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

Dipl.-Phys. Felix Alexander Klein

born in Friedberg (Hessen)

Oral-examination:

21.11.2014

Interaction and differential analysis of genomic high-throughput technologies

Referees:

Dr. Paul Bertone

and

Prof. Dr. Caroline Friedel

To Carolina

Contents

Abstract

High-throughput technologies are powerful tools for studying fundamental biological processes and for biomedical research. Analysis tools are needed to extract the biological information contained in the generated data sets.

In this dissertation I describe methods that I developed for the analysis of data from two high-throughput technologies. The first technology, Circularized Chromosome Conformation Capture (4C), allows to study the 3D chromatin interactions of a certain genomic region (viewpoint) with the rest of the genome. With the second technique, high-throughput microscopy screening, large scale phenotypic screens can be performed to investigate the influences of perturbations, such as compound treatment, on cell phenotypes.

Chapter 2 comprises the analysis of three studies which used 4C to study the influence of chromatin 3D structure on gene regulation. The 4C signal shows a strong dependence on the genomic distance from the viewpoint. To address the computational tasks of finding specific interactions, which are superimposed on the regular signal, and to detect changes between interaction profiles of different conditions I developed the R package `FourCSeq`. The package has been submitted to www.bioconductor.org.

Based on my analysis of the interaction profiles from 103 viewpoints, we could show that long-range chromatin interactions were widespread throughout the compact *Drosophila* genome. Furthermore, the comparison of the interaction profiles from different developmental time points and tissue types revealed that the chromatin configuration was mostly stable across time points and tissue types.

In two further 4C sequencing projects, I analyzed the influence of large genomic rearrangements on the chromatin structure of two genomic loci in mice. The first project focused on the locus of the *Shh*. My analysis showed that upon genomic inversion, the chromatin structure of the locus collapsed and contacts were redistributed. The second project studied the chromatin structure of the *Ap2-γ* and *Bmp7* locus. By analyzing the 4C profiles of 4 viewpoints spaced throughout this locus and determining the primary interaction domains for each viewpoint, I showed that the locus is partitioned by a small transition zone into two distinct domains. Analysis of the interaction profiles from alleles carrying large chromosomal rearrangements further supported this view that the transition zone played an important role in partitioning the locus. Together, both studies show that the chromatin structure is important for long-range gene regulation and allocation of enhancers to their target genes.

In Chapter 3, I describe the analysis of a high-throughput microscopy compound screen in a panel of isogenic human colorectal cancer cell lines. In this Pharmacogenetic Phenome Compendium (PGPC) project, we investigated chemical-genetic interactions between compounds and the genetic backgrounds of the isogenic cell line panel. Using high-throughput microscopy, we screened 1280 bioactive compounds in 12 isogenic colon cancer cell lines with specific mutations in signaling pathways. After image segmentation and feature extraction, I used a fea-

ture selection algorithm to select a non-redundant set of 20 phenotypic features for further analysis. My analysis of the phenotypic chemical-genetic interaction data allowed us to predict synergistic drug-combinations, uncover connections between signaling pathways, and cluster functionally related compounds and biological processes. Furthermore, I showed that the combined approach of high-throughput microscopy and chemical-genetic screening is more sensitive for interactome mapping than either alone. For easy access to the data and reproducibility of the results, I generated the `PGPC` **R** package that will be submitted to www.bioconductor.org.

Zusammenfassung

High-throughput Technologien werden eingesetzt, um grundlegende biologische Prozesse zu studieren und die biomedizinische Forschung voranzutreiben. Damit die biologischen Informationen, die in den Datensätzen enthalten sind, extrahiert und verwendet werden können, sind Methoden zur Analyse der Daten nötig.

