frame, yielding 13 time-frames for the five days of screening. The area



(a) Before normalization



(b) After normalization

Figure 3.2: Cell arrays before and after normalization. The color key shows the distribution of the cell counts over the array. (a) A cell array before normalization, showing the edge effects with high cell counts in the most upper row. (b) The same cell array after B-score normalization. It shows a smoothing of the edge effects. Blue boxes represent empty spots which were not a part of the screen.

under the curve (AUC) (integral of the phenotype counts for each time-frame) was computed for each of these time-frames. AUC of a time-frame was defined as the phenotypic signal for that time-frame. AUCs were computed using the R-package caTools [Tuszynski, 2012].

3.5.3 Estimating the phenotypic score

We assigned a significance score to the phenotype signal of each time-frame in the form of p-values. We computed significance values (p-values) by a non-parametric test (Wilcoxon rank test) instead of using Z-scores, as a significant p-value (≤ 0.05) indicates reproducibility of the siRNA effect and are less sensitive to outliers [Boutros et al., 2006]. The two populations subjected to the test were four replicates of a gene per siRNA, and the overall population acting as the negative control.

3.6 Analyzing phenotype profiles

3.6.1 Clustering of phenotype profiles

For each gene, a phenotype profile was defined by the time-frame with the most significant phenotype signal and the first time-frame with a significant phenotype signal. For further analysis we considered three phenotypic scores, higher cell death, higher mitotic index and lower interphase counts than the overall population. Clusters based on phenotype profiles were generated using Euclidean distance as dissimilarity measure. Multistep bootstrap sampling was used via the R package pvclust [Suzuki and Shimodaira, 2006] to assign a confidence level to each cluster. In this sampling process many sets of bootstrap replicates with varying sample size are generated. The approximate-unbiased p-values are calculated from the change in frequency over changing sample size [Suzuki and Shimodaira, 2006]. The p-value calculated by multi-scale bootstrap sampling is less biased than the bootstrap probability. For the clustering we used, 1000 replications of bootstrap, with relative sample size varying from 0.33 to 1.33 and 95% confidence level.

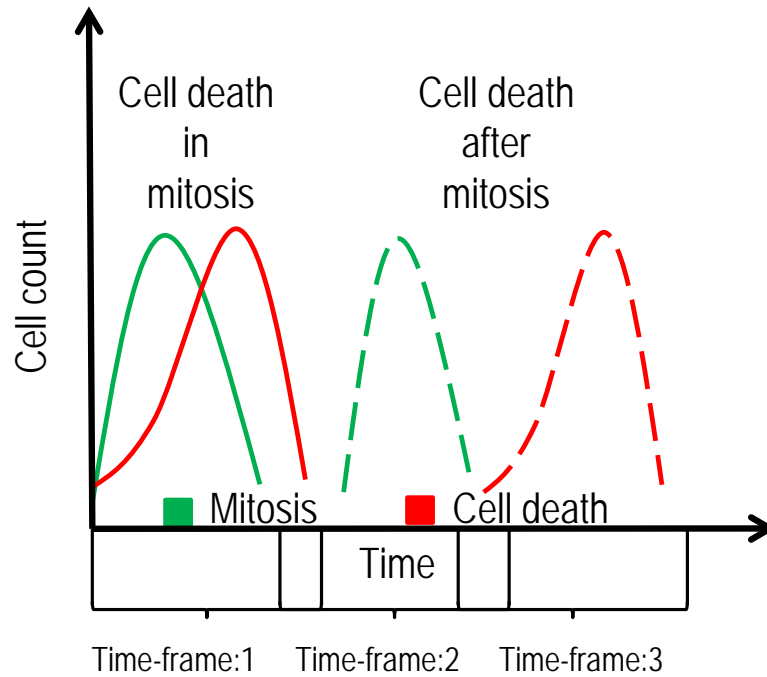


Figure 3.3: Determination of hits that show cell death during or after mitosis. The population response to a knockdown was computed to identify the sequence of phenotype occurrence.

3.6.2 Tracking of phenotype events

Genes which showed a high mitosis count as well as a high count of cell deaths were further investigated to determine the sequence of the occurrence of these phenotypes. A phenotype profile of a gene consisted of the p-value (Wilcoxon rank test, $p\text{-value} \leq 0.05$) of each time-frame for the two phenotypes under consideration. Mitotic defects were indicated by the occurrence of phenotypes in the same time-frame or occurrence of high cell death in the time-frame next to the time-frame with high mitosis counts (Figure 5). We selected the genes with significantly high occurrence of mitosis phenotypes in time-frame t_0 and significantly high occurrence of cell death phenotypes either at the same time-frame (t_0) or at time-frame t_{0+1} .

3.7 Estimating the periodicity

To estimate the periodicity of the cell lines, we performed a non-linear fit to the overall temporal distribution of all interphase counts (including all controls and knockdowns), using the `nlinfit` function of Matlab (www.matlab.com). To smooth the data, the entire time series was reduced to forty time-frames. Each frame represented integration of three hours of imaging data. For the fit, a non-linear function combining a sinus function and a linear function was used,

$$y = \alpha \sin(\beta x + \phi) + mx + c \quad (3.6)$$

where α is the amplitude, β the time period, x the interphase counts in a time series, ϕ the phase, m the linear slope and c a constant. Fitting values of the parameters were $\alpha = 0.03$, $\beta = 0.18$, $c = 0.4$, $m = 0.001$ for SH-EP cells and $\alpha = 0.02$, $\beta = 0.2$, $c = 0.3$, $m = 0.05$ for SK-N-BE(2)-C cells.

3.8 Expression analysis of the hits

Gene expression profiles for the hits were extracted from whole genome single-color microarray profiles of 478 pre-treatment primary neuroblastoma tumors analyzed as part of the MAQC-II project [Oberthuer et al., 2010]. Data was normalized using the quantile method using the R-package `limma` [Smyth, 2004]. Two tumors were removed from the survival analysis as the overall survival data and cause of death were unknown. To split the tumors into high and low risk groups we used the R-package `maxstat` [Hothorn and Lausen, 2003]. We used a 10-fold cross-validation, i.e. we divided the data set into 10 parts and used the cutoff value from 9 parts to assign the group label to the tumors of the 10th part. Overall survival analysis was performed using the R-package `survival` [Therneau, 2012]. Statistical significance of the curves was determined using the log-rank test.

3.9 Enrichment Analysis

Pathway Enrichment

Enrichment tests were done for each pathway in Reactome on the selected genes compared with all genes from the 11K microarray as background (universe) using the software DAVID [Huang da et al., 2009]. EASE Scores (from a modified Fisher's exact test) were used for obtaining the significance values [Hosack et al., 2003].

Gene Ontology enrichment

Gene Ontology enrichment analysis was performed using the Bioconductor package topGO [Alexa et al., 2006] and the weight algorithm.

Kinase enrichment

Kinase enrichment analysis was done using the Kinase Enrichment Analysis (KEA) tool. It employs a kinase-substrate database, compiled from several experimental resources (for details, see [Lachmann and Ma'ayan, 2009]). Given a list of genes, KEA identifies kinases for which a significant enrichment of their substrates can be found in the gene list (using Fisher's exact tests). P-values from all these enrichment tests were corrected for multiple testing using the method of Benjamini-Hochberg [Benjamini and Hochberg, 1995]. After multiple testing corrections, p-values ≤ 0.05 were considered significant.

3.10 Validation experiments

The validation experiments were performed by our collaborators, by Dr. Sina Gogolin and Dr. Tobias Paffhausen in the Lab of PD. Dr. Frank Westermann, at Tumor Genetics division of DKFZ. For detailed materials and methods please see Gogolin, Batra et.al Cancer Letter, 2012.

3.10.1 Ploidy and Cell cycle analysis

Ploidy analysis was done to assess the DNA index for each tumor sample. Clinical data and tumor samples from 483 patients enrolled in the German Neuroblastoma Trial and diagnosed between 1998 and 2010 were used in this study. Informed consent was collected within the trial protocol. Native cryo-conserved tumor samples were minced with scissors in 2.1% citric acid 0.5% Tween-20. Cells were permeabilised using phosphate buffer (7.2 g Na₂HPO₄ x 2H₂O in 100 ml distilled water, pH 8.0) and then treated with fluorescent dye, diamino-2-phenylindole (DAPI), to stain the DNA. DNA content of each tumor was assessed using high resolution flow cytometric analyses performed on the Galaxy pro flow cytometer (Partec, Mnster, Germany) equipped with a mercury vapor lamp 100W and DAPI filter [Ehemann et al., 1999][Ehemann et al., 2003].

Cell cycle analysis was performed to determine cells in different phases of cell cycle. Cells were prepared in the same manner as above for ploidy analysis. Cells were incubated in 75cm² flasks. After 24 hours of incubation cells were induced using doxycycline and/or treated with vincristine or doxorubicin. Cell cycle phases were identified using same cytometer as used for ploidy analysis.

3.10.2 Cytogenetic analysis

Four-color FISH analysis was performed to verify the polyploidy in the cells. Centromeric regions of chromosomes 3, 6, 8 and 18 were localized with fluorescently labeled plasmids (chromosome 3: pAE0.68 - Cy3 (GE Healthcare), chromosome 6: pEDZ6 - Cy3.5, chromosome 8: pZ8.4 FITC (Molecular Probes, Eugene, Oregon, USA), and chromosome 18: 2Xba - DEAC (Molecular Probes)) [Henegariu et al., 2000][Savelyeva et al., 2006]. The cells were imaged and analyzed using a Zeiss axiophot microscope and IPLap 10 software.

3.10.3 MAD2L1 knockdown

Two types of clones were created with pTER+ vector, one with shRNA targeting MAD2L1 (AATACGGACTCACCTTGCTTG, Gen Bank TM accession number

NM.002358) and other with control/scramble shRNA (AACAGTCGCGTTTGC-GACTGG, Ambion)[van de Wetering et al., 2003]. Two SH-EP cell lines, *SH - EP^{MYCN}* and parental SH-EP cells were transfected with doxycycline-inducible pcDNA6TR repressor, using manufacturers (Invitrogen) protocol. *SH - EP^{MYCN}* pcDNA6TR or SH-EPpcDNA6TR were transfected with pTER+ vector harboring the shRNAs using Effectene (QIAGEN, Hilden, Germany). Thus four clones were created *SH - EP^{MYCN}* shMAD2L1, *SH - EP^{MYCN}* scramble shRNA, SH-EP-shMAD2L1, SH-EP scramble shRNA. Western blotting was used to determine effective down regulation of MAD2L1 in *SH - EP^{MYCN}* shMAD2L1 and SH-EP-shMAD2L1.

CHAPTER 4

RESULTS AND DISCUSSION

Time-lapse image based RNAi screens generate multiparametric readout. We present a pipeline to analyze such a screen. We present (1) an elaborate protocol to optimize classification of phenotypes using Support Vector Machines, (2) a novel method to identify cell fate using knockdown screens, and (3) novel candidate genes whose inhibition cause mitosis-linked-cell-death. The major steps of this study are depicted in Figure: 4.1 (page: 40).

4.1 Selecting relevant genes for knockdown screening

In a previous study by Oberthuer et al.[Oberthuer et al., 2006], a predictive-signature comprising of 144 genes was established to predict the course of the disease for neuroblastoma patients. In a follow-up study by Westermann et al.[Westermann et al., 2008] a genome-wide search of MYCN/MYC target genes using a MYCN-inducible neuroblastoma cell line was performed recording time series of gene expression after MYCN induction. The profiles were clustered yielding gene sets with similar gene expression profiles. For our screen, we selected two sets of genes from these clusters, one set from clusters enriched in genes that belonged to the 144-gene predictor signature. The second set of genes was se-

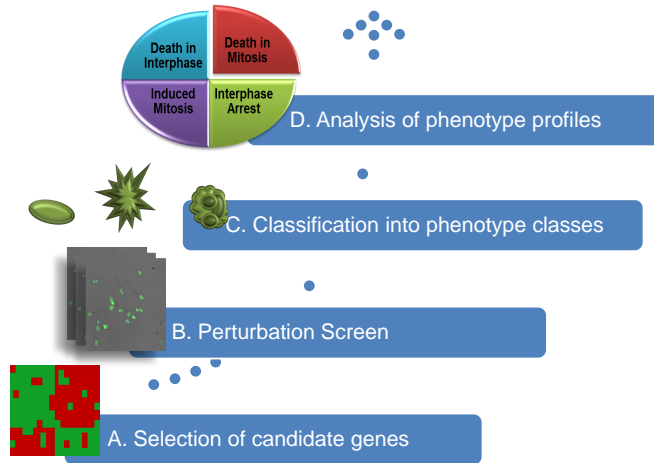


Figure 4.1: The workflow. (A) Neuroblastoma associated genes were selected based on gene expression profiles of neuroblastoma tumors and cell lines, (B) selected genes were subjected to image-based time-lapse siRNA knockdown screens, (C) each cell in an image was classified into one of the phenotype classes: interphase, mitosis, or cell death, and (D) time series of the phenotypes were assembled into phenotype profiles to determine gene function of each gene knockdown.

lected from clusters enriched ($p\text{-value} \leq 0.05$, adjusted for multiple testing) in the E-Box motif (binding motif of MYC family), indicating direct MYC family targets [Westermann et al., 2008]. Details on the selection are given in section: 3.1 (page: 23) and the gene list is provided in appendix A.

Using the selected 240 neuroblastoma associated genes, we performed enrichment tests for the pathway definitions of the Reactome database [Croft et al., 2011] and Gene Ontology (www.geneontology.org). We found four Reactome pathways to be significantly enriched with the candidate genes comprising cell cycle associated pathways (mitotic cell cycle, cell cycle checkpoints and APC-Cdc20 mediated degradation of Nek2A) (Table: 4.1, page: 41). For Gene Ontology, four out of the top five Gene Ontology terms were linked to cell cycle (mitosis, cell division, mitotic spindle organization, and mitotic cell cycle checkpoint), demonstrating that the gene selection procedure properly captured genes relevant to the cell cycle of neuroblastoma.

Table 4.1: Functional enrichment of screened genes. Pathways of Reactome and gene groups from Gene Ontology which were enriched in the screened genes.

ID	Term description	Number of candidate genes in pathway/process	P-value
<i>Reactome pathway</i>			
152	Cell Cycle	37	4.2E-14
1538	Cell cycle checkpoints	14	4.78E-05
8017	APC-CDC20 mediated degradation of NEK2A	5	0.0070
1698	Metabolism of nucleotides	8	0.030
<i>Gene Ontology</i>			
7067	Mitosis	47	7.61E-26
51301	Cell division	45	3.89E-22
7052	Mitotic spindle organization	09	5.66E-08
7093	Mitotic cell cycle checkpoints	11	1.15E-06
6260	DNA replication	23	1.49-06

4.2 Data and Data processing

The screen was performed in two neuroblastoma cell lines, SH-EP and SK-N-BE(2)-C (Figure: 4.2, page: 42). SH-EP has a single copy MYCN and functional p53. SK-N-BE(2)-C has MYCN amplification and p53 mutation. The screen was conducted for 120 hours with four LabTeks having 275 spots per cell line. Images were taken every 35-40 min, generating 180-220 images per spot, resulting in 55000 image sequence (total number of images: 440000). Both cell lines stably expressed GFP tagged histones, which enabled us to track the nuclei in the images. Each single cell nucleus was segmented from images and characterized by texture descriptors, e.g. Haralick texture, Zernike moments, granularity, grayscale invariants, wavelet features and by morphological descriptors, e.g. shape, size and circularity.

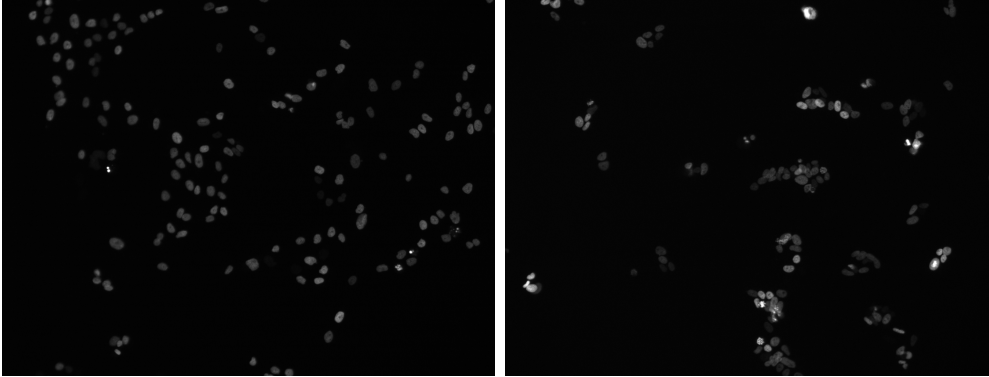


Figure 4.2: Sample images of cell lines. On the left is an image of SH-EP cells, and on the right of SK-N-BE(2)-C cells.

4.3 Automated classification of cellular phenotypes

To track mitotic events, each cell was classified into one out of four distinct phenotype classes (Figure: 3.1, page: 27): interphase (round or elliptical object with smooth boundaries), mitosis (dividing cell comprising prometaphase, metaphase, and anaphase), cell death (small and bright fragments of the nuclei), and artifact (clusters of cells that could not be further subdivided, or small-dark objects; these objects were not used for a further functional analysis). The classifier was trained using Support Vector Machines (SVMs) to distinguish these four phenotype classes on a training set of manually annotated nuclei.

4.3.1 Classification performance based on training set

To assess the performance of the classifiers, a cross-validation procedure was performed, yielding an overall accuracy of 95.3% for the SH-EP cell line and 87% for SK-N-BE(2)-C. These results outperform previous investigations with HeLa cells (accuracy for HeLa cells: 93.9% to 94.7% [Harder et al., 2008]) even though imaging and image analysis of neuroblastoma cells was more challenging due to the higher tendency to cluster and higher cell motility. Note that the stated performance of our approach was determined using well separable objects of the training set. The confusion matrices of the two neuroblastoma cell lines are given

in Table: 4.2 (page: 43).