In dieser Dissertation beschreibe ich die Methoden zur Datenanalyse von zwei High-throughput-Technologien, die ich entwickelt habe. Die erste Technologie, Circularized Chromosome Conformation Capture (4C), ermöglicht es, die 3D Interaktionen von einer bestimmten genomischen Region (Viewpoint) mit dem restlichen Genom zu erforschen. Bei der zweiten Technologie handelt es sich um High-througput microscopy screening. Umfangreiche phänotypische Screenings wie dieses werden durchgeführt, um die Einflüsse von äußeren Störungen, wie zum Beispiel die Behandlung mit chemischen Substanzen, auf den Phänotyp der Zelle zu studieren.

Kapitel 3 enthält die Analyse von drei Projekten, bei denen die 4C Technologie eingesetzt wurde, um den Einfluss der 3D Chromatin-Struktur auf die Genregulation zu untersuchen. Das 4C Signal ist stark von der genomischen Distanz zum Viewpoint abhängig. Um spezifische Interaktionen zu detektieren, die dem normalen Signal überlagert sind, und um Veränderungen zwischen den Interaktionsprofilen verschiedener Bedingungen zu detektieren, habe ich das R-Paket `FourCSeq` entwickelt. Das Paket wurde bei www.bioconductor.org eingereicht.

Basierend auf meiner Analyse der Interaktionsprofile von 103 Viewpoints konnten wir zeigen, dass Interaktionen mit langer Reichweite im Chromatin des kompakten *Drosophila*-Genoms weit verbreitet sind. Außerdem hat der Vergleich der Interaktionsprofile von verschiedenen Zeitpunkten und Gewebetypen ergeben, dass die Konfiguration des Chromatins größtenteils stabil bleibt.

Für zwei weitere 4C Projekte habe ich den Einfluss von großen genomischen Neuanordnungen auf die Chromatinstruktur von zwei Regionen im Mausgenom analysiert. Der Schwerpunkt des ersten Projekts liegt auf der Region um das *Shh* Gen. Meine Analyse zeigt, dass bei einer genomischen Inversion die Chromatinstruktur dieser Region zusammenbricht und die Kontakte neu verknüpft werden. Das zweite Projekt untersucht die Chromatinstruktur der Region um die beiden Gene *Ap2-γ* und *Bmp7*. Anhand der Analyse von 4C Profilen von vier Viewpoints, die in dieser Region verteilt sind, und der Ermittlung primärer Interaktionsdomänen für jeden dieser Viewpoints, konnte ich zeigen, dass diese Region von einer sogenannten Übergangszone (transition zone) deutlich in zwei Domänen geteilt wird. Desweiteren zeigt die Analyse der Interaktionsprofile von Allelen, die große chromosomale Umordnungen aufweisen, dass die Übergangszone eine wichtige Rolle für die Aufteilung der Region spielt. Alles in allem legen die beiden Studien dar, wie wichtig die Chromatinstruktur für Genregulation über lange Reichweiten und die Zuordnung von Enhancern zu ihren Zielgenen ist.

In Kapitel 4 beschreibe ich die Analyse eines High-throughput-Screens von chemischen Substanzen mit isogenen menschlichen kolorektalen Krebszelllinien. Mit diesem Pharmacogenetic Phenome Compendium (PGPC) Projekt haben wir

chemisch-genetische Interaktionen zwischen chemischen Substanzen und dem genetischen Hintergrund der jeweiligen isogenen Zelllinien untersucht. Mit Hilfe von High-throughput-Mikroskopie haben wir 1280 bioaktive Stoffe in zwölf isogenen Darmkrebszelllinien analysiert, die bestimmte Mutationen in Signalwegen aufweisen. Nach der Bildsegmentierung und der Extrahierung von Bildinformationen habe ich einen Algorithmus benutzt um einen redundanzfreien Satz von 20 phänotypischen Bildinformationen für die weitere Analyse auszuwählen. Meine Untersuchung der phänotypischen chemisch-genetischen Interaktionsdaten erlaubte es uns, die synergistischen Substanz-Kombinationen vorherzusagen, die Verbindungen zwischen Signalwegen aufzudecken und funktional verwandte Substanzen und biologische Prozesse zu gruppieren. Außerdem habe ich gezeigt, dass die kombinierte Anwendung von High-throughput-Mikroskopie und chemisch-genetischem Screening sensibler ist um das Interaktom zu entschlüsseln, als die jeweilige Methode allein. Für einen leichten Zugang zu den Daten und die Möglichkeit die Ergebnisse zu reproduzieren habe ich das R-Paket `PGPC` entwickelt, welches bei www.bioconductor.org eingereicht wird.

Abbreviations

3C Chromosome Conformation Capture. 12

4C Circularized Chromosome Conformation Capture. 1

AUC area under the curve. 66

BAM binary alignment/map. 19

BI Bliss independence. 67

CCLE Cancer Cell Line Encyclopedia. 94

CGP Cancer Genome Project. 95

ChIP-Seq Chromatin Immunoprecipitation sequencing. 10

CK2 Casein Kinase 2. 83

CLL Chronic Lymphocytic Leukemia. 83

CRISPR clustered, regularly interspaced, short palindromic repeats. 96

CRM *cis*-regulatory module. 11

CT chromosome territory. 11

CTCF CCCTC-binding factor. 57

CTNNB1 β-catenin. 61

CTRP Cancer Therapeutics Response Portal. 95

DMSO DMSO Dimethyl sulfoxide. 61

EGFR epidermal growth factor receptor. 83

ESC embryonic stem cell. 11

GRO-seq genomic run-on sequencing. 32

GWAS genome-wide association study. 58, 97

HDR homology-directed repair. 96

Hi-C 3C combined with paired-end high-throughput sequencing. 12

HSA highest single agent. 67

iPS cell induced pluripotent stem cell. 96

KO knockout. 61

LCR local control region. 11

MAD median absolute deviation. 25, 54

NHEJ non-homologous end joining. 96

NPA normalized proteasome activity. 68

NPI normalized percent inhibition. 67

PGPC Pharmacogenetic Phenome Compendium. 1

polII RNA polymerase II. 10

RPGC reads per genomic content. 35

RPKM reads per kilobase per million mapped reads. 35

RPM reads-per-million. 22

SB transposon sleeping beauty transposon. 37

SD standard deviation. 23

TAD topologically associating domain. 37

TF transcription factor. 11

TK Thymidine Kinase. 79

TSS transcription start site. 10

TZ transition zone. 44

WT wild type. 40

ZPA Zone of Polarising Activity. 37

ZRS ZPA Regulatory Sequence. 37

Gene names

Ap2-γ Activating Enhancer Binding Protein 2 Gamma

ap apterous

Bmp7 Bone Morphogenetic Protein 7

Shh Sonic Hedgehog

Chapter 1

Introduction

In recent years, with the advent of high-throughput technologies, biology has turned into a datadriven field. Nowadays, many areas can be covered by using different high-throughput techniques, such as sequencing, mass spectrometry, and automated microscopy, to name some examples. On the one hand, this allows one to study fundamental biological processes, for instance, gene regulation, in more detail, helping to increase our understanding and knowledge of these processes. On the other hand, these technologies are used in biomedical research to reveal the molecular basis of common diseases and the development of new therapeutics to treat diseases.

The generation of huge and complex data sets has reached a scale where data handling and analysis has become more and more challenging. These data sets allow one to investigate cell systems, networks, and connections of biological pathways and processes. The bioinformatic analysis and integration of data from different sources is central to our efforts to extract new biological findings and connect data obtained from the different technologies (Figure 1.1). To cope with the data, current methods need to be adapted and new analysis tools and pipelines have to be developed.