Table 4.2: Confusion matrix of the classification results after 5-fold cross validation and non-normalized features

(a) SH-EP cell line					(b) SK-N-BE(2)-C cell line				
True	Predicted				True	Predicted			
	I	M	A	Ar		I	M	A	Ar
I	163	3	0	4	I	223	4	1	2
M	0	79	11	0	M	19	52	9	0
A	0	21	179	0	A	28	14	78	0
Ar	9	0	0	106	Ar	31	0	0	69

4.3.2 Optimization of feature normalization

Feature normalization was performed to avoid dominance of features with larger numerical range over features with smaller range [wei Hsu et al., 2010]. A classifiers performance can be significantly improved by scaling the features before classification [Hur and Weston, 2011]. Normalization can further improve the performance when the data is generated in several experiments and the given training set may not contain samples from all the experiments. The normalization was performed using Z-score normalization which shifts the distribution of feature values to a mean value of 0 and standard deviation of 1. This normalization brings the features in comparable numerical ranges. We generated four classifiers three with three different normalization schemes and one without any normalization. For details on the normalization schemes please refer to section: 3.4.2 (page: 28). The normalization of features improved the results over the non-normalized features (Table: 4.3, page: 44). Figure: 4.3 (page: 45) shows that normalization with transformation parameters derived from systematically sampled objects showed best performance parameters in case of SH-EP cells. For SK-N-BE(2)-C cells, normalization of each image independently gave best results.

Table 4.3: Confusion matrix of the classification results after 5-fold cross validation and normalized features. (a, b) with features normalized by training set, (c, d) with features normalized by each image, (e, f) with features normalized by systematically selected samples.

(a) SH-EP cell line

True	Predicted			
	I	M	A	Ar
I	155	5	0	10
M	2	78	10	0
A	1	11	186	2
Ar	13	0	0	102

(b) SK-N-BE(2)-C cell line

True	Predicted			
	I	M	A	Ar
I	226	1	0	3
M	4	63	13	0
A	3	18	99	0
Ar	8	0	0	92

(c) SH-EP cell line

True	Predicted			
	I	M	A	Ar
I	159	1	2	8
M	5	77	7	1
A	1	12	186	1
Ar	16	0	0	99

(d) SK-N-BE(2)-C cell line

True	Predicted			
	I	M	A	Ar
I	221	0	1	8
M	2	67	11	0
A	5	14	99	2
Ar	6	0	0	94

(e) SH-EP cell line

True	Predicted			
	I	M	A	Ar
I	155	6	0	9
M	1	85	4	0
A	1	9	189	1
Ar	17	0	0	98

(f) SK-N-BE(2)-C cell line

True	Predicted				
	I	M	A	C	Ar
I	220	3	3	2	2
M	3	64	13	0	0
A	9	26	85	0	0
C	5	0	0	95	0
Ar	1	0	0	0	44

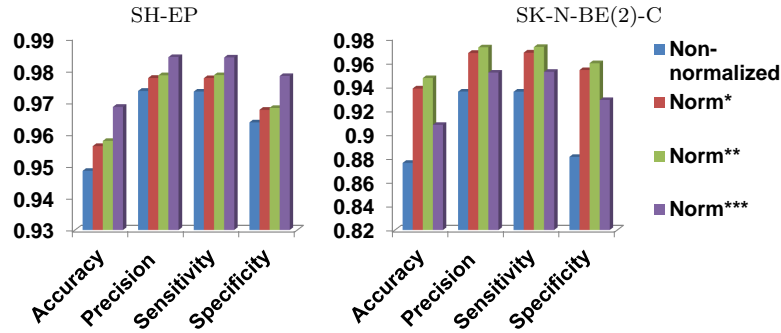


Figure 4.3: Classification performance after 5-fold cross validation with normalized and non-normalized features. Norm* - normalized by training set, Norm** - normalized by each image independently, Norm*** - normalized by systematically sampled objects.

4.3.3 Manual evaluation of classification performance

To determine performance on all objects including hardly distinguishable samples, we randomly selected a test set comprising of any segmented objects from our data. Manual verification of the classified phenotypes showed that our classifiers well distinguished the phenotypes. Nevertheless, separation of mitosis and interphase and of mitosis and cell death was comparably low. The performance of these four classifiers were compared and the best performing one was used (Table: 4.4, page: 46). For the SH-EP cell line none of the normalization schemes proved better than non-normalized features (Figure: 4.4, page: 45). For SK-N-BE(2)-C the normalization scheme based on systematically sampled set was marginally better (Figure: 4.4, page: 45).

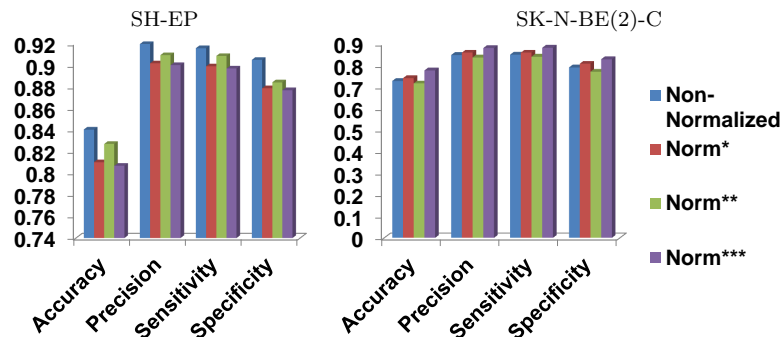


Figure 4.4: Manually evaluated performance of the classifiers. Norm* - normalized by training set, Norm** - normalized by each image independently, Norm*** - normalized by systematically sampled objects.

Table 4.4: Confusion matrix of the classification results after manual evaluation. (a, b) with features normalized by training set, (c, d) with features normalized by each image, (e, f) with features normalized by systematically selected samples, (g, h) non-normalized.

(a) SH-EP cell line

True	Predicted				
	I	M	A	Ar	D
I	193	64	15	42	17
M	3	25	10	0	3
A	3	25	119	7	4
Ar	39	52	36	137	5

(b) SK-N-BE(2)-C cell line

True	Predicted				
	I	M	A	Ar	D
I	142	38	18	30	6
M	0	23	6	0	1
A	20	41	68	4	7
Ar	32	10	16	115	9

(c) SH-EP cell line

True	Predicted				
	I	M	A	Ar	D
I	214	56	19	31	11
M	5	22	11	0	3
A	3	20	124	9	2
Ar	49	29	38	127	26

(d) SK-N-BE(2)-C cell line

True	Predicted				
	I	M	A	Ar	D
I	122	19	51	35	7
M	2	12	11	3	2
A	19	26	77	10	8
Ar	45	10	15	107	5

(e) SH-EP cell line

True	Predicted				
	I	M	A	Ar	D
I	194	68	14	44	11
M	3	25	11	1	1
A	3	25	121	7	2
Ar	42	47	38	132	10

(f) SK-N-BE(2)-C cell line

True	Predicted					
	I	M	A	C	Ar	D
I	124	24	10	31	29	16
M	1	17	1	0	7	4
A	19	31	57	4	21	8
C	38	6	10	106	9	13
Ar	57	21	29	8	88	10

(g) SH-EP cell line

True	Predicted				
	I	M	A	Ar	D
I	217	60	10	27	17
M	3	28	6	0	4
A	0	27	121	6	4
Ar	55	40	31	120	23

(h) SK-N-BE(2)-C cell line

True	Predicted				
	I	M	A	Ar	D
I	137	32	24	33	8
M	0	19	9	0	2
A	21	40	67	5	7
Ar	31	6	16	120	9

4.3.4 Automatic error correction

In order to improve the separation of the challenging cases such as Mitosis \leftrightarrow Interphase and Mitosis \leftrightarrow Cell death, we designed an automated correction scheme employing tracking information (section: 3.4.6, page: 30). In addition, we applied a filter to discard objects for which the predicted phenotypes were ambiguous. As a result, the filter discarded 8% of the objects. For SH-EP cells, these automatic corrections improved overall accuracy from 84% to 89%. Precision and sensitivity increased from 91% to 94%. These corrections did not improve the results for SK-N-BE(2)-C cells (Figure: 4.5).

Table 4.5: Confusion matrix of the classification results after automatic corrections. These corrections were performed with the normalization which performed the best for the cells lines.

True	Predicted			
	I	M	A	Ar
I	241	5	28	25
M	11	16	5	0
A	10	12	119	6
Ar	63	10	33	104

True	Predicted					
	I	M	A	Ar	C	D
I	124	24	10	29	31	16
M	1	17	1	7	0	4
A	19	31	57	21	4	8
Ar	57	21	29	88	8	10
C	38	6	10	9	106	13

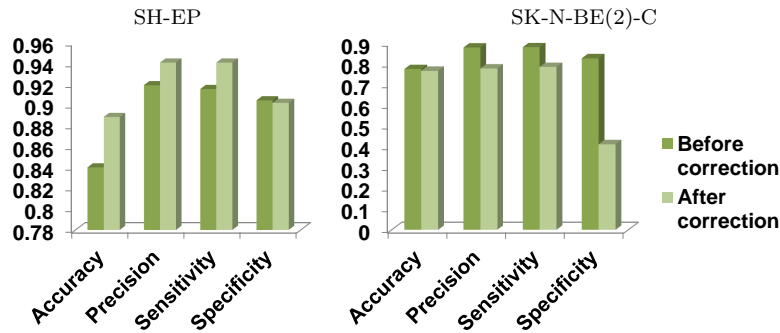


Figure 4.5: Classification performance after automatic correction.

4.3.5 Summary

In summary, we obtained reliable results by improving automated classification of phenotypes from image data of neuroblastoma cell lines. We have modified the protocol for nuclei classification. As shown in Figure: 4.6 (page: 48), the usual procedure to assess the performance of a classifier is shown in light green, where as we added the steps shown in dark green. We have added a validation step in which the classifiers performance was verified on realistic data distributions and correction rules were formulated to rectify the errors observed in the manual verification process.

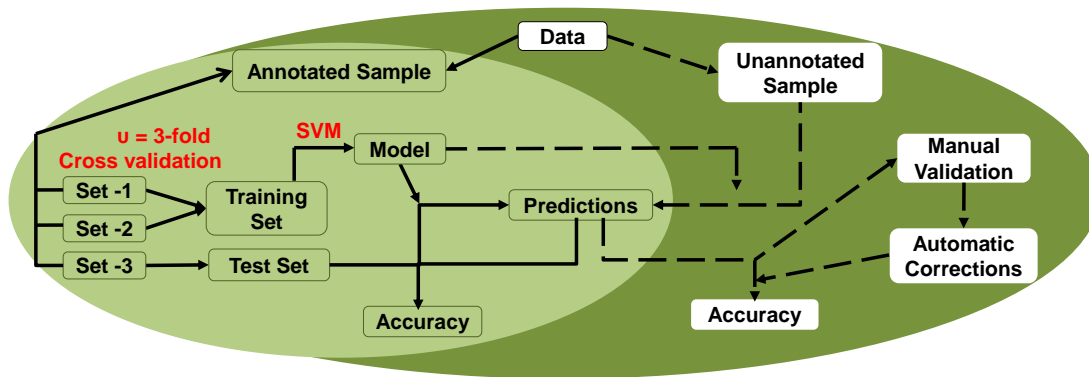


Figure 4.6: Protocol for nuclei classification. The light green portion indicates the common steps taken in phenotype classification. The dark green part represents the additional steps that we followed for optimization and improvement.

4.4 Quality control of the experimental set-up

As a quality control, we compared proliferation dynamics of positive and negative controls over the entire period of the screening (Figure: 4.7, page: 50). We selected positive controls with a distinct apoptotic phenotype [Neumann et al., 2006] in pilot screen (KIF11, PLK1, INCENP, data not shown). We used these as positive controls and two scramble siRNA constructs as negative controls. To obtain a measure of proliferation dynamics, we counted the number of interphase cells in each image over the investigated time-frames. For the SH-EP cell line, we found a significantly reduced proliferation of the positive controls in comparison to the negative controls in all time-frames (p -value ≤ 0.05). For the other cell

line (SK-N-BE(2)-C), we found a significantly reduced proliferation in the later time-frames (56-120 hours, Figure: 4.7, page: 50) indicating a delayed effect of the perturbation.

4.5 Estimating cell cycle kinetics

Cell cycle kinetics has been used as a parameter for optimization of cancer treatment schedules. Interestingly, treatment schedules matching the integer multiple of the cell cycle duration reduces damage to normal cells [Bernard et al., 2010]. We were interested if our time series analysis allowed us to estimate the cell cycle duration of our cell lines. We examined the cell cycle behavior of the cell culture, assuming that siRNA transfection causes synchronization of the cells. The cell cycle duration of a cell line can be computed either by the mitotic index or by S-phase dynamics [Baguley and Marshall, 2004]. In our approach, interphase phases G1, G2, and S were not differentiated therefore we studied the interphase dynamics as a whole. The interphase population was averaged over all replicates and knockdowns. In accordance with our expectation we observed periodicity. We identified a cell cycle duration of 35 hours for SH-EP cells (Figure: 4.8, page: 51) and of 31 hours for SK-N-BE(2)-C (Figure: 4.8, page: 51). Note that in earlier studies using HeLa cells, shorter cell cycle duration of 17 hours was reported [Zhang et al., 2011]. Our finding that neuroblastoma cells synchronize as well opens the possibility to study population response dynamics for each knockdown (next sections).

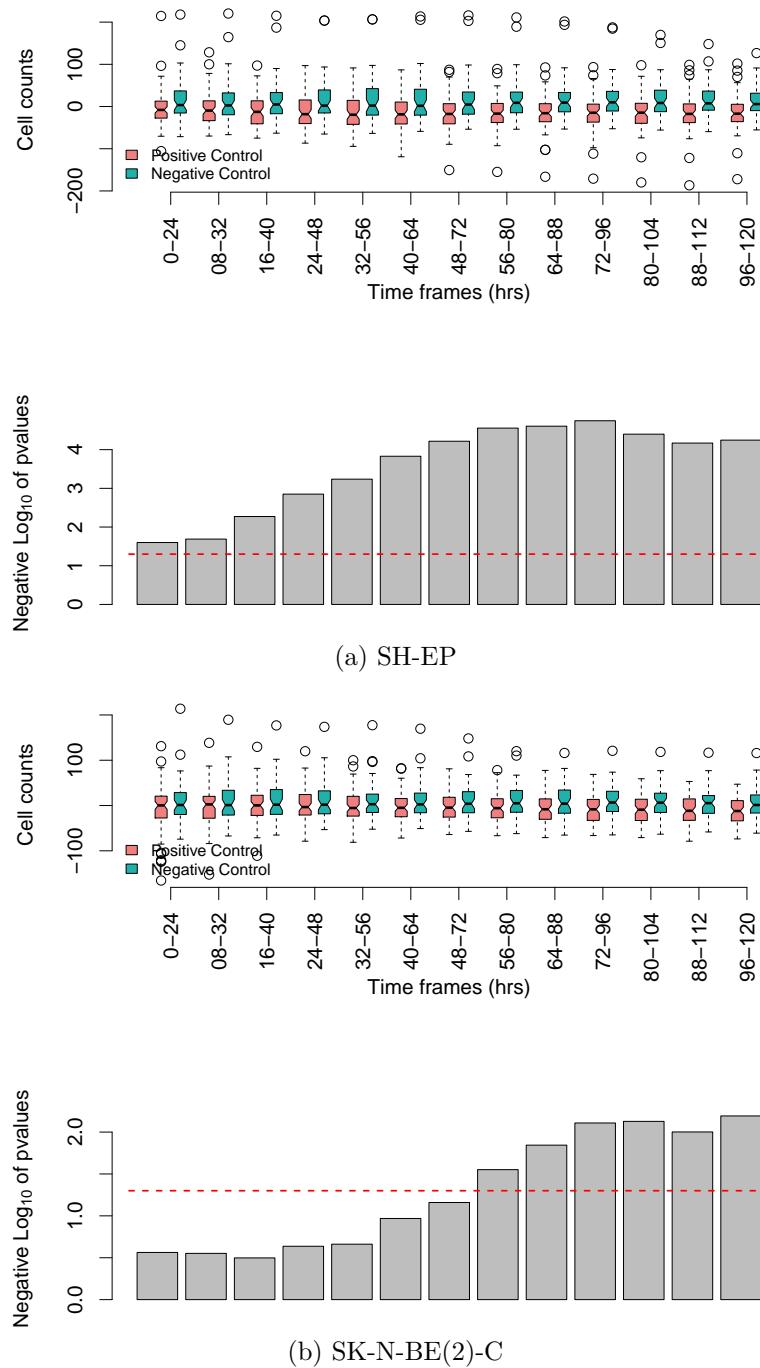
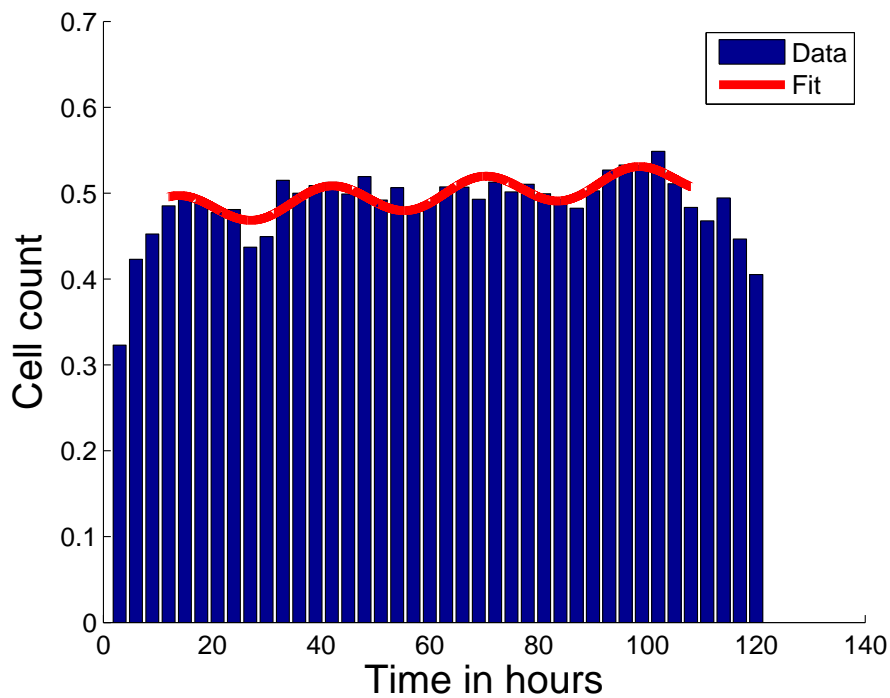
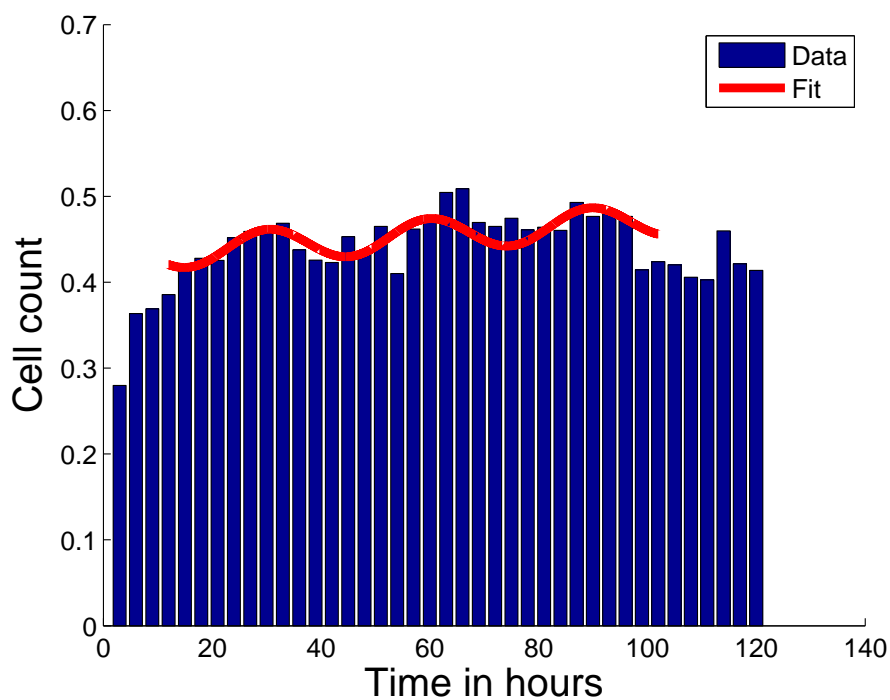


Figure 4.7: Experimental quality control. Top: cell counts of positive (coral red) and negative (coral blue) controls are plotted by boxplots for all time-frames. Bottom: Significance values of the differences of positive and negative controls are given for each time-frame by negative log₁₀ p-values. A significance threshold (p-value = 0.05) is indicated by a red dashed line. SH-EP cell line: The positive controls show significant lower counts for all time-frames. SK-N-BE(2)-C cell line: The positive controls show significant lower counts for time frames post 56 hrs.



(a) SH-EP



(b) SK-N-BE(2)-C

Figure 4.8: Time series of interphase cells during five days of screening. The SH-EP population shows a periodicity of ~ 35 hours and SK-N-BE(2)-C population show periodicity of ~ 31 hours representing the cell cycle duration (blue bars: interphase counts (normalized by B-Score normalization) of all screened cells for each time-frame, red curve: fitting curve)

4.6 Clustering phenotype profiles

Clusters of genes with similar phenotype profile will be denoted as phenoclusters in the following. We generated phenoclusters for both cell lines using the phenotypes of interest i.e. low cell proliferation, high cytotoxicity and high mitosis. Note that the phenotype was observed in the context of the population response with each time-frame. The phenotypes observed in each gene and each cell line is give in Appendix B and C. The clustering resulted in seven clusters with the following phenotypes. Each cluster was hypothetically associated with a possible phenotype as shown in the Figure: 1.1 (page: 5).

1. Interphase arrest: As the interphase count is low but high cell death is not observed we defined this knockdown effect with low interphase count as the interphase arrest phenotype.
2. Mitotic arrest: Mitosis is a 20 min process in a cell cycle of 24 hrs. We have a time-lapse of 35-40 min, which indicates the mitotic cells captured in the images are likely to be prolonged or in arrest. Thus we defined the knockdown effect with high mitotic index as the mitotic arrest phenotype.
3. Cytotoxic: The knockdown with high cell death was defined as the cytotoxic phenotype.
4. Mitosis-linked-cell-death: The knockdown that causes both high mitosis and high cell death phenotypes was called mitosis-linked-cell-death phenotype.
5. Secondary apoptosis: The knockdown with low interphase count and high cell death was defined as the secondary apoptosis phenotype.
6. Mixed*: The knockdown with low interphase count and high mitotic count was labeled mixed*. It may happen that the cells in interphase arrest escape into mitosis and arrest or vice versa.
7. Mixed**: Also the knockdown with low interphase count, high mitotic count and high cell death was labeled mixed**.

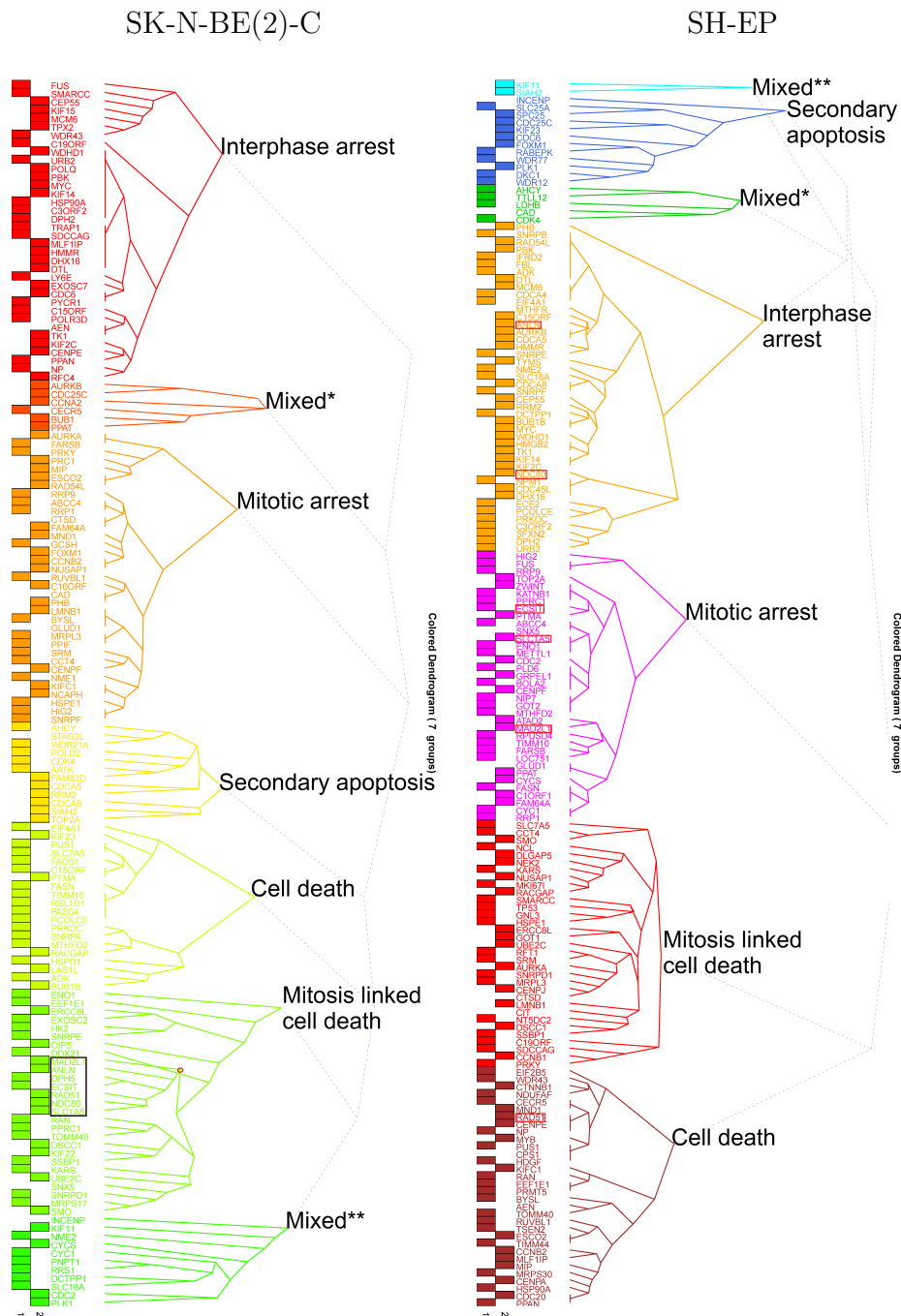


Figure 4.9: Phenoclusters. Clusters based on phenotypes of interest. The rectangular boxes marked 1 and 2 represent the two sources of selection of genes as explained in section: 3.1 (page: 23). The black box marked in the SK-N-BE(2)-C cluster represents the genes of interest, phenotypes of these genes are also marked in the SH-EP cluster by red boxes.

4.6.1 Phenocluster: Knockdowns dependent on MYCN and p53

Two cell lines were used in this screen, SH-EP with single copy MYCN and wild type p53 and SK-N-BE(2)-C with MYCN amplification and mutated p53. From the phenotype cluster of SK-N-BE(2)-C (while focusing on mitosis-linked-cell-death phenocluster) we identified a sub cluster of seven genes that show mitosis-linked-cell-death phenotype in SK-N-BE(2)-C (black box in Figure: 4.9, page: 53). The sub cluster consists of MAD2L1, ANLN, NCD80, RAD51, DPH5, ECSIT and SLC1A5. DPH5, ECSIT and SLC1A5 are not associated with mitosis or related functions. MAD2L1, ANLN, NCD80, RAD51 are functionally associated with Metaphase-Anaphase (M-A) checkpoint regulation, component of anaphase promoting complex, mitotic spindle checkpoint signaling, and DNA repair, respectively [Shah and Cleveland, 2000, Zhao and Fang, 2005, Martin-Lluesma et al., 2002, Thacker, 2005]. Phenotype profile tracking shows that out of the seven genes RAD51 does not exhibit mitosis-linked-cell-death, instead these are two independent events happening at distant time-frames. Thus, we skipped it in further analysis. We compared the phenotypes these genes exhibit in SH-EP cells. All the six genes show mitotic arrest or interphase arrest (red boxes in Figure: 4.9, page: 53).

Next, we wanted to determine if the MYCN amplification or the dysfunctional p53-p21 signaling plays a role in the knockdown phenotype. We started with MAD2L1, an M-A checkpoint gene, which when knockdown leads to mitosis-linked-cell-death in neuroblastoma cells with amplified MYCN and defective p53-p21 signaling system, i.e. SK-N-BE(2)C cells. To this end, we used two cell lines SH-EP with single copy of MYCN and SH-EP^{MYCN} which stably expresses MYCN transgene. Note that both these cell lines have functional p53-p21 signaling. To test the effect of deregulated M-A checkpoint in these cell lines we used doxycycline inducible shRNA targeting MAD2L1. With this we tested the effect of deregulation of M-A checkpoint in a functional p53-p21 scenario. The proportion of cells in different phases of cell cycle, in both these scenarios SH-EP-shMAD2L1 and SH-EP^{MYCN}-shMAD2L1, were same. These results show that MYCN amplification alone does not induce M-A checkpoint failure.

Further, to test the effect of weakened p53-p21 signaling, we treated the cells with vincristine. Vincristine disrupts microtubule formation and nuclear accumulation of p53, thus mimics weak p53-p21 signaling. Flow cytometric analysis reveals the presence of 8N cells in SH-EP^{MYCN}-shMAD2L1, which is indicative of cycling tetraploidy. 5% of the cells in SH-EP-shMAD2L1 were also in 8N stage. Tetraploidy may arise due to cytokinesis failure or mitotic slippage. The tetraploidy was further confirmed by cytogenetic analysis. This shows that MYCN amplification and nonfunctional p53 signaling together support aneuploidy.

MAD2L1 protein is involved in mitotic spindle assembly checkpoint. It prevents the onset of anaphase until all the chromosomes are properly aligned on the metaphase plate, i.e. it induces an anaphase stop signal if even a single kinetochore is unattached. On one hand, partial loss of MAD2L1 also induces aneuploidy and results in tumorigenesis [Michel et al., 2004]. On the other hand, MAD2L1 up regulation mediates chromosome instability, such as tetraploidy, in absence of p53 [Schvartzman et al., 2011]. Michel et. al [Michel et al., 2004], show that complete elimination of MAD2L1 in certain cancer cells leads to p53 independent cell death. Thus, above experiments show that inhibition of MAD2L1 induces tetraploidy in neuroblastoma cells in presence of MYCN amplification and dysfunctional p53-p21 signaling, i.e. p53 inactivating mutation.

4.7 Temporal tracking of phenotype events

Death in mitosis, i.e. cell death before completion of the mitotic process, has been reported as the most promising component in cell cycle for drug design [Manchado et al., 2012]. This can be explained by the concept that inhibitors affecting the initial phase of the cell cycle lead to cells in quiescence. Inhibitors leading to high cell death on the other hand also affect normal cells causing severe side effects during therapy. Therefore, we tracked the sequence of phenotypes in the population to select genes either with a high number of cells in mitosis and cell death at the same time-frame or a high number of cells in mitosis followed by cell death (Figure: 3.3, page: 34). Accordingly, these genes either showed mitotic cell death or mitotic slippage preceding cell death. Table: 4.6 (page: 56)

lists the genes which cause mitosis-linked-cell-death upon knockdown.

Table 4.6: Genes which cause mitosis-linked-cell-death phenotype.

SH-EP		SK-N-BE(2)-C	
Gene Symbol	Entrez Id	Gene Symbol	Entrez Id
AEN	64782	ANLN	54443
ATAD2	29028	CDC2	983
AURKA	6790	CYC1	1537
C19orf48	84798	CYCS	54205
CCNB1	891	DCTPP1	79077
CENPJ	55835	DDX21	9188
CIT	11113	DLGAP5	9787
CTSD	1509	DPH5	51611
DLGAP5	9787	DSCC1	79075
DSCC1	79075	ECSIT	51295
GNL3	26354	EEF1E1	9521
GOT1	2805	ENO1	2023
KARS	3735	ERCC6L	54821
LMNB1	4001	FAM64A	54478
MKI67IP	84365	FASN	2194
MRPL3	11222	GOT2	2806
MTHFD2	10797	HK2	3099
NCL	4691	INCENP	3619
NEK2	4751	KIF22	3835
NT5DC2	64943	MAD2L1	4085
NUSAP1	51203	MND1	84057
RACGAP1	29127	MRPS17	51373
RFT1	91869	NDC80	10403
SMARCC1	6599	NME2	4831
SMO	6608	OIP5	11339
SNRPD1	6632	PLK1	5347
SRM	6723	PNPT1	87178
SSBP1	6742	PPRC1	23082
TP53	7157	RAN	5901
UBE2C	11065	RRS1	23212
		SLC1A5	6510
		SMO	6608
		SNRPD1	6632
		SSBP1	6742
		TOMM40	10452
		UBE2C	11065

4.7.1 Predicting upstream kinase regulators

Protein phosphorylation by kinases is a common regulatory mechanism in signaling of cell cycle progression and mitotic processes. The fact that most tumors show alterations makes kinases attractive therapeutic targets [Harrison et al., 2009]. We performed statistical enrichment analysis (using KEA [Lachmann and Ma'ayan, 2009]) for the proteins encoded by candidate genes for being substrates of the regulatory kinase (Table: 4.7, page: 57). We focused this prediction on genes that cause mitosis-linked-cell-death phenotype upon knockdown. In both cell lines the Aurora kinase family showed a significant enrichment of substrates among our candidate genes. For the SH-EP cell line, the top three kinase families identified were AUR, GSK and CDK (p-value: 0.0003, 0.005 and 0.006, respectively). AUR and CDK kinase families have been well associated with neuroblastoma. GSK has been found recently to be associated with neuroblastoma. In the following sections we discuss these three kinase families and targets of GSK family in detail.

Table 4.7: Upstream kinase enrichment of candidate genes

Kinase family	Target genes in candidate list	P-value
<i>SH-EP</i>		
AUR	AURKA, TP53, RACGAP1	3.00E-04
GSK	MKI67IP, TP53, LMNB1, NCL, SMARCC1	0.005
CDK	CCNB1, LMNB1, MYB, NCL, RAC-GAP1, SMARCC1, TP53	0.006
RCK	SMARCC1, TP53, LMNB1, CIT	0.015
RSK	CCNB1, TP53	0.013
PKA	TP53, LMNB1, AURKA	0.03
CAMLK	TP53, LMNB1	0.044
MAPK	NEK2, TP53, LMNB1, NCL, SMARCC1	0.044
<i>SK-N-BE(2)C</i>		
WEE	CDC2, PLK1	0.0045
NEK	NDC80, RAN	0.021
AUR2	NDC80, INCENP	0.02

AUR family

The aurora kinase family includes Aurora kinases A, B and C. These are known to

regulate cellular division, chromosome segregation, spindle integrity and centrosome regulation [Carmena and Earnshaw, 2003]. Aurora kinase inhibitors have a strong therapeutic potential. Inhibitors of these kinases have been designed and clinical trials are undergoing [Kitzen et al., 2010]. MLN8054 is a selective inhibitor of AURKA which is in preclinical trials for neuroblastoma. AURKB can be inhibited by AZD1152 which is selectively cytotoxic to neuroblastoma tumor-initiating cells. This inhibitor currently is in a clinical trial for acute myelogenous leukemia [Morozova et al., 2010].

CDK family

Cyclin-dependent kinases (CDKs) require cyclin subunits for their kinase activity. CDKs include CDK 1-14, 16, 20 and they are involved in each phase of the cell cycle. CDKs are often deregulated in cancer and have been extensively studied as therapeutic targets. Their inhibitors are in clinical trials [Fu, 2010, Malumbres and Barbacid, 2007, Shapiro, 2006]. Roscovitine is a selective CDK inhibitor. It is in late phase-II trials against non-small cell lung cancer and nasopharyngeal cancer. Roscovitine and its analog CR8 induce cell death in neuroblastoma cells by down-regulating the CDK dependent survival factor Mcl-1 [Bettayeb et al., 2010].

GSK family

Interestingly, we also found the GSK family, which has not been associated with neuroblastoma therapy, as prominently as the CDKs and AURs. The family of GSKs consists of multifunctional serine-threonine kinases GSK3 α and GSK3 β [Doble and Woodgett, 2003]. Their role in cancer and chromosome assembly on the metaphase plate has been recently discovered [Korur et al., 2009, Wakefield et al., 2003, Wang et al., 2008]. It has been shown that GSK3 β inhibition leads to G2/M accumulation and increased apoptosis in the neuroblastoma cell line SK-N-SH [Dickey et al., 2011]. In glioma cells, inhibition of GSK3 induces pro-apoptotic effects, inhibits pro-survival signals, and induces mitochondrial permeability [Kotliarova et al., 2008].

GSK target genes among our candidate genes are NIFK, LMNB1, NCL, SMARCC1 and TP53. NIFK interacts with the forkhead-associated domain of the Ki-67 antigen in a mitosis-specific manner. It is a putative RNA-binding protein and may play a role in mitosis and cell cycle progression [Takagi et al., 2001].