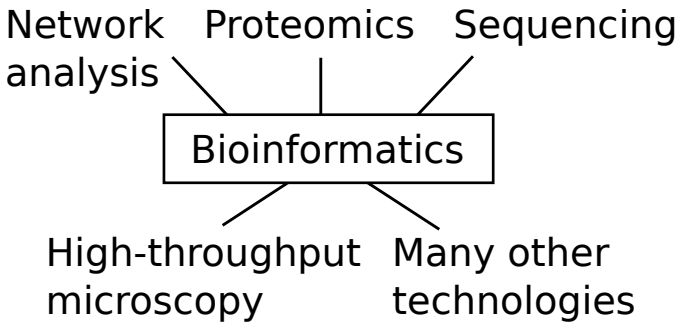


Figure 1.1: Bioinformatics is the connection between the different technology platforms.

In the following I give a brief introduction to the biological background and the technologies which were used to generate the data I analyzed.

1.1 Regulation of gene transcription

The genetic information stored in the DNA, is transcribed into complementary RNA in the processes of transcription. RNA polymerases are the enzymes that produce complementary RNA molecules from the DNA template. Several forms of RNA polymerases exist, of which RNA polymerase II (polII) is responsible for the transcription of coding transcripts from the nuclear DNA (Alberts *et al.*, 2007).

Transcription consists of three molecular steps (Alberts *et al.*, 2007). First, during transcription initiation, RNA polymerase is recruited to the DNA at the transcription start site (TSS) by the pre-initiation complex formed by general transcription factors. After recruitment, the polymerase starts the synthesis of the complementary RNA molecule with the first nucleotide. Second, during transcript elongation, the polymerase synthesizes the complementary RNA molecule while moving along the DNA template. Third, transcription is terminated and the polymerase stops the RNA synthesis and releases the RNA molecule for further processing.

Transcription is regulated at each of the three steps (Alberts *et al.*, 2007). Before a DNA sequence from highly condensed chromatin can be transcribed it has to be made accessible for binding of the polymerase. Next, the polymerase has to be recruited to the TSS by the pre-initiation complex and other transcription factors. After binding, the polymerase has to be enabled to move along the DNA template for proper transcript elongation. Finally, the release of the synthesized RNA molecule is controlled by the disassembly of the transcription machinery.

With more and more genomes being sequenced and improved methods to explore the transcriptome of cells in unbiased ways, there have been great advances in recent years, uncovering the existence and function of non-coding RNAs. The first of these studies used tiling arrays to identify the transcribed regions of the genome (Bertone *et al.*, 2004; David *et al.*, 2006), followed by studies using RNA sequencing technology (Nagalakshmi *et al.*, 2008). Nowadays, new protocols, combined with sequencing, further investigate the complex transcriptome, shining light on the complex transcriptome and function of non-coding RNAs (for current reviews see (Pelechano and Steinmetz, 2013), (Hausser and Zavolan, 2014) and (Fatica and Bozzoni, 2014)).

1.1.1 Histones and chromatin

In eukaryotes, DNA is tightly wrapped around nucleosomes in stretches of 146 bp, forming highly condensed chromatin. Nucleosomes are protein octameres formed by four dimeres of the H2A, H2B, H3 and H4 histone proteins. Enzymes can modify the N-terminal tails of these proteins, which are protruding from the

nucleosome. These modifications have implications on the interactions between different nucleosomes and the binding of proteins to DNA wrapped around nucleosomes. Chromatin Immunoprecipitation sequencing (ChIP-Seq) studies have revealed functional consequences of histone modifications, for example activating or repressing transcription (see (Bannister and Kouzarides, 2011) for a current review).

1.1.2 Transcription factors and long-range gene regulation

Transcription factor (TF)s are proteins that bind to the DNA at promoter regions or regulatory sequences. These regulatory sequences can be located far, as measured in linear genetic distance, from the actual TSS, acting as distal enhancers or repressors depending on whether they enhance or repress transcriptional output. TFs can bind to DNA directly, histones, or other TFs already bound to DNA. The general or basal transcription factors are required to form the pre-initiation complex and subsequently the transcription machinery. Other tissue specific TFs are only expressed in certain tissues to regulate transcription and generate a tissue specific transcriptional output. Temporal and spatial dynamics in binding of several TFs to *cis*-regulatory module (CRM)s tightly regulate transcription during development (for a current review see (Spitz and Furlong, 2012)).