LMNB1 belongs to the lamin protein family that forms the nuclear membrane. Lamins are essential for various nuclear functions, like the assembly of the nuclear envelope, and DNA replication, and they provide structural integrity and support [Gruenbaum et al., 2005]. Depletion of lamin B results in a disorganized spindle and spindle poles, chromosome mis-segregation and prolonged prometaphase [Tsai et al., 2006]. NCL (nucleolin) is a phosphoprotein abundantly found in the nucleolus. It is involved in ribosome biogenesis, cell proliferation and growth, embryogenesis, cytokinesis and nucleogenesis [Ginisty et al., 1999, Morimoto et al., 2007, Srivastava and Pollard, 1999]. SMARCC1 is a member of the SWI/SNF family of proteins, with helicase and ATPase activities. Its transitional inactivation and reactivation is required for the formation of a repressed chromatin structure during mitosis [Sif et al., 1998]. It has been associated with colorectal cancer and outcome of the disease. High levels of SMARCC1 proteins are associated with better overall survival [Anderson et al., 2009]. TP53 is well known to regulate cell cycle arrest, apoptosis, senescence, DNA repair, and changes in tumor metabolism [Kohn, 1999].

In summary, the three families of kinases are substantially involved in regulation of the cell cycle and have therapeutic potential. The substrates of these kinases need to be explored further for their role in neuroblastoma and its therapy.

4.7.2 Comparing the two neuroblastoma cell lines

We compared the candidate genes which exhibit mitosis-linked-cell-death in SHEP and SK-N-BE(2)-C to find common genes (Table: 4.6, page: 56). We found 6 such genes common in the two cell lines: DSCC1, DLGAP5, UBE2C, SSBP1, SNRPD1, and SMO. These genes showed similar phenotype in both neuroblastoma cell lines, therefore it can be said that these phenotypes are independent of MYCN copy number and p53 functional state. We did not find a corresponding phenotype in a genome-wide HeLa cell screen (Mitocheck database, <http://www.mitocheck.org/cgi-bin/mtc>). It indicates that the phenotype may be specific to neuroblastoma cells or the difference in screening and analysis played a major role. In the following sections, we discuss the literature summary and the gene expression analysis of these hits.

Literature report

A functional interpretation of the six identified genes is given in the following:

1. DLGAP5 (Discs, Large homolog-Associated Protein 5) is a known mitotic regulator. It stabilizes microtubules and ensures bipolar spindle formation. AURKA regulates its activity by phosphorylation [Wong et al., 2008]. DLGAP5 depleted HeLa cells have shown a delay in mitotic progression and their mitotic exit resulted in an unequal segregation of chromosomes [Wong and Fang, 2006].
2. SSBP1 (Single-Stranded DNA Binding Protein 1) is a housekeeping gene associated with mitochondrial biogenesis. It interacts with tumor-suppressor TP53 to enable DNA repair in mitochondria during oxidative stress [Wong et al., 2009]. Its inhibition causes genomic instability and negatively affects cell cycle checkpoint activation [Richard et al., 2008].
3. SNRPD1 encodes a small nuclear ribonucleoprotein that belongs to the SNRNP core protein family. It acts as a charged protein scaffold to promote SNRNP assembly and it strengthens SNRNP-SNRNP interactions through non-specific electrostatic contacts with RNA [Yamanaka et al., 2000]. snRNPs are major components of the spliceosome [Nilsen, 2003].
4. UBE2C is an E2 ubiquitin-conjugating enzyme. It is required for degradation of mitotic cyclins and for cell cycle progression [Yamanaka et al., 2000]. Its knockdown in U251 glioma cells results in arrest at G2/M phase and apoptosis through induction of Bax and p53 [Jiang et al., 2010].
5. DSCC1 (Defective in Sister Chromatid Cohesion 1 homolog) is one of the components of the replication factor C (RFC) complex with an important role during S phase of the cell cycle. DSCC1 double mutants terminated proliferation and showed premature senescence (increased size, flattened morphology) [Terret et al., 2009].
6. SMO (Smoothed) is a G protein-coupled receptor that interacts with PTCH, a receptor for hedgehog proteins. The hedgehog signaling pathway regulates cell proliferation, differentiation and tissue patterning during

embryonic development [Pasca di Magliano and Hebrok, 2003]. SMO has been identified as a potential drug target in osteosarcoma, as its inhibitor cyclopamine promotes G1 arrests and represses expression of cyclin D1, cyclin E1, SKP2, and pRb [Hirotzu et al., 2010]. Deregulation of the hedgehog signaling pathway has been discovered in brain, lung and skin cancers [Pasca di Magliano and Hebrok, 2003]. Inhibitors targeting SMO for curing medulloblastoma tumors are in clinical trials [Yauch et al., 2009].

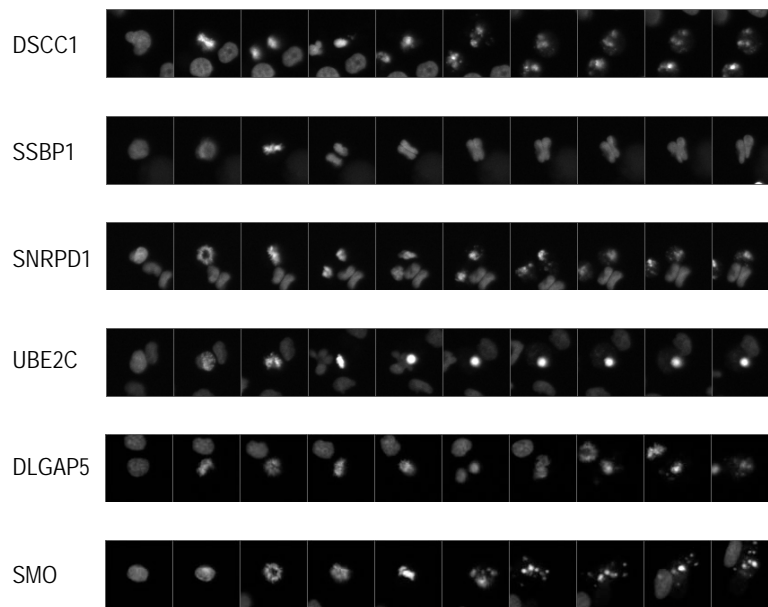


Figure 4.10: Selection of time-lapse images illustrating cell fate observed in the SHEP cell line for the six hits. The image sequence of knockdown of DSCC1 shows a cell in interphase, mitosis (metaphase), interphase (daughter nuclei), deformation of the nucleus (cell death), and cell death. The sequence of knockdown of SSBP1 shows a cell in interphase, mitosis (prometaphase), mitosis (metaphase), mitosis (anaphase), and finally daughter nuclei sticking together in arrest. The sequence of knockdown of SNRPD1 shows a cell in interphase, mitosis (prometaphase), mitosis (metaphase), daughter nuclei and cell death. The sequence of knockdown of UBE2C shows a cell in interphase, mitosis (prometaphase), mitosis (anaphase), and cell death. The sequence of knockdown of DLGAP5 shows a cell in interphase, mitosis (metaphase), daughter nuclei, deformation, and cell death. The sequence of SMO knockdown shows a cell in interphase, mitosis (prometaphase), mitosis (metaphase) and cell death.

Figure: 4.10 (page: 61) depicts a selection of typical time-lapses of cells with these gene knockdowns. In summary, four (DLGAP5, DSCC1, SSBP1, UBE2C) of these six proteins are directly involved in cell cycle and one indirectly (SMO) which is involved in cell cycle regulation. As such, the functional interpretation

of the six candidate genes provide indications that monitoring cell cycle dynamics enables identification of drug targets for neuroblastoma cells.

Gene expression analysis

Interestingly, all these six genes were highly up-regulated (p-value ≤ 0.01) in aggressive neuroblastoma tumors (stage 4, with MYCN amplification) in comparison to non-aggressive tumors (stage 1 without MYCN amplification) (Table: 4.8, page: 62). Furthermore, all six genes showed a good prediction performance for overall survival as shown in the Figure: 4.11 (page: 63).

Table 4.8: Gene expression analysis of the six hits.

Entrez ID	Gene symbol	Up-regulated MYCN	AMP	in tu- mors	P-value
79075	DSCC1	6.2E-18			3.7E-20
9787	DLGAPP5	1.1E-17			5.4E-24
11065	UBEC2	3.5E-18			4.1E-26
6742	SSBP1	1.1E-15			7.3E-18
6632	SNPRD1	2.3E-21			3.4E-30
6608	SMO	2.2E-15			7.3E-18

4.8 Data access via data repository iCHIP

All original data from this study is publicly available in the web-based database iCHIP. It can be accessed at <https://ichip.bioquant.uni-heidelberg.de> (User: guest; Password: sHeY82Nu). Each movie and each image can be observed and downloaded. Access to the images is achieved by selection of a gene in the query page. Associated gene and siRNA information is also available as well as the calculated phenotype scoring and related quality measures.

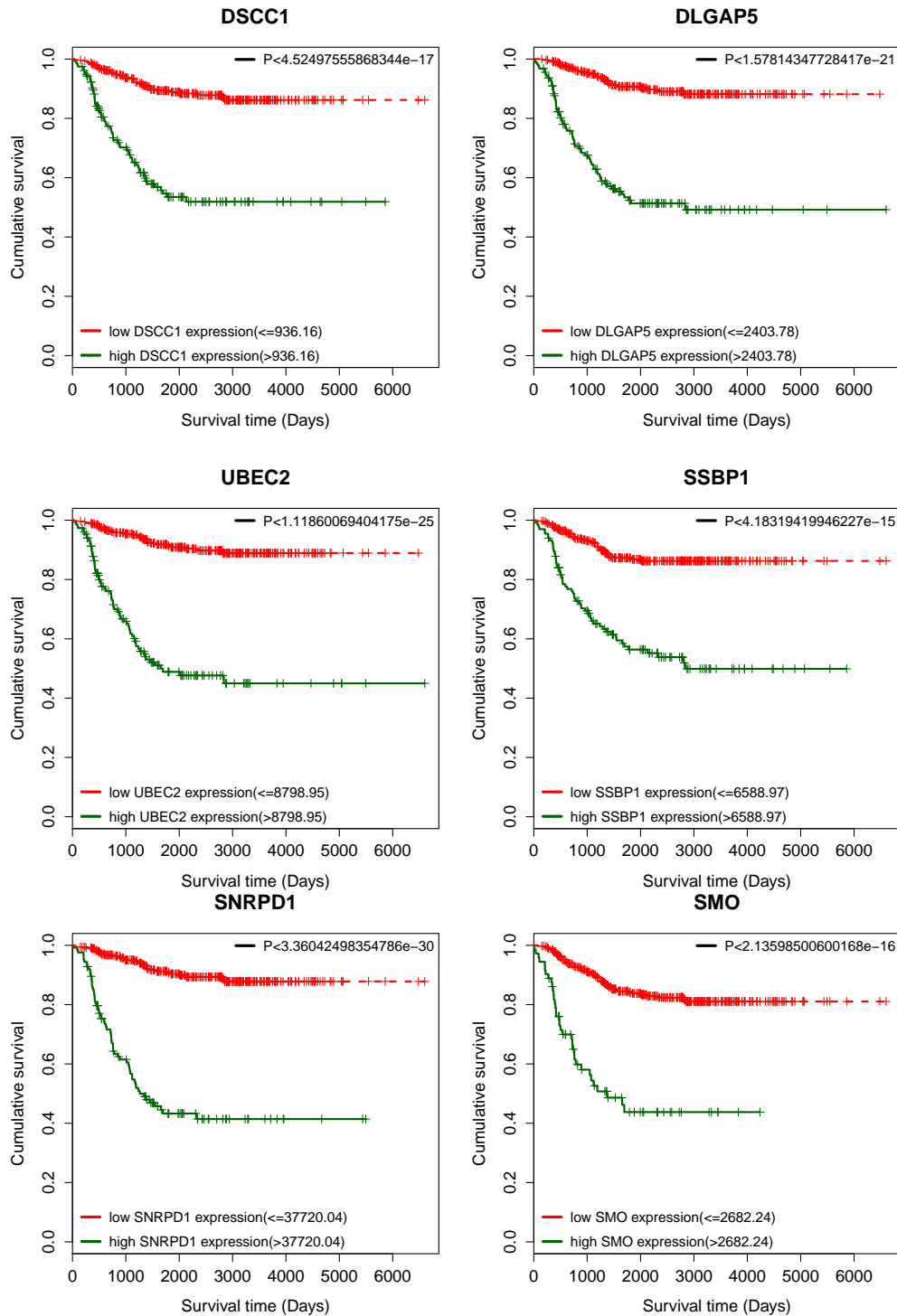


Figure 4.11: Kaplan Meier plots of the six hits. The log-rank p-values are shown on the top right of the plots.

CHAPTER 5

CONCLUSION

We have developed a processing pipeline for time-lapse microscopy screens from raw bitmaps to detailed perturbation analysis and identification of drug targets for tumors. The methodological contributions are threefold. First, integration of gene expression and gene knockdown analysis enables overcoming challenges posed by large genome-wide time-lapse studies. Second, optimization of the classification of cellular phenotypes improves prediction. Third, a novel analysis technique to track knockdown phenotype kinetics to monitor cellular dynamics.

Genome-wide siRNA screens are costly, need large data storage capacities, are very time consuming, and may still lead to ambiguous results ([Brass et al., 2008, Bushman et al., 2009, Konig et al., 2008, Zhou et al., 2008]). In contrast, kinome screens are less resource intensive and focus on a subset of genes (kinases) of the human genome [Duan et al., 2012, Cole et al., 2011]. In line with kinome screens, we focused our screen on a set of genes which are involved in cell cycle progression and tumorigenesis of neuroblastoma cells. We identified a list of candidates by analyzing large sets of publicly available gene expression data. From this list, we selected a set of 240 genes for knockdown studies, which have a potential role in neuroblastoma tumor progression.

With the selected set of 240 genes, we performed time-lapse image-based knock-down screens. These screens were performed in two neuroblastoma cell lines, i.e.

SH-EP and SK-N-BE(2)-C. The methods from screening, image processing and further analysis to identify knockdown phenotypes were optimized and developed specifically for these neuroblastoma cell lines. In a collaborative venture, existing image segmentation methods were optimized and new tracking method has been successfully developed [Harder et al., 2011].

To track mitotic aberrations after gene knockdown, we monitored well defined phenotypic classes (interphase, mitosis, cell death) of cell nuclei. In the proposed pipeline, classification of the cells into distinct phenotypes using image-based screens is crucial, as any follow-up interpretation is based upon this. We applied a filter which removed ambiguous samples from the data. We also normalized the features using various scaling schemes to gain better results. We did not solely rely on the cross-validation accuracy values to assess classification performance. The classified phenotypes were manually evaluated on a randomly selected independent test set. In addition, assignment of the mitosis and interphase classes was improved by an automated correction scheme employing tracking information. We found that filtering improved the classifiers performance, feature scaling was beneficial for SK-N-B-E(2)C cell line, and manual evaluation reveals the errors which could not be traced with cross validation and automatic corrections help improve the classifiers performance.

Subsequently, these phenotypes were analyzed in a time dependent manner. Tracking mitosis with a time-lapse of 40 min at a single cell level is challenging given the fact that the mitosis takes approximately 20 min in human cells. Our finding that neuroblastoma cells synchronize their cell cycle opens the possibility to monitor phenotype kinetics of the whole population. We tracked population response and observed the consequence of gene perturbation considering integration of overlapping time-frames. We identified six genes (DLGAP5, DSCC1, SSBP1, UBE2C, SNRPD1, and SMO) with an important role in prevention of aberrant cell cycle progression in both cell lines and hence six potential drug targets for silencing in cancer therapy. These genes were significantly up-regulated in aggressive neuroblastoma tumors and are good predictors for clinical outcome.

In addition, phenotype profiles of genes were clustered using unsupervised clustering to identify genes with same knockdown phenotype profile. We identified a sub cluster of 6 genes that show mitosis-linked-cell-death phenotype only in SK-

N-BE(2)-C cell lines. Of these six genes we validated the knockdown phenotype of MAD2L1, which is an M-A checkpoint gene. MAD2L1 knockdown in SK-N-BE(2)-C cells causes mitosis-linked-cell-death. We performed cytometric and cytogenetic analyses to confirm the presence of aneuploid cells upon MAD2L1 inhibition in presence of amplified MYCN and dysfunctional p53-p21 signaling. Aneuploid cells are supported by MAD2L1 in absence of p53 signaling [Schvartzman et al., 2011] and thus its inhibition may be a therapeutic option in neuroblastomas with overactive M-A checkpoint and dysfunctional p53-p21 signaling [Gogolin et al., 2012]. Similar analysis of other genes can contribute more to our understanding of the dynamics of neuroblastoma cells.

In summary, we developed a general method to characterize cell fate upon knock-down using high-throughput time-lapse imaging data, and applied the pipeline to neuroblastoma cells. The analysis identified six novel candidates which were not previously associated with cell cycle of neuroblastoma cells. In this study, we employed the neuroblastoma cell lines SH-EP and SK-N-BE(2)-C. As a future aspect, our findings need validations using a larger set of different neuroblastoma cell lines and cells from primary tumor material.

Appendices

APPENDIX A

APPENDIX: SCREENED GENES

Table A.1: Screened gene list

Entrez ID	Official Gene Symbol	Gene Name
132	ADK	adenosine kinase
191	AHCY	adenosylhomocysteinase
641	BLM	Bloom syndrome, RecQ helicase-like
661	POLR3D	polymerase (RNA) III (DNA directed) polypeptide D, 44kDa
699	BUB1	budding uninhibited by benzimidazoles 1 homolog (yeast)
701	BUB1B	budding uninhibited by benzimidazoles 1 homolog beta (yeast)
705	BYSL	bystin-like
790	CAD	carbamoyl-phosphate synthetase 2, aspar- tate transcarbamylase, and dihydroorotase
890	CCNA2	cyclin A2
891	CCNB1	cyclin B1
983	CDK1	cell division cycle 2, G1 to S and G2 to M
990	CDC6	cell division cycle 6 homolog (<i>S. cerevisiae</i>)

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
991	CDC20	cell division cycle 20 homolog (<i>S. cerevisiae</i>)
995	CDC25C	cell division cycle 25 homolog C (<i>S. pombe</i>)
1019	CDK4	cyclin-dependent kinase 4
1033	CDKN3	cyclin-dependent kinase inhibitor 3
1058	CENPA	centromere protein A
1062	CENPE	centromere protein E, 312kDa
1063	CENPF	centromere protein F, 350/400ka (mitosin)
1373	CPS1	carbamoyl-phosphate synthetase 1, mitochondrial
1499	CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa
1509	CTSD	cathepsin D
1537	CYC1	cytochrome c-1
1736	DKC1	dyskeratosis congenita 1, dyskerin
1802	DPH2	DPH2 homolog (<i>S. cerevisiae</i>)
1973	EIF4A1	similar to eukaryotic translation initiation factor 4A; small nucleolar RNA, H/ACA box 67; eukaryotic translation initiation factor 4A, isoform 1
2023	ENO1	enolase 1, (alpha)
2091	FBL	fibrillarin
2194	FASN	fatty acid synthase
2305	FOXM1	forkhead box M1
2521	FUS	fusion (involved in t(12;16) in malignant liposarcoma)
2746	GLUD1	glutamate dehydrogenase 1
2805	GOT1	glutamic-oxaloacetic transaminase 1, soluble (aspartate aminotransferase 1)
2806	GOT2	glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2)

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
2875	GPT	glutamic-pyruvate transaminase (alanine aminotransferase)
3068	HDGF	hepatoma-derived growth factor (high-mobility group protein 1-like)
3099	HK2	hexokinase 2 pseudogene; hexokinase 2
3148	HMGB2	high-mobility group box 2
3161	HMMR	hyaluronan-mediated motility receptor (RHAMM)
3326	HSP90AB1	heat shock protein 90kDa alpha (cytosolic), class B member 1
3329	HSPD1	heat shock 60kDa protein 1 (chaperonin) pseudogene 5; heat shock 60kDa protein 1 (chaperonin) pseudogene 6; heat shock 60kDa protein 1 (chaperonin) pseudogene 1; heat shock 60kDa protein 1 (chaperonin) pseudogene 4; heat shock 60kDa protein 1 (chaperonin)
3336	HSPE1	heat shock 10kDa protein 1 (chaperonin 10)
3619	INCENP	inner centromere protein antigens 135/155kDa
3735	KARS	lysyl-tRNA synthetase
3832	KIF11	kinesin family member 11
3833	KIFC1	kinesin family member C1
3835	KIF22	kinesin family member 22
3945	LDHB	lactate dehydrogenase B
3992	FADS1	fatty acid desaturase 1
4001	LMNB1	lamin B1
4061	LY6E	lymphocyte antigen 6 complex, locus E
4085	MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
4175	MCM6	minichromosome maintenance complex component 6
4193	MDM2	Mdm2 p53 binding protein homolog (mouse)
4234	METTL1	methyltransferase like 1
4284	MIP	major intrinsic protein of lens fiber
4524	MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)
4548	MTR	5-methyltetrahydrofolate-homocysteine methyltransferase
4602	MYB	v-myb myeloblastosis viral oncogene homolog (avian)
4609	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
4613	MYCN	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)
4691	NCL	nucleolin
4751	NEK2	NIMA (never in mitosis gene a)-related kinase 2
4830	NME1	non-metastatic cells 1, protein (NM23A)
4831	NME2	non-metastatic cells 2, protein (NM23B)
4860	PNP	nucleoside phosphorylase
4869	NPM1	nucleophosmin 1 (nucleolar phosphoprotein B23, numatrin) pseudogene 21; hypothetical LOC100131044; similar to nucleophosmin 1; nucleophosmin (nucleolar phosphoprotein B23, numatrin)
5036	PA2G4	proliferation-associated 2G4, 38kDa; proliferation-associated 2G4 pseudogene 4

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
5198	PFAS	phosphoribosylformylglycinamidine synthase
5245	PHB	prohibitin
5347	PLK1	polo-like kinase 1 (Drosophila)
5425	POLD2	polymerase (DNA directed), delta 2, regulatory subunit 50kDa
5427	POLE2	polymerase (DNA directed), epsilon 2 (p59 subunit)
5471	PPAT	phosphoribosyl pyrophosphate amidotransferase
5496	PPM1G	protein phosphatase 1G (formerly 2C), magnesium-dependent, gamma isoform
5591	PRKDC	similar to protein kinase, DNA-activated, catalytic polypeptide; protein kinase, DNA-activated, catalytic polypeptide
5616	PRKY	protein kinase, Y-linked
5757	PTMA	hypothetical LOC728026; prothymosin, alpha; hypothetical gene supported by BC013859; prothymosin, alpha pseudogene 4 (gene sequence 112)
5831	PYCR1	pyrroline-5-carboxylate reductase 1
5888	RAD51	RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)
5901	RAN	RAN, member RAS oncogene family
5984	RFC4	replication factor C (activator 1) 4, 37kDa
6241	RRM2	ribonucleotide reductase M2 polypeptide
6469	SHH	sonic hedgehog homolog (Drosophila)
6472	SHMT2	serine hydroxymethyltransferase 2 (mitochondrial)
6478	SIAH2	seven in absentia homolog 2 (Drosophila)

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
6510	SLC1A5	solute carrier family 1 (neutral amino acid transporter), member 5
6566	SLC16A1	solute carrier family 16, member 1 (mono-carboxylic acid transporter 1)
6599	SMARCC1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1
6608	SMO	smoothened homolog (Drosophila)
6626	SNRPA	small nuclear ribonucleoprotein polypeptide A
6628	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1
6632	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa; hypothetical protein LOC100129492
6635	SNRPE	small nuclear ribonucleoprotein polypeptide E-like 1; small nuclear ribonucleoprotein polypeptide E; similar to hCG23490
6636	SNRPF	small nuclear ribonucleoprotein polypeptide F
6723	SRM	spermidine synthase
6742	SSBP1	single-stranded DNA binding protein 1
6790	AURKA	aurora kinase A; aurora kinase A pseudogene 1
7083	TK1	thymidine kinase 1, soluble
7153	TOP2A	topoisomerase (DNA) II alpha 170kDa
7157	TP53	tumor protein p53
7272	TTK	TTK protein kinase
7298	TYMS	thymidylate synthetase
7371	UCK2	uridine-cytidine kinase 2

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
7866	IFRD2	interferon-related developmental regulator 2
8140	SLC7A5	solute carrier family 7 (cationic amino acid transporter, y+ system), member 5
8260	NAA10	ARD1 homolog A, N-acetyltransferase (<i>S. cerevisiae</i>)
8318	CDC45	CDC45 cell division cycle 45-like (<i>S. cerevisiae</i>)
8438	RAD54L	RAD54-like (<i>S. cerevisiae</i>)
8449	DHX16	DEAH (Asp-Glu-Ala-His) box polypeptide 16
8508	NIPSNAP1	nipsnap homolog 1 (<i>C. elegans</i>)
8568	RRP1	ribosomal RNA processing 1 homolog (<i>S. cerevisiae</i>)
8607	RUVBL1	RuvB-like 1 (<i>E. coli</i>)
8893	EIF2B5	eukaryotic translation initiation factor 2B, subunit 5 epsilon, 82kDa
9055	PRC1	protein regulator of cytokinesis 1
9133	CCNB2	cyclin B2
9136	RRP9	ribosomal RNA processing 9, small subunit (SSU) processome component, homolog (yeast)
9156	EXO1	exonuclease 1
9188	DDX21	DEAD (Asp-Glu-Ala-Asp) box polypeptide 21
9212	AURKB	aurora kinase B
9391	CIAO1	cytosolic iron-sulfur protein assembly 1 homolog (<i>S. cerevisiae</i>)
9493	KIF23	kinesin family member 23
9521	EEF1E1	eukaryotic translation elongation factor 1 epsilon 1

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
9625	AATK	apoptosis-associated tyrosine kinase
9718	ECE2	endothelin converting enzyme 2
9787	DLGAP5	discs, large (<i>Drosophila</i>) homolog-associated protein 5
9816	URB2	URB2 ribosome biogenesis 2 homolog (<i>S. cerevisiae</i>)
9833	MELK	maternal embryonic leucine zipper kinase
9928	KIF14	kinesin family member 14
10056	FARSB	phenylalanyl-tRNA synthetase, beta subunit
10105	PPIF	peptidylprolyl isomerase F
10131	TRAP1	TNF receptor-associated protein 1
10244	RABEPK	Rab9 effector protein with kelch motifs
10257	ABCC4	ATP-binding cassette, sub-family C (CFTR/MRP), member 4
10300	KATNB1	katanin p80 (WD repeat containing) subunit B 1
10403	NDC80	NDC80 homolog, kinetochore complex component (<i>S. cerevisiae</i>)
10419	PRMT5	protein arginine methyltransferase 5
10452	TOMM40	translocase of outer mitochondrial membrane 40 homolog (yeast)
10469	TIMM44	translocase of inner mitochondrial membrane 44 homolog (yeast)
10471	PFDN6	prefoldin subunit 6
10575	CCT4	chaperonin containing TCP1, subunit 4 (delta)
10721	POLQ	polymerase (DNA directed), theta

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
10797	MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
10807	SDCCAG3	serologically defined colon cancer antigen 3; similar to Serologically defined colon cancer antigen 3
10856	RUVBL2	RuvB-like 2 (E. coli)
10884	MRPS30	mitochondrial ribosomal protein S30
10885	WDR3	WD repeat domain 3
11004	KIF2C	kinesin family member 2C
11065	UBE2C	ubiquitin-conjugating enzyme E2C
11113	CIT	citron (rho-interacting, serine/threonine kinase 21)
11130	ZWINT	ZW10 interactor
11169	WDHD1	WD repeat and HMG-box DNA binding protein 1
11222	MRPL3	mitochondrial ribosomal protein L3
11339	OIP5	Opa interacting protein 5
22974	TPX2	TPX2, microtubule-associated, homolog (Xenopus laevis)
23016	EXOSC7	exosome component 7
23082	PPRC1	peroxisome proliferator-activated receptor gamma, coactivator-related 1
23107	MRPS27	mitochondrial ribosomal protein S27
23160	WDR43	WD repeat domain 43
23170	TTL12	tubulin tyrosine ligase-like family, member 12
23212	RRS1	RRS1 ribosome biogenesis regulator homolog (S. cerevisiae)
23397	NCAPH	non-SMC condensin I complex, subunit H

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
23404	EXOSC2	exosome component 2
24147	FJX1	four jointed box 1 (Drosophila)
25813	SAMM50	sorting and assembly machinery component 50 homolog (S. cerevisiae)
26094	DCAF4	WD repeat domain 21A
26156	RSL1D1	ribosomal L1 domain containing 1
26354	GNL3	guanine nucleotide binding protein-like 3 (nucleolar)
26519	TIMM10	translocase of inner mitochondrial membrane 10 homolog (yeast)
26577	PCOLCE2	procollagen C-endopeptidase enhancer 2
27131	SNX5	sorting nexin 5
27166	PRELID1	PRELI domain containing 1; similar to Px19-like protein (25 kDa protein of relevant evolutionary and lymphoid interest) (PRELI)
27346	TMEM97	transmembrane protein 97
27440	CECR5	cat eye syndrome chromosome region, candidate 5
29028	ATAD2	ATPase family, AAA domain containing 2
29127	RACGAP1	Rac GTPase activating protein 1 pseudogene; Rac GTPase activating protein 1
29923	HILPDA	chromosome 7 open reading frame 68
50814	NSDHL	NAD(P) dependent steroid dehydrogenase-like
51187	RSL24D1	ribosomal L24 domain containing 1; similar to ribosomal protein L24-like
51203	NUSAP1	nucleolar and spindle associated protein 1
51295	ECSIT	ECSIT homolog (Drosophila)
51373	MRPS17	mitochondrial ribosomal protein S17

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
51388	NIP7	nuclear import 7 homolog (<i>S. cerevisiae</i>)
51514	DTL	denticleless homolog (<i>Drosophila</i>)
51611	DPH5	DPH5 homolog (<i>S. cerevisiae</i>)
54205	CYCS	cytochrome c, somatic
54443	ANLN	anillin, actin binding protein
54478	FAM64A	family with sequence similarity 64, member A
54821	ERCC6L	excision repair cross-complementing rodent repair deficiency, complementation group 6-like
54865	GPATCH4	G patch domain containing 4
55038	CDCA4	cell division cycle associated 4
55143	CDCA8	cell division cycle associated 8
55165	CEP55	centrosomal protein 55kDa
55646	LYAR	Ly1 antibody reactive homolog (mouse)
55732	C1orf112	chromosome 1 open reading frame 112
55759	WDR12	WD repeat domain 12
55789	DEPDC1B	DEP domain containing 1B
55835	CENPJ	centromere protein J
55872	PBK	PDZ binding kinase
56342	PPAN	peter pan homolog (<i>Drosophila</i>)
56905	C15orf39	chromosome 15 open reading frame 39
56992	KIF15	kinesin family member 15
57405	SPC25	SPC25, NDC80 kinetochore complex component, homolog (<i>S. cerevisiae</i>)
64782	AEN	apoptosis enhancing nuclease
64943	NT5DC2	5'-nucleotidase domain containing 2
79075	DSCC1	defective in sister chromatid cohesion 1 homolog (<i>S. cerevisiae</i>)
79077	DCTPP1	dCTP pyrophosphatase 1

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
79084	WDR77	WD repeat domain 77
79682	MLF1IP	MLF1 interacting protein
79723	SUV39H2	suppressor of variegation 3-9 homolog 2 (Drosophila)
80097	MZT2B	family with sequence similarity 128, member B
80273	GRPEL1	GrpE-like 1, mitochondrial (E. coli)
80324	PUS1	pseudouridylate synthase 1
80746	TSEN2	tRNA splicing endonuclease 2 homolog (S. cerevisiae)
81610	FAM83D	family with sequence similarity 83, member D
81887	LAS1L	LAS1-like (S. cerevisiae)
83540	NUF2	NUF2, NDC80 kinetochore complex component, homolog (S. cerevisiae)
84057	MND1	meiotic nuclear divisions 1 homolog (S. cerevisiae)
84319	C3orf26	chromosome 3 open reading frame 26
84365	MKI67IP	MKI67 (FHA domain) interacting nucleolar phosphoprotein
84798	C19orf48	chromosome 19 open reading frame 48
84881	RPUSD4	RNA pseudouridylate synthase domain containing 4
87178	PNPT1	polyribonucleotide nucleotidyltransferase 1
90417	C15orf23	chromosome 15 open reading frame 23
91869	RFT1	RFT1 homolog (S. cerevisiae)
91942	NDUFAF2	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, assembly factor 2
113130	CDCA5	cell division cycle associated 5
115286	SLC25A26	solute carrier family 25, member 26

Continued on next page

Table A.1 – continued from previous page

Entrez ID	Official Gene Symbol	Gene Name
116028	RMI2	chromosome 16 open reading frame 75
118980	SFXN2	sideroflexin 2
133015	PACRGL	PARK2 co-regulated-like
151246	SGOL2	shugoshin-like 2 (S. pombe)
157570	ESCO2	establishment of cohesion 1 homolog 2 (S. cerevisiae)
201164	PLD6	phospholipase D family, member 6
442578	STAG3L3	aminoacyl tRNA synthetase complex-interacting multifunctional protein 2; stromal antigen 3-like 3
552900	BOLA2	bolA homolog 2 (E. coli); bolA homolog 2B (E. coli)
728833	FAM72D	family with sequence similarity 72, member D
729533	FAM72A	family with sequence similarity 72, member A
751071	METTL12	methyltransferase like 12
100008586	GAGE12F	G antigen 2A; G antigen 2B; G antigen 12I; G antigen 12F; G antigen 2E; G antigen 12G; G antigen 12D; G antigen 1; G antigen 2C; G antigen 12E; G antigen 2D; G antigen 12B; G antigen 3; G antigen 4; G antigen 12C; G antigen 5; G antigen 6; G antigen 7; G antigen 8

APPENDIX B

APPENDIX: PHENOTYPES IN SHEP CELL LINE

Table B.1: Phenotypes in SH-EP cell lines.

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis /Mitosis tion	Arrest Induc- tion
AATK	-	-	-	
ABCC4	-	-	0.00348	
ADK	-	0.01505	-	
AEN	0.01276	-	0.04341	
AHCY	-	-	-	
ANLN	-	0.01152	-	
ARD1A	-	-	0.00968	
AURKB	-	0.02552	-	
BLM	-	-	0.03472	
BOLA2	-	-	-	
BUB1	-	-	0.02797	
BUB1B	-	0.00219	-	
BYSL	0.03405	-	-	

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Induc- tion	Arrest Induc- tion
C15orf15	-	-	-	
C15orf23	-	0.00967	-	
C15orf39	-	-	-	
C16orf75	-	-	-	
C1orf112	-	-	0.00821	
C3orf26	-	0.02414	-	
CAD	-	0.03594	0.01115	
CCNA2	-	-	-	
CCNB2	0.00318	-	-	
CCT4	0.00237	-	0.01265	
CDC2	0.0372	-	-	
CDC20	0.00182	-	0.03566	
CDC25C	0.04038	0.04052	-	
CDC45L	-	0.04556	-	
CDC6	0.01458	0.00367	-	
CDCA4	-	0.01176	-	
CDCA5	-	0.00741	-	
CDCA8	-	0.00903	-	
CDK4	-	-	-	
CDK4	-	-	-	
CDK4	-	-	-	
CDKN3	-	-	-	
CECR5	0.04018	-	-	
CENPA	-	-	-	
CENPE	-	-	-	
CENPF	-	-	0.01265	
CEP55	-	0.01603	-	
CIAO1	-	-	-	

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
CPS1	-	-	-
CTNNB1	0.02785	-	-
CYC1	-	0.04107	-
CYCS	-	-	0.01788
DCTPP1	-	0.01004	-
DDX21	-	-	-
DEPDC1B	-	-	-
DHX16	-	-	-
DKC1	-	0.00638	-
DPH2	-	0.03329	-
DPH5	-	-	-
DTL	-	0.00142	-
ECE2	-	0.02695	-
ECSIT	-	-	-
EEF1E1	0.01852	-	-
EIF2B5	0.01655	-	-
EIF4A1	-	0.0493	-
ENO1	-	-	-
ERCC6L	0.00995	-	0.00802
ESCO2	0.02361	-	-
EXO1	-	-	-
EXOSC2	-	-	-
EXOSC7	-	-	-
FADS1	-	-	0.02892
FAM128B	-	-	-
FAM64A	-	-	0.01634
FAM72A	-	-	-
FAM83D	-	-	-

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
FARSB	-	-	0.03518
FASN	-	-	-
FBL	-	0.01977	-
FJX1	-	-	-
FOXM1	0.01056	0.00741	-
FUS	-	-	0.00426
GAGE12F	-	-	-
GCSH	-	-	-
GLUD1	-	-	0.03281
GOT2	-	-	0.03031
GPATCH4	-	-	-
GPT	-	-	-
GRPEL1	-	-	0.01203
HDGF	0.02566	-	-
HIG2	0.04018	-	0.01751
HK2	-	-	-
HMGB2	-	0.00119	-
HMMR	-	0.01603	-
HSP90AB1	0.01888	-	-
HSPD1	-	-	-
HSPE1	-	-	0.02713
IFRD2	-	0.00983	-
INCENP	-	0.00641	-
INCENP	-	0.00641	-
INCENP	-	0.00641	-
KATNB1	-	-	0.01347
KIF11	-	-	-
KIF11	-	-	-

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis tion	Arrest Induc- tion
KIF11	-	-	-	
KIF14	-	0.01329	-	
KIF15	-	-	-	
KIF22	-	-	-	
KIF23	0.04511	0.01896	-	
KIF2C	-	0.02266	-	
KIFC1	-	-	-	
LAS1L	0.04571	-	0.01654	
LDHB	-	0.02908	-	
LOC751071	-	-	0.01859	
LY6E	0.03435	0.04066	-	
LYAR	-	-	-	
MAD2L1	-	-	-	
MCM6	-	0.00092	-	
MDM2	-	-	0.03006	
MELK	-	-	-	
METTTL1	-	-	0.01689	
MIP	0.00886	-	-	
MLF1IP	0.01867	-	-	
MND1	0.00838	-	-	
MRPS17	-	-	0.03115	
MRPS27	0.03831	-	0.04208	
MRPS30	-	-	0.04476	
MTHFR	-	0.03644	-	
MTR	-	-	-	
MYB	0.01137	-	-	
MYC	-	0.00345	-	
MYCN	-	-	-	

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis tion	Arrest Induc- tion
NCAPH	-	-	-	
NDC80	-	0.00166	-	
NDUFAF2	-	-	-	
NIP7	-	-	0.03316	
NIPSNAP1	-	-	-	
NME1	-	-	-	
NME2	-	0.00591	-	
NP	-	0.02961	0.04198	
NPM1	-	-	-	
NSDHL	-	-	-	
NUF2	-	-	-	
OIP5	-	-	0.03751	
PA2G4	-	-	-	
PACRGL	-	-	-	
PBK	-	0.02745	-	
PCOLCE2	0.04851	0.03582	-	
PFAS	-	-	-	
PFDN6	-	-	-	
PHB	-	0.03411	-	
PLD6	-	-	0.02479	
PLK1	-	0.00397	-	
PLK1	-	0.00397	-	
PLK1	-	0.00397	-	
PNPT1	-	-	-	
POLD2	-	-	0.04408	
POLE2	-	-	0.03167	
POLQ	-	-	-	
POLR3D	-	-	0.03023	

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis tion	Arrest Induc- tion
PPAN	0.01511	-	-	
PPAT	0.03545	-	0.01394	
PPIF	-	-	-	
PPM1G	-	-	-	
PPRC1	-	-	0.01442	
PRC1	-	-	-	
PRELID1	-	-	-	
PRKDC	-	0.04648	-	
PRKY	0.02296	-	0.01599	
PRMT5	0.03695	-	-	
PTMA	-	-	0.00889	
PUS1	-	-	-	
PYCR1	-	-	0.04375	
RABEPK	-	0.01817	-	
RAD51	0.01258	-	-	
RAD54L	0.01446	0.02094	-	
RAN	-	-	-	
RFC4	-	-	-	
RPUSD4	-	-	0.01273	
RRM2	-	0.00895	-	
RRP1	-	-	0.0114	
RRP9	0.02146	-	0.01442	
RRS1	-	-	0.01837	
RSL1D1	-	-	-	
RUVBL1	0.01796	-	-	
RUVBL2	-	-	-	
SAMM50	-	-	-	
SDCCAG3	0.04291	-	0.02465	

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Induc- tion	Arrest Induc- tion
SFXN2	-	-	-	
SGOL2	-	-	-	
SHH	-	-	-	
SHMT2	0.0215	-	-	
SIAH2	-	-	0.04144	
SLC16A1	-	0.01661	-	
SLC1A5	-	-	0.01815	
SLC25A26	0.0211	0.02637	-	
SLC7A5	0.00689	-	0.029	
SNRPA	-	-	-	
SNRPB	-	0.00591	-	
SNRPE	-	0.04163	-	
SNRPF	-	0.04496	-	
SNX5	-	-	0.01892	
SPC25	-	0.03236	-	
STAG3L3	-	0.04452	-	
SUV39H2	-	-	-	
TIMM10	0.04819	-	0.03984	
TIMM44	0.03471	-	-	
TK1	-	0.00503	-	
TMEM97	0.03276	-	-	
TOP2A	-	-	0.02382	
TPX2	-	-	-	
TRAP1	-	-	-	
TSEN2	0.03323	-	-	
TTK	-	-	-	
TTLL12	-	0.01993	0.02185	
TYMS	-	0.02971	-	

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Induc- tion	Arrest Induc- tion
UCK2	-	-	0.0458	
URB2	-	0.02961	-	
WDHD1	-	0.01069	-	
WDR12	0.01566	0.00706	-	
WDR21A	-	-	-	
WDR3	-	-	-	
WDR43	0.02795	-	-	
WDR77	0.01212	0.01152	-	
ZWINT	0.03411	-	0.00805	
ATAD2	-	-	-	
AURKA	0.00942	-	0.00776	
C19orf48	0.00498	-	-	
CCNB1	0.00808	-	0.00669	
CENPJ	0.00196	-	-	
CIT	-	-	-	
CTSD	0.01741	-	0.02423	
DLGAP5	0.0118	-	0.01151	
DSCC1	0.0095	-	0.00649	
GNL3	0.01687	-	0.01533	
GOT1	0.01903	-	0.0091	
KARS	0.0152	-	0.00598	
LMNB1	-	-	0.00511	
MKI67IP	0.02831	-	-	
MRPL3	0.04248	-	0.01155	
MTHFD2	0.03545	-	-	
NCL	0.03857	-	-	
NEK2	0.00084	-	0.00258	
NT5DC2	0.00157	-	0.00841	

Continued on next page

Table B.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
NUSAP1	0.01171	-	0.00528
RACGAP1	0.00147	-	0.02382
RFT1	0.01171	-	0.03307
SMARCC1	0.01505	-	0.01122
SMO	0.00459	-	0.00726
SNRPD1	0.00658	-	0.01949
SRM	0.00618	-	0.00473
SSBP1	-	-	0.0026
TOMM40	0.03435	-	0.02837
TP53	0.02666	-	0.00546
UBE2C	0.00717	-	0.00349

APPENDIX C

APPENDIX: PHENOTYPES IN BE2C CELL LINE

Table C.1: Phenotypes in SK-N-BE(2)-C cell lines.

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
AATK	0.02207	0.00983	-
ABCC4	-	-	-
ADK	0.04093	-	-
AEN	-	0.04602	-
AHCY	0.02232	0.02815	-
ARD1A	-	-	-
ATAD2	-	-	-
AURKA	-	-	0.04334
AURKB	-	-	-
BLM	-	-	-
BOLA2	-	-	-
BUB1	-	-	0.014
BUB1B	0.01133	-	-

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
BYSL	-	-	0.02971
C15orf15	-	-	-
C15orf23	-	-	0.04378
C15orf39	-	0.04107	-
C16orf75	-	-	0.01268
C19orf48	-	0.01147	-
C1orf112	-	-	-
C3orf26	-	0.02735	-
CAD	-	-	0.04234
CCNA2	-	0.0186	-
CCNB1	-	-	0.04393
CCNB2	-	-	0.00385
CCT4	-	-	0.01642
CDC20	-	-	-
CDC25C	-	-	-
CDC45L	-	-	-
CDC6	-	0.00899	-
CDCA4	-	0.04556	-
CDCA5	-	0.01292	-
CDCA8	-	0.03746	-
CDK4	-	-	-
CDK4	-	-	-
CDK4	-	-	-
CDKN3	-	-	-
CECR5	-	0.01017	-
CENPA	0.00416	-	-
CENPE	-	0.04526	-
CENPF	-	-	0.01863

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis tion	Arrest Induc- tion
CENPJ	-	-	-	
CEP55	-	0.01867	-	
CIAO1	-	-	-	
CIT	-	-	-	
CPS1	-	-	-	
CTNNB1	0.03435	-	-	
CTSD	-	-	-	
DEPDC1B	-	-	-	
DHX16	-	0.0045	-	
DKC1	-	-	-	
DPH2	-	0.0045	-	
DTL	-	0.01166	-	
ECE2	-	-	-	
EIF2B5	-	-	0.04556	
EIF4A1	0.00565	-	-	
ESCO2	-	-	0.01364	
EXO1	-	-	-	
EXOSC2	0.01824	-	0.04087	
EXOSC7	-	0.00589	-	
FADS1	0.03545	-	-	
FAM128B	-	-	-	
FAM72A	-	-	0.04995	
FAM83D	-	0.01051	-	
FARSB	-	-	-	
FBL	-	-	-	
FJX1	-	-	-	
FOXM1	-	-	0.01321	
FUS	-	0.01463	-	

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
GAGE12F	-	-	-
GCSH	-	-	-
GLUD1	-	-	-
GNL3	-	-	-
GOT1	-	-	-
GPATCH4	0.01161	-	-
GPT	-	-	-
GRPEL1	-	-	-
HDGF	-	-	0.02552
HIG2	-	-	0.00431
HMGB2	-	-	-
HMMR	-	0.00209	-
HSP90AB1	0.02642	0.02352	-
HSPD1	-	-	-
HSPE1	0.0417	-	0.00152
IFRD2	-	-	-
KARS	0.02496	-	0.03991
KATNB1	-	-	-
KIF11	-	0.00031	-
KIF11	-	0.00031	-
KIF11	-	0.00031	-
KIF14	-	0.01114	0.04906
KIF15	-	-	-
KIF23	0.00395	-	-
KIF2C	-	0.00446	-
KIFC1	0.03208	-	0.01572
LAS1L	0.01087	-	-
LDHB	-	-	-

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
LMNB1	-	-	0.00911
LOC751071	-	-	-
LY6E	-	0.03943	-
LYAR	-	-	-
MCM6	-	-	-
MDM2	-	-	-
MELK	-	-	-
METTL1	-	-	-
MIP	-	-	-
MKI67IP	-	-	-
MLF1IP	-	0.02216	-
MRPL3	-	-	-
MRPS27	-	-	-
MRPS30	-	-	0.03576
MTHFD2	0.04291	-	-
MTHFR	-	-	-
MTR	-	-	-
MYB	0.011	-	0.04298
MYC	-	0.01903	-
MYCN	-	-	-
NCAPH	-	-	-
NCL	-	-	-
NDUFAF2	-	-	-
NEK2	-	-	-
NIP7	-	-	-
NIPSNAP1	-	-	-
NME1	-	-	0.02608
NP	-	0.0194	-

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis tion	Arrest Induc- tion
NPM1	-	-	0.04205	
NSDHL	-	-	-	
NT5DC2	-	-	-	
NUF2	-	-	-	
NUSAP1	-	-	0.00725	
PA2G4	-	-	-	
PACRGL	-	-	-	
PBK	-	0.007	-	
PCOLCE2	-	-	-	
PFAS	-	-	-	
PFDN6	-	-	-	
PHB	-	-	0.01381	
PLD6	-	-	-	
POLD2	-	0.02705	-	
POLE2	-	-	-	
POLQ	-	0.00266	-	
POLR3D	-	0.02343	-	
PPAN	-	0.00524	-	
PPAT	-	0.03124	0.00269	
PPIF	-	-	0.00843	
PPM1G	-	-	-	
PRC1	-	-	-	
PRELID1	-	-	-	
PRKDC	-	-	-	
PRKY	-	-	0.01704	
PRMT5	-	-	-	
PTMA	0.00886	-	-	
PUS1	-	-	-	

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis tion	Arrest Induc- tion
PYCR1	-	0.0308	-	
RABEPK	-	-	-	
RACGAP1	-	-	-	
RAD51	0.00869	-	0.02805	
RAD54L	0.03818	-	0.01929	
RFC4	-	0.02552	-	
RFT1	-	-	-	
RPUSD4	-	-	0.02066	
RRM2	0.03772	-	-	
RRP1	-	-	0.01412	
RRP9	-	-	0.01575	
RSL1D1	0.00762	-	-	
RUVBL1	-	-	0.01004	
RUVBL2	-	-	-	
SAMM50	-	-	-	
SDCCAG3	0.04686	0.03047	-	
SFXN2	-	-	-	
SGOL2	-	-	-	
SHH	-	-	-	
SHMT2	-	-	-	
SIAH2	0.01642	0.02971	-	
SLC16A1	0.01114	0.01323	0.03844	
SLC25A26	-	-	-	
SLC7A5	-	-	-	
SMARCC1	-	0.01446	-	
SNRPA	-	-	-	
SNRPB	-	-	-	
SNRPE	-	-	-	

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Induc- tion	Arrest Induc- tion
SNRPF	-	-	0.02166	
SNX5	0.00446	-	0.00224	
SPC25	-	-	-	
SRM	-	-	0.02249	
STAG3L3	-	0.00211	-	
SUV39H2	-	-	0.0427	
TIMM10	-	-	-	
TIMM44	-	-	-	
TK1	-	0.03271	-	
TMEM97	-	-	0.03271	
TOP2A	-	0.04163	-	
TP53	-	-	-	
TPX2	-	-	-	
TRAP1	-	0.00975	-	
TSEN2	-	-	-	
TTK	-	-	-	
TTLL12	-	-	-	
TYMS	-	-	-	
UCK2	-	-	0.04946	
URB2	-	0.02505	-	
WDHD1	-	0.01082	-	
WDR12	-	-	-	
WDR21A	0.00225	0.00731	-	
WDR3	-	-	-	
WDR43	-	-	-	
WDR77	-	-	-	
ZWINT	-	-	-	
ANLN	0.03346	-	0.00288	

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis tion	Arrest Induc- tion
CDC2	0.01395	-	0.00555	
CYC1	0.00147	0.0036	-	
CYCS	0.00738	-	-	
DCTPP1	0.00765	0.01888	0.01645	
DDX21	0.00649	-	0.00573	
DLGAP5	-	-	-	
DPH5	-	-	0.00376	
DSCC1	0.01012	-	0.00108	
ECSIT	0.00349	-	0.00305	
EEF1E1	-	-	-	
ENO1	0.01629	-	-	
ERCC6L	-	-	0.00969	
FAM64A	0.03923	-	0.00954	
FASN	-	-	-	
GOT2	0.03984	-	0.04803	
HK2	-	-	0.01867	
INCENP	0.02287	-	0.02361	
INCENP	0.02287	-	0.02361	
INCENP	0.02287	-	0.02361	
KIF22	0.02477	-	0.00336	
MAD2L1	-	-	0.00088	
MND1	0.04979	-	0.007	
MRPS17	0.02387	-	0.02352	
NDC80	0.00667	-	0.00289	
NME2	-	0.01499	-	
OIP5	0.02423	-	-	
PLK1	-	0.01161	-	
PLK1	-	0.01161	-	

Continued on next page

Table C.1 – continued from previous page

GeneName	Cytotoxic	Antiproliferative	Mitotic /Mitosis Arrest Induc- tion
PLK1	-	0.01161	-
PNPT1	-	0.03213	0.02496
PPRC1	0.02552	-	0.00565
RAN	0.01372	-	0.00772
RRS1	-	0.02326	-
SLC1A5	0.02477	-	0.00628
SMO	0.00892	-	0.00105
SNRPD1	0.00658	-	0.00649
SSBP1	0.02361	-	0.02094
TOMM40	0.00164	-	0.00275
UBE2C	-	-	0.02086

Aneuploid

A set of chromosomes which does not contain an exact multiple of haploid sets of chromosomes..

Central dogma

All living cells have a common basic phenomenon of transfer of information between these sequential-information-carrying biopolymers, referred to as Central Dogma. It states that information can be copied from DNA to DNA (DNA replication). It can be transferred from DNA to RNA(transcription) and RNA can be read to synthesize proteins (translation). In certain viruses an enzyme called reverse transcriptase allows the information transfer from RNA to DNA as well. The relationship between a sequence of DNA and the sequence of the corresponding protein is called the genetic code. The genetic code is read in triplet nucleotides called codons.[Lewin, 2004]. Each codon has a defined meaning, there are 64 known codons. ATG is a called start codon, as it marks the beginning of a gene. TAA, TAG or TGA are called termination codons, it marks the end of a gene..

Decision boundary

Decision boundary is the hypersurface that separates the underlying vector space into two sets..

DNA

Deoxyribonucleic acid (DNA) is a polymer of nucleotides. It is a double helix molecule held together by hydrogen bonds. The hydrogen bonds are formed between bases in the opposite strands such that following bonds are formed: A-T, G-C, A-U. Thus, A and T are called complementary bases, G and C are also complementary. DNA is very long, unbroken fiber along the chromosome, which would be 2 inches long if laid out straight. It is highly coiled such that 2cm long DNA fits in a cell $10\mu m$. The packing of DNA is such that 200 nucleotides are coiled around eight histone proteins. The histone proteins are positively charged and thus balance the negative charge on the nucleotides owing to the phosphate groups..

Feature space

In pattern recognition a feature space is an abstract space where each pattern sample is represented as a point in n-dimensional space..

Gene and Gene expression

The fundamental unit of information in living systems is the gene. Gene is a segment of Deoxyribonucleic acid (DNA), marked by a transcription start site in the beginning and termination site at the end. There are 20,000 genes in human DNA. Most of these genes are recipes to form functional product called protein, some also lead to the formation of RNAs. Genes guide a cells function and traits. The information in a gene is processed in two steps (a) a RNA molecule whose nucleotide sequence is complementary to the DNA sequence is assembled via RNA polymerase enzyme (b) a polypeptide is synthesized whose amino acid sequence is based on the nucleotide sequence in RNA..

Hyperplane

Hyperplane is a geometrical concept. It is a generalization of plane in different number of dimensions. It separates space into two half spaces..

Nucleotide

Each nucleotide is composed of nitrogenous base, a pentose and a phosphate. The ribose in DNA is deoxyzised therefore the name deoxyribonu-

clieic acid. There are five nitrogenous bases Adenine(A), Tyrosine(T), Cytosine(C), Guananie(G) and Uracil(U). RNA has A,U,C,G and DNA has A,T,C,G..

RNA

RNA differs from DNA in three ways (a) the pentose sugar which is deoxy in DNA (b) the nitrogenous base Thymine(T) is replaced by Uracil (U) in RNA (c) it is single stranded. It is not packed as DNA, since its a smaller molecule..

Support vectors

The data points that lie on hyperplane are called suport vectors. The solution changes if these points are removed..

Transcription

Transcription is the first step of the central dogma. It is initiated by binding of an enzyme RNA polymerase on a specific region called promoter, located at the beginning of a gene. DNA is a double helix molecule and it is unwind during the transcription process. The strand on which RNA polymerase binds is called template or anti-sense strand, and the other strand is called sense or coding strand. RNA polymerase reads each nucleotide on DNA template strand and adds a complementary nucleotide to nascent RNA molecule. At the end of the gene there is a stop signal which disengages RNA polymerase from the DNA strand. The RNA so produced is called messenger RNA (mRNA). mRNA carried the information from nucleus to cytoplasm and translates it to produce a protein[Raven, 2007]. .

Translation

Translation is the second step of the central dogma. It is initiated by the transport of mRNA molecule from nucleus to cytoplasm. In this process, mRNA strand bind ribosomes molecules and is read by the transfer RNA (tRNA). The ribosome then moves along the mRNA reading a codon (three nucleotides) at once. Ribosome induces tRNAs with anti-codon i.e. complementary sequence to that of the codon to bind the mRNA. tRNA carries a specific amino acid and thus the neighboring amino acids are chained to

form a polypeptide. This polypeptide chain is further processed in the cell to form a functional protein[Raven, 2007]. .

BIBLIOGRAPHY

- [Alexa et al., 2006] Alexa, A., Rahnenfuhrer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics (Oxford, England)*, 22(13):1600–7.
- [Anderson et al., 2009] Anderson, K., Lai, Z., McDonald, O. B., Stuart, J. D., Nartey, E. N., Hardwicke, M. A., Newlander, K., Dhanak, D., Adams, J., Patrick, D., Copeland, R. A., Tummino, P. J., and Yang, J. (2009). Biochemical characterization of GSK1070916, a potent and selective inhibitor of aurora b and aurora c kinases with an extremely long residence time. *Biochem J*, 420(2):25965.
- [Baguley and Marshall, 2004] Baguley, B. C. and Marshall, E. S. (2004). In vitro modelling of human tumour behaviour in drug discovery programmes. *Eur J Cancer*, 40(6):794–801.
- [Bakal et al., 2007] Bakal, C., Aach, J., Church, G., and Perrimon, N. (2007). Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science (New York, N.Y.)*, 316(5832):1753–1756. PMID: 17588932.
- [Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300.

- [Bernard et al., 2010] Bernard, S., Cajavec Bernard, B., Levi, F., and Herzl, H. (2010). Tumor growth rate determines the timing of optimal chronomodulated treatment schedules. *PLoS Comput Biol*, 6(3):e1000712.
- [Bettayeb et al., 2010] Bettayeb, K., Baunbaek, D., Delehouze, C., Loaec, N., Hole, A. J., Baumli, S., Endicott, J. A., Douc-Rasy, S., Benard, J., Oumata, N., Galons, H., and Meijer, L. (2010). CDK inhibitors roscovitine and CR8 trigger mcl-1 down-regulation and apoptotic cell death in neuroblastoma cells. *Genes Cancer*, 1(4):369–80.
- [Bharadwaj and Yu, 2004] Bharadwaj, R. and Yu, H. (2004). The spindle checkpoint, aneuploidy, and cancer. *Oncogene*, 23(11):2016–2027.
- [Birmingham et al., 2009] Birmingham, A., Selfors, L. M., Forster, T., Wrobel, D., Kennedy, C. J., Shanks, E., Santoyo-Lopez, J., Dunican, D. J., Long, A., Kelleher, D., Smith, Q., Beijersbergen, R. L., Ghazal, P., and Shamu, C. E. (2009). Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods*, 6(8):569–75.
- [Bolstad et al., 2003] Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–193. PMID: 12538238.
- [Boutros et al., 2006] Boutros, M., Bras, L. P., and Huber, W. (2006). Analysis of cell-based RNAi screens. *Genome Biol*, 7(7):R66.
- [Brass et al., 2008] Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N., Engelman, A., Xavier, R. J., Lieberman, J., and Elledge, S. J. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, 319(5865):921–6.
- [Brideau et al., 2003] Brideau, C., Gunter, B., Pikounis, B., and Liaw, A. (2003). Improved statistical methods for hit selection in high-throughput screening. *Journal of biomolecular screening*, 8(6):634–47.
- [Brodeur, 2003] Brodeur, G. M. (2003). Neuroblastoma: biological insights into a clinical enigma. *Nature reviews*, 3(3):203–16.

- [Brodeur et al., 1984] Brodeur, G. M., Seeger, R. C., Schwab, M., Varmus, H. E., and Bishop, J. M. (1984). Amplification of n-myc in untreated human neuroblastomas correlates with advanced disease stage. *Science (New York, N.Y.)*, 224(4653):1121–1124. PMID: 6719137.
- [Bushman et al., 2009] Bushman, F. D., Malani, N., Fernandes, J., D’Orso, I., Cagney, G., Diamond, T. L., Zhou, H., Hazuda, D. J., Espeseth, A. S., Konig, R., Bandyopadhyay, S., Ideker, T., Goff, S. P., Krogan, N. J., Frankel, A. D., Young, J. A., and Chanda, S. K. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog*, 5(5):e1000437.
- [Capasso et al., 2009] Capasso, M., Devoto, M., Hou, C., Asgharzadeh, S., Glessner, J. T., Attiyeh, E. F., Mosse, Y. P., Kim, C., Diskin, S. J., Cole, K. A., Bosse, K., Diamond, M., Laudenslager, M., Winter, C., Bradfield, J. P., Scott, R. H., Jagannathan, J., Garris, M., McConville, C., London, W. B., Seeger, R. C., Grant, S. F. A., Li, H., Rahman, N., Rappaport, E., Hakonarson, H., and Maris, J. M. (2009). Common variations in BARD1 influence susceptibility to high-risk neuroblastoma. *Nature Genetics*, 41(6):718–723. PMID: 19412175.
- [Carmena and Earnshaw, 2003] Carmena, M. and Earnshaw, W. C. (2003). The cellular geography of aurora kinases. *Nat Rev Mol Cell Biol*, 4(11):842–54.
- [Clark, 2003] Clark, T.G., B. M.-L. S. A. D. (2003). Survival analysis part i: Basic concepts and first analyses.
- [Cole et al., 2011] Cole, K. A., Huggins, J., Laquaglia, M., Hulderman, C. E., Russell, M. R., Bosse, K., Diskin, S. J., Attiyeh, E. F., Sennett, R., Norris, G., Laudenslager, M., Wood, A. C., Mayes, P. A., Jagannathan, J., Winter, C., Mosse, Y. P., and Maris, J. M. (2011). RNAi screen of the protein kinome identifies checkpoint kinase 1 (CHK1) as a therapeutic target in neuroblastoma. *Proc Natl Acad Sci U S A*, 108(8):3336–41.
- [Conrad et al., 2004] Conrad, C., Erfle, H., Warnat, P., Daigle, N., Lorch, T., Ellenberg, J., Pepperkok, R., and Eils, R. (2004). Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res*, 14(6):1130–6.
- [Cooper, 2000] Cooper, G. (2000). *The Cell: A molecular approach*. Sunderland(MA), Sinauer Associates.

- [Croft et al., 2011] Croft, D., O’Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D’Eustachio, P., and Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*, 39(Database issue):D691–7.
- [David L. Nelson, 2005] David L. Nelson, M. M. C. (2005). *Lehninger: Principles of biochemistry*. W.H. Freeman and company.
- [Deyell and Attiyeh, 2011] Deyell, R. J. and Attiyeh, E. F. (2011). Advances in the understanding of constitutional and somatic genomic alterations in neuroblastoma. *Cancer Genet*, 204(3):113–21.
- [Dickey et al., 2011] Dickey, A., Schleicher, S., Leahy, K., Hu, R., Hallahan, D., and Thotala, D. K. (2011). GSK-3beta inhibition promotes cell death, apoptosis, and in vivo tumor growth delay in neuroblastoma neuro-2A cell line. *J Neurooncol*, 104(1):145–53.
- [Diskin et al., 2009] Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., Cole, K., Moss, Y. P., Wood, A., Lynch, J. E., Pecor, K., Diamond, M., Winter, C., Wang, K., Kim, C., Geiger, E. A., McGrady, P. W., Blakemore, A. I. F., London, W. B., Shaikh, T. H., Bradfield, J., Grant, S. F. A., Li, H., Devoto, M., Rappaport, E. R., Hakonarson, H., and Maris, J. M. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, 459(7249):987–991. PMID: 19536264.
- [Doble and Woodgett, 2003] Doble, B. W. and Woodgett, J. R. (2003). GSK-3: tricks of the trade for a multi-tasking kinase. *J Cell Sci*, 116(Pt 7):1175–86.
- [Duan et al., 2012] Duan, Z., Zhang, J., Choy, E., Harmon, D., Liu, X., Nielsen, P., Mankin, H., Gray, N. S., and Hornicek, F. J. (2012). Systematic kinome shRNA screening identifies CDK11 (PITSLRE) kinase expression is critical for osteosarcoma cell growth and proliferation. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 18(17):4580–4588. PMID: 22791884.

- [Echeverri and Perrimon, 2006] Echeverri, C. J. and Perrimon, N. (2006). High-throughput RNAi screening in cultured cells: a user's guide. *Nature Reviews Genetics*, 7(5):373–384. PMID: 16607398.
- [Ehemann et al., 1999] Ehemann, V., Hashemi, B., Lange, A., and Otto, H. F. (1999). Flow cytometric DNA analysis and chromosomal aberrations in malignant glioblastomas. *Cancer letters*, 138(1-2):101–106. PMID: 10378780.
- [Ehemann et al., 2003] Ehemann, V., Sykora, J., Vera-Delgado, J., Lange, A., and Otto, H. F. (2003). Flow cytometric detection of spontaneous apoptosis in human breast cancer using the TUNEL-technique. *Cancer letters*, 194(1):125–131. PMID: 12706866.
- [Erfle et al., 2007] Erfle, H., Neumann, B., Liebel, U., Rogers, P., Held, M., Walter, T., Ellenberg, J., and Pepperkok, R. (2007). Reverse transfection on cell arrays for high content screening microscopy. *Nature protocols*, 2(2):392–9.
- [Fu, 2010] Fu, J. (2010). Collaboration of mitotic kinases in cell cycle control. *Nature Education*, 3(9):82.
- [Fuchs et al., 2010] Fuchs, F., Pau, G., Kranz, D., Sklyar, O., Budjan, C., Steinbrink, S., Horn, T., Pedal, A., Huber, W., and Boutros, M. (2010). Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Molecular Systems Biology*, 6:370. PMID: 20531400.
- [Gascoigne and Taylor, 2008] Gascoigne, K. E. and Taylor, S. S. (2008). Cancer cells display profound intra- and interline variation following prolonged exposure to antimetabolic drugs. *Cancer Cell*, 14(2):111–22.
- [Gaudray et al., 1992] Gaudray, P., Szepietowski, P., Escot, C., Birnbaum, D., and Theillet, C. (1992). DNA amplification at 11q13 in human cancer: from complexity to perplexity. *Mutation research*, 276(3):317–328. PMID: 1374524.
- [Ginisty et al., 1999] Ginisty, H., Sicard, H., Roger, B., and Bouvet, P. (1999). Structure and functions of nucleolin. *J Cell Sci*, 112 (Pt 6):761–72.
- [Gogolin et al., 2012] Gogolin, S., Batra, R., Harder, N., Ehemann, V., Paffhausen, T., Diessl, N., Sagulenko, V., Gade, S., Nolte, I., Rohr, K., Knig, R., and Westermann, F. (2012). MYCN-mediated overexpression of mitotic

- spindle regulatory genes and loss of p53-p21 function jointly support the survival of tetraploid neuroblastoma cells. *Cancer letters*. PMID: 23186832.
- [Gordon et al., 2012] Gordon, D. J., Resio, B., and Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. *Nature reviews. Genetics*, 13(3):189–203. PMID: 22269907.
- [Griffiths AJF, 2000] Griffiths AJF, Miller JH, S. D. e. a. (2000). *An Introduction to Genetic Analysis. 7th edition*. W. H. Freeman.
- [Gruenbaum et al., 2005] Gruenbaum, Y., Margalit, A., Goldman, R. D., Shumaker, D. K., and Wilson, K. L. (2005). The nuclear lamina comes of age. *Nature Reviews. Molecular Cell Biology*, 6(1):21–31. PMID: 15688064.
- [Harder et al., 2011] Harder, N., Batra, R., Gogolin, S., Diessl, N., Eils, R., Westermann, F., Knig, R., and Rohr, K. (2011). Large-scale tracking for cell migration and proliferation analysis and experimental optimization of high-throughput screens. *Proc. 6th Internat. Workshop on Microscopic Image Analysis with Applications in Biology (MIAAB '11)*.
- [Harder et al., 2008] Harder, N., Eils, R., and Rohr, K. (2008). Automated classification of mitotic phenotypes of human cells using fluorescent proteins. In *METHODS IN CELL BIOLOGY*. Elsevier.
- [Harder et al., 2006] Harder, N., Mora-Bermdez, F., Godinez, W. J., Ellenberg, J., Eils, R., and Rohr, K. (2006). Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 9(Pt 1):840–848. PMID: 17354969.
- [Harder et al., 2009] Harder, N., Mora-Bermudez, F., Godinez, W. J., Wunsche, A., Eils, R., Ellenberg, J., and Rohr, K. (2009). Automatic analysis of dividing cells in live cell movies to detect mitotic delays and correlate phenotypes in time. *Genome Research*, 19(11):2113–2124.
- [Harrison et al., 2009] Harrison, M. R., Holen, K. D., and Liu, G. (2009). Beyond taxanes: a review of novel agents that target mitotic tubulin and microtubules, kinases, and kinesins. *Clin Adv Hematol Oncol*, 7(1):54–64.

- [Henegariu et al., 2000] Henegariu, O., Bray-Ward, P., and Ward, D. C. (2000). Custom fluorescent-nucleotide synthesis as an alternative method for nucleic acid labeling. *Nature biotechnology*, 18(3):345–348. PMID: 10700155.
- [Hirotsu et al., 2010] Hirotsu, M., Setoguchi, T., Sasaki, H., Matsunoshita, Y., Gao, H., Nagao, H., Kunigou, O., and Komiya, S. (2010). Smoothed as a new therapeutic target for human osteosarcoma. *Mol Cancer*, 9:5.
- [Holzel et al., 2010] Holzel, M., Huang, S., Koster, J., Ora, I., Lakeman, A., Caron, H., Nijkamp, W., Xie, J., Callens, T., Asgharzadeh, S., Seeger, R. C., Messiaen, L., Versteeg, R., and Bernards, R. (2010). NF1 is a tumor suppressor in neuroblastoma that determines retinoic acid response and disease outcome. *Cell*, 142(2):218–29.
- [Hosack et al., 2003] Hosack, D. A., Dennis, Glynn, J., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome biology*, 4(10):R70. PMID: 14519205.
- [Hothorn and Lausen, 2003] Hothorn, T. and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics and amp; Data Analysis*, 43(2):121 – 137.
- [Huang da et al., 2009] Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57.
- [Hur and Weston, 2011] Hur, A. B. and Weston, J. (2011). A user’s guide to support vector classification.
- [Jiang et al., 2010] Jiang, S., Katayama, H., Wang, J., Li, S. A., Hong, Y., Radvanyi, L., Li, J. J., and Sen, S. (2010). Estrogen-induced aurora kinase-a (AURKA) gene expression is activated by GATA-3 in estrogen receptor-positive breast cancer cells. *Horm Cancer*, 1(1):1120.
- [Kanda et al., 1998] Kanda, T., Sullivan, K. F., and Wahl, G. M. (1998). Histone-GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells. *Current biology: CB*, 8(7):377–385. PMID: 9545195.

- [Kavallaris, 2010] Kavallaris, M. (2010). Microtubules and resistance to tubulin-binding agents. *Nat Rev Cancer*, 10(3):194–204.
- [Kishore, 2010] Kishore, J., G. M. K. P. (2010). Understanding survival analysis: Kaplan-meier estimate. *International Journal of Ayurveda Research*.
- [Kitzen et al., 2010] Kitzen, J. J., de Jonge, M. J., and Verweij, J. (2010). Aurora kinase inhibitors. *Crit Rev Oncol Hematol*, 73(2):99110.
- [Kohn, 1999] Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular biology of the cell*, 10(8):2703–2734. PMID: 10436023.
- [Konig et al., 2008] Konig, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M., Irelan, J. T., Chiang, C. Y., Tu, B. P., De Jesus, P. D., Lilley, C. E., Seidel, S., Opaluch, A. M., Caldwell, J. S., Weitzman, M. D., Kuhen, K. L., Bandyopadhyay, S., Ideker, T., Orth, A. P., Miraglia, L. J., Bushman, F. D., Young, J. A., and Chanda, S. K. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1):49–60.
- [Korur et al., 2009] Korur, S., Huber, R. M., Sivasankaran, B., Petrich, M., Morin, P., Hemmings, B. A., Merlo, A., and Lino, M. M. (2009). GSK3beta regulates differentiation and growth arrest in glioblastoma. *PLoS One*, 4(10):e7443.
- [Kotliarova et al., 2008] Kotliarova, S., Pastorino, S., Kovell, L. C., Kotliarov, Y., Song, H., Zhang, W., Bailey, R., Maric, D., Zenklusen, J. C., Lee, J., and Fine, H. A. (2008). Glycogen synthase kinase-3 inhibition induces glioma cell death through c-MYC, nuclear factor-kappaB, and glucose regulation. *Cancer Res*, 68(16):6643–51.
- [Lachmann and Ma’ayan, 2009] Lachmann, A. and Ma’ayan, A. (2009). KEA: kinase enrichment analysis. *Bioinformatics*, 25(5):684–6.
- [Le, 2003] Le, T. (2003). *Introductory statistics*. Wiley Interscience.
- [Lewin, 2004] Lewin, B. (2004). *Genes VIII*. Pearson Education International.

- [Li and Li, 2006] Li, J. J. and Li, S. A. (2006). Mitotic kinases: the key to duplication, segregation, and cytokinesis errors, chromosomal instability, and oncogenesis. *Pharmacol Ther*, 111(3):974–84.
- [Malumbres and Barbacid, 2007] Malumbres, M. and Barbacid, M. (2007). Cell cycle kinases in cancer. *Curr Opin Genet Dev*, 17(1):60–5.
- [Manchado et al., 2012] Manchado, E., Guillaumot, M., and Malumbres, M. (2012). Killing cells by targeting mitosis. *Cell Death Differ*, 19(3):369–77.
- [Maris, 2010] Maris, J. M. (2010). Recent advances in neuroblastoma. *The New England Journal of Medicine*, 362(23):2202–2211. PMID: 20558371.
- [Maris et al., 2008] Maris, J. M., Mosse, Y. P., Bradfield, J. P., Hou, C., Monni, S., Scott, R. H., Asgharzadeh, S., Attiyeh, E. F., Diskin, S. J., Laudenslager, M., Winter, C., Cole, K. A., Glessner, J. T., Kim, C., Frackelton, E. C., Casalunovo, T., Eckert, A. W., Capasso, M., Rappaport, E. F., McConville, C., London, W. B., Seeger, R. C., Rahman, N., Devoto, M., Grant, S. F. A., Li, H., and Hakonarson, H. (2008). Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *The New England Journal of Medicine*, 358(24):2585–2593. PMID: 18463370.
- [Markowetz, 2010] Markowetz, F. (2010). How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput Biol*, 6(2):e1000655.
- [Martin-Lluesma et al., 2002] Martin-Lluesma, S., Stucke, V. M., and Nigg, E. A. (2002). Role of *hec1* in spindle checkpoint signaling and kinetochore recruitment of Mad1/Mad2. *Science (New York, N. Y.)*, 297(5590):2267–2270. PMID: 12351790.
- [Meyer, 2012] Meyer, D, D. E. H. K. W. A. L. F. (2012). CRAN - package e1071. <http://cran.r-project.org/web/packages/e1071/index.html>.
- [Michel et al., 2004] Michel, L., Benezra, R., and Diaz-Rodriguez, E. (2004). MAD2 dependent mitotic checkpoint defects in tumorigenesis and tumor cell death: a double edged sword. *Cell cycle (Georgetown, Tex.)*, 3(8):990–992. PMID: 15254432.

- [Morimoto et al., 2007] Morimoto, H., Ozaki, A., Okamura, H., Yoshida, K., Amorim, B. R., Tanaka, H., Kitamura, S., and Haneji, T. (2007). Differential expression of protein phosphatase type 1 isotypes and nucleolin during cell cycle arrest. *Cell Biochem Funct*, 25(4):369–75.
- [Morozova et al., 2010] Morozova, O., Vojvodic, M., Grinshtein, N., Hansford, L. M., Blakely, K. M., Maslova, A., Hirst, M., Cezard, T., Morin, R. D., Moore, R., Smith, K. M., Miller, F., Taylor, P., Thiessen, N., Varhol, R., Zhao, Y., Jones, S., Moffat, J., Kislinger, T., Moran, M. F., Kaplan, D. R., and Marra, M. A. (2010). System-level analysis of neuroblastoma tumor-initiating cells implicates AURKB as a novel drug target for neuroblastoma. *Clin Cancer Res*, 16(18):4572–82.
- [Mosse et al., 2004] Mosse, Y. P., Laudenslager, M., Khazi, D., Carlisle, A. J., Winter, C. L., Rappaport, E., and Maris, J. M. (2004). Germline PHOX2B mutation in hereditary neuroblastoma. *Am J Hum Genet*, 75(4):727–30.
- [Mosse et al., 2008] Mosse, Y. P., Laudenslager, M., Longo, L., Cole, K. A., Wood, A., Attiyeh, E. F., Laquaglia, M. J., Sennett, R., Lynch, J. E., Perri, P., Laureys, G., Speleman, F., Kim, C., Hou, C., Hakonarson, H., Torkamani, A., Schork, N. J., Brodeur, G. M., Tonini, G. P., Rappaport, E., Devoto, M., and Maris, J. M. (2008). Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature*, 455(7215):930–5.
- [Neumann et al., 2006] Neumann, B., Held, M., Liebel, U., Erfle, H., Rogers, P., Pepperkok, R., and Ellenberg, J. (2006). High-throughput RNAi screening by time-lapse imaging of live human cells. *Nature methods*, 3(5):385–90.
- [Nilsen, 2003] Nilsen, T. W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, 25(12):1147–9.
- [NIST, 2003] NIST (2003). e-handbook of statistical methods.
- [Obaya et al., 1999] Obaya, A. J., Mateyak, M. K., and Sedivy, J. M. (1999). Mysterious liaisons: the relationship between c-myc and the cell cycle. *Oncogene*, 18(19):2934–41.
- [Oberthuer et al., 2006] Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Knig, R., Haas, S., Eils, R., Schwab, M.,

- Brors, B., Westermann, F., and Fischer, M. (2006). Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 24(31):5070–5078. PMID: 17075126.
- [Oberthuer et al., 2010] Oberthuer, A., Juraeva, D., Li, L., Kahlert, Y., Westermann, F., Eils, R., Berthold, F., Shi, L., Wolfinger, R. D., Fischer, M., and Brors, B. (2010). Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients. *The pharmacogenomics journal*, 10(4):258–266. PMID: 20676065.
- [Pasca di Magliano and Hebrok, 2003] Pasca di Magliano, M. and Hebrok, M. (2003). Hedgehog signalling in cancer formation and maintenance. *Nat Rev Cancer*, 3(12):903–11.
- [Raven, 2007] Raven, P. (2007). *Biology by Peter Raven (NASTA Hardcover Reinforced High School Binding) Student Edition*. McGraw-Hill Companies, Incorporated.
- [Richard et al., 2008] Richard, D. J., Bolderson, E., Cubeddu, L., Wadsworth, R. I., Savage, K., Sharma, G. G., Nicolette, M. L., Tsvetanov, S., McIlwraith, M. J., Pandita, R. K., Takeda, S., Hay, R. T., Gautier, J., West, S. C., Paull, T. T., Pandita, T. K., White, M. F., and Khanna, K. K. (2008). Single-stranded DNA-binding protein hSSB1 is critical for genomic stability. *Nature*, 453(7195):677–81.
- [Savelyeva et al., 2006] Savelyeva, L., Sagulenko, E., Schmitt, J. G., and Schwab, M. (2006). The neurobeachin gene spans the common fragile site FRA13A. *Human genetics*, 118(5):551–558. PMID: 16244873.
- [Schvartzman et al., 2011] Schvartzman, J.-M., Duijf, P. H. G., Sotillo, R., Coker, C., and Benezra, R. (2011). Mad2 is a critical mediator of the chromosome instability observed upon rb and p53 pathway inhibition. *Cancer cell*, 19(6):701–714. PMID: 21665145.

- [Shah and Cleveland, 2000] Shah, J. V. and Cleveland, D. W. (2000). Waiting for anaphase: Mad2 and the spindle assembly checkpoint. *Cell*, 103(7):997–1000. PMID: 11163175.
- [Shapiro, 2006] Shapiro, G. I. (2006). Cyclin-dependent kinase pathways as targets for cancer treatment. *J Clin Oncol*, 24(11):1770–83.
- [Sif et al., 1998] Sif, S., Stukenberg, P. T., Kirschner, M. W., and Kingston, R. E. (1998). Mitotic inactivation of a human SWI/SNF chromatin remodeling complex. *Genes Dev*, 12(18):2842–51.
- [Smyth, 2004] Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3:Article3. PMID: 16646809.
- [Srivastava and Pollard, 1999] Srivastava, M. and Pollard, H. B. (1999). Molecular dissection of nucleolin’s role in growth and cell proliferation: new insights. *FASEB J*, 13(14):1911–22.
- [Stark, 1997] Stark, P. B. (1997). Statistics tools for internet and classroom instruction with a graphical user interface.
- [Suzuki and Shimodaira, 2006] Suzuki, R. and Shimodaira, H. (2006). Pvclost: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics (Oxford, England)*, 22(12):1540–1542. PMID: 16595560.
- [T.-F. Wu, 2004] T.-F. Wu, C.-J. L. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, pages 5:975–1005.
- [Takagi et al., 2001] Takagi, M., Sueishi, M., Saiwaki, T., Kametaka, A., and Yoneda, Y. (2001). A novel nucleolar protein, NIFK, interacts with the forkhead associated domain of ki-67 antigen in mitosis. *J Biol Chem*, 276(27):25386–91.
- [Terret et al., 2009] Terret, M. E., Sherwood, R., Rahman, S., Qin, J., and Jallepalli, P. V. (2009). Cohesin acetylation speeds the replication fork. *Nature*, 462(7270):231–4.

- [Thacker, 2005] Thacker, J. (2005). The RAD51 gene family, genetic instability and cancer. *Cancer letters*, 219(2):125–135. PMID: 15723711.
- [Therneau, 2012] Therneau, T. (2012). CRAN - package survival. <http://cran.r-project.org/web/packages/survival/index.html>.
- [Thomaz and Gillies, 2011] Thomaz, C. and Gillies, D. (2011). Intelligent data analysis and probabilistic inference.
- [Tsai et al., 2006] Tsai, M. Y., Wang, S., Heidinger, J. M., Shumaker, D. K., Adam, S. A., Goldman, R. D., and Zheng, Y. (2006). A mitotic lamin b matrix induced by RanGTP required for spindle assembly. *Science*, 311(5769):1887–93.
- [Tuszynski, 2012] Tuszynski, J. (2012). CRAN - package caTools. <http://cran.r-project.org/web/packages/caTools/index.html>.
- [V. Kovalev and Rohr, 2006] V. Kovalev, N. H., B. N. M. H. U. L. H. E. J. E. R. E. and Rohr, K. (2006). Feature selection for evaluating fluorescence microscopy images in genome-wide cell screens. *Proceedings of IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*.
- [Vakifahmetoglu et al., 2008] Vakifahmetoglu, H., Olsson, M., and Zhivotovsky, B. (2008). Death through a tragedy: mitotic catastrophe. *Cell death and differentiation*, 15(7):1153–62.
- [van de Wetering et al., 2003] van de Wetering, M., Oving, I., Muncan, V., Pon Fong, M. T., Brantjes, H., van Leenen, D., Holstege, F. C. P., Brummelkamp, T. R., Agami, R., and Clevers, H. (2003). Specific inhibition of gene expression using a stably integrated, inducible small-interfering-RNA vector. *EMBO reports*, 4(6):609–615. PMID: 12776180.
- [Wakefield et al., 2003] Wakefield, J. G., Stephens, D. J., and Tavare, J. M. (2003). A role for glycogen synthase kinase-3 in mitotic spindle dynamics and chromosome alignment. *J Cell Sci*, 116(Pt 4):637–46.
- [Walter et al., 2010] Walter, T., Held, M., Neumann, B., Heriche, J. K., Conrad, C., Pepperkok, R., and Ellenberg, J. (2010). Automatic identification and

- clustering of chromosome phenotypes in a genome wide RNAi screen by time-lapse imaging. *J Struct Biol*, 170(1):19.
- [Wang et al., 2008] Wang, Z., Smith, K. S., Murphy, M., Piloto, O., Somervaille, T. C., and Cleary, M. L. (2008). Glycogen synthase kinase 3 in MLL leukaemia maintenance and targeted therapy. *Nature*, 455(7217):1205–9.
- [wei Hsu et al., 2010] wei Hsu, C., chung Chang, C., and jen Lin, C. (2010). A practical guide to support vector classification.
- [Westermann et al., 2008] Westermann, F., Muth, D., Benner, A., Bauer, T., Henrich, K. O., Oberthuer, A., Brors, B., Beissbarth, T., Vandesompele, J., Pattyn, F., Hero, B., Konig, R., Fischer, M., and Schwab, M. (2008). Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol*, 9(10):R150.
- [Willingham et al., 2004] Willingham, A. T., Deveraux, Q. L., Hampton, G. M., and Aza-Blanc, P. (2004). RNAi and HTS: exploring cancer by systematic loss-of-function. *Oncogene*, 23(51):8392–400.
- [Wong and Fang, 2006] Wong, J. and Fang, G. (2006). HURP controls spindle dynamics to promote proper interkinetochore tension and efficient kinetochore capture. *J Cell Biol*, 173(6):879–91.
- [Wong et al., 2008] Wong, J., Lerrigo, R., Jang, C. Y., and Fang, G. (2008). Aurora a regulates the activity of HURP by controlling the accessibility of its microtubule-binding domain. *Mol Biol Cell*, 19(5):2083–91.
- [Wong et al., 2009] Wong, T. S., Rajagopalan, S., Townsley, F. M., Freund, S. M., Petrovich, M., Loakes, D., and Fersht, A. R. (2009). Physical and functional interactions between human mitochondrial single-stranded DNA-binding protein and tumour suppressor p53. *Nucleic Acids Res*, 37(2):568–81.
- [Yamanaka et al., 2000] Yamanaka, A., Hatakeyama, S., Kominami, K., Kitagawa, M., Matsumoto, M., and Nakayama, K. (2000). Cell cycle-dependent expression of mammalian e2-c regulated by the anaphase-promoting complex/cyclosome. *Mol Biol Cell*, 11(8):2821–31.

- [Yauch et al., 2009] Yauch, R. L., Dijkgraaf, G. J., Alicke, B., Januario, T., Ahn, C. P., Holcomb, T., Pujara, K., Stinson, J., Callahan, C. A., Tang, T., Bazan, J. F., Kan, Z., Seshagiri, S., Hann, C. L., Gould, S. E., Low, J. A., Rudin, C. M., and de Sauvage, F. J. (2009). Smoothened mutation confers resistance to a hedgehog pathway inhibitor in medulloblastoma. *Science*, 326(5952):572–4.
- [Zhang et al., 2011] Zhang, J. D., Koerner, C., Bechtel, S., Bender, C., Keklikoglou, I., Schmidt, C., Irsigler, A., Ernst, U., Sahin, O., Wiemann, S., and Tschulena, U. (2011). Time-resolved human kinome RNAi screen identifies a network regulating mitotic-events as early regulators of cell proliferation. *PloS One*, 6(7):e22176. PMID: 21765947.
- [Zhao and Fang, 2005] Zhao, W.-M. and Fang, G. (2005). Anillin is a substrate of anaphase-promoting complex/cyclosome (APC/C) that controls spatial contractility of myosin during late cytokinesis. *The Journal of biological chemistry*, 280(39):33516–33524. PMID: 16040610.
- [Zhou et al., 2008] Zhou, Y. B., Cao, J. B., Wan, B. B., Wang, X. R., Ding, G. H., Zhu, H., Yang, H. M., Wang, K. S., Zhang, X., and Han, Z. G. (2008). hBolA, novel non-classical secreted proteins, belonging to different BolA family with functional divergence. *Mol Cell Biochem*, 317(1-2):618.

ACKNOWLEDGMENTS

There is no such thing as a self-made man. We are made up of thousands of others. Everyone who has ever done a kind deed for us, or spoken one word of encouragement to us, has entered into the makeup of our character and our thoughts, as well as our success. **George Adams.**

My thesis is no exception, many people helped me reach this far. I would like to thank..

Prof. Dr. Roland Eils, for creating an encouraging and scientifically stimulating environment here in Eils's Lab.

PD. Dr. Rainer König, for giving me a challenging research project and for all the scientific discussions. For being so supportive and having a friendly and warm *Network modeling* group.

PD. Dr. Karl Rohr, Dr. Nathalie Harder and Dr. Petr Matula from the *Biomedical computer vision* group for the fruitful collaboration. Warm and especial thanks to Nathalie, it was learning experience to work with her.

PD. Dr. Frank Westermann and Dr. Sina Gogolin from the *Tumor genetics* group at DKFZ for timely help and fruitful collaboration.

Heidelberg Graduate School for Mathematical and Computational sciences (HGS MATH COMP) for awarding me the PhD stipend.

PD. Dr. Stefan Wiemann, Prof. Ursula Kummer, and Dr. Vytaute Starkuviene-Erfele for their contribution to the completion of this work.

Previous and current members of *Network modeling* group Dr. Anna-Lena Kranz, Dr. Zita Soons, Dr. Marcus Oswald, Mr. Ashwini Sharma, Dr. Kitiporn Plaimas, and Mr. Volker Ast, for the friendly and helpful atmosphere.

Dr. Rosario Piro for all the helpful scientific discussions.

Dr. Apichat Surataneer for helping me in the confusing era of my PhD.

Dr. Gunnar Schramm for the joy he spreads.

Dr. Tobias Bauer for his humor and all the work hours he saved me by helping me with R. And for bringing Cathy, Florian and Christian to the office. They always used to lit up my mood.

Mr. Moritz Aschoff for sharing the Grad life ups and downs.

Ms. Anna Katharina Dieckmann, Mr. Lorenz Maier and their son Little David, for being so helpful, kind and warm.

Mr. Karl-Heinz Groß for always being so kind and helpful. Mr. Ralf Kabbe for the excellent support and timely help.

Ms. Corinna Sprengart and Ms. Manuela Schäfer for being so understanding, cheerful and helpful. Both of them took care of endless administrative affairs with ease and a smile on their face. It made me ask them once if they were trained for being that way. Hats off to their efforts.

Dr. Jan Eufinger and Ms. Ulrike Conrad for the wonderful management of many many events that I attended during these years.

Words fall short when it comes to thanking my loved ones.

My friends for being such lovely friends as they are, Sandeep and Maria for their love, support, assurances and encouragement. Numerous dishes we cooked together to endless discussions where we killed each other. Life in Heidelberg would not have been this memorable without them.

Corinna for being the source of positive energy. Talking to her recharge me every time.

Sownya for sharing the other side of the Grad life and motivating me to do better.

My parents, Sushil & Sarita, and my siblings, Satyam & Ruchi, who immensely loved me, always believed in me, motivated me and inspired me to be a better self. I owe it all to them.

aneuploidy, 4

cell clusters, 26, 28, 42

cell cycle, 9

cell tracking, 26, 30, 47, 66

classification, 7, 15, 17, 27, 42, 48, 66

classifier, 4, 6, 15, 17, 28

clustering, 5

cytokinesis, 10

enrichment, 40, 57

fishers test, 20, 24, 36

gene expression clusters, 24, 40

hyperplane, vc dimension, 18

interphase, 10

kaplan meier plots, 14, 63

kinase, 23, 36, 57, 65

kinase enrichment, 36, 57

log-rank test, 14, 15, 63

mitosis, 10

mitotic slippage, 6, 55

MYCN, 3, 24, 40, 59, 67

normal distribution, 19

phenocluster, 33, 52

phenoclusters, 67

phenotype tracking, 34, 55, 59, 66

RNA interference, 12

siRNA, 5, 33

survival analysis, 14

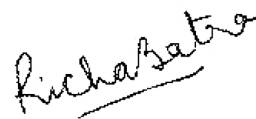
SVM, 15, 29

wilcoxon test, 19, 33, 34

Erklärung:

Ich versichere, daß ich diese Arbeit selbständig verfaßt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den March 4, 2013

A handwritten signature in black ink that reads "Richa Batra". The signature is written in a cursive style and is underlined.

.....
(Richa Batra)

All I know is that I know nothing

Socrates