1.1.3 The role of chromatin 3D structure in gene regulation

To which extent chromatin 3D structure, and its changes, influence gene transcription is still unclear, but an increasing set of evidence links the spatial organization of chromatin and the regulation of gene transcription.

During interphase the cell nucleus is compartmentalized. Individual chromosomes occupy distinct chromosome territory (CT)s in the nucleus (Cremer and Cremer, 2001). Furthermore, it is known that the arrangement of CTs relative to each other and in respect to the center and periphery of the cell nucleus is tissue specific (Parada *et al.*, 2004). Importantly, CTs were shown to display a marked intermingling making inter-chromosomal interaction possible (Branco and Pombo, 2006).

The dialog between CRMs such as promoters, enhancers, silencers and insulators was proposed to act through direct association via DNA looping. The developmental stagespecific transcriptome profile can thus be linked to a specific set of DNA loops making up the cell type characteristic spatial arrangement of chromatin during interphase. Indeed, it was shown that the spatial arrangement can change during differentiation and a whole locus can be displaced in this process. For example, the *HoxB* locus changes significantly during differentiation (Chambeyron and Bickmore, 2004). After treatment of embryonic stem cells (ESCs) with retinoic acid, the *HoxB* chromatin fiber decondenses and loops out of the MMU11 CT. This process is accompanied by the sequential expression of the different *HoxB* genes, suggesting a strong spatio-temporal correlation. Despite the compact *Drosophila*

melanogaster genome it was shown that long-range interactions are established between the two *Hox* complexes which are separated by 10 Mb of DNA on the same chromosome arm (Bantignies *et al.*, 2011). These contacts increase during development depending on Polycomb group (PcG) proteins, showing that there are changes in chromosome conformation. Just recently, a study showed that the looping of the local control region (LCR) to the β-globin promoter, mediated by GATA1, in erythroblasts is required for polII recruitment and phosphorylation (Deng *et al.*, 2012).

Further studies in a genome-wide context are needed to further understand the influence of CRMs, epigenetic modifications, gene looping, and intra- and inter-chromosomal interactions on chromatin 3D structure and to elucidate what are the actual driving forces that organize the 3D structure of chromatin in the context of gene regulation.

1.2 Chromosome conformation capture

The development of the Chromosome Conformation Capture (3C) protocol (Dekker *et al.*, 2002) dramatically increased the possibilities to study the 3D structure of chromosomes in recent years. The combination of the 3C method with high-throughput sequencing now provides the possibility to study chromatin interactions on a genome-wide scale (Zhao *et al.*, 2006; Stadhouders *et al.*, 2013; Dostie *et al.*, 2006; Lieberman-Aiden *et al.*, 2009; Fullwood *et al.*, 2009). All protocols consist of 5 main steps (Figure 1.2). First, interacting loci are captured using formaldehyde cross-linking, followed by DNA fragmentation by either restriction enzyme digestion or sonication. In the next step, fragments are ligated under dilute conditions, which favor intra-complex ligation, generating unique ligation products of interacting loci. These ligation products are then purified and further processed and detected by different techniques.

The original 3C method employs regular PCR with two selected primers to detect pairwise interaction products (Dekker *et al.*, 2002). The Circularized Chromosome Conformation Capture (4C) methods use inverse PCR amplification followed by either microarray detection or high-throughput sequencing to detect all fragments ligated to a viewpoint of choice (Zhao *et al.*, 2006; Stadhouders *et al.*, 2013). The 5C protocol allows one to detect interactions for a large number of fragments by employing multiplexed ligation mediated amplification with a pool of primers for thousands of fragments (Dostie *et al.*, 2006). The protocol of 3C combined with paired-end high-throughput sequencing (Hi-C) first uses restriction enzyme digestion to fragment the DNA after cross-linking, followed by filling in the sticky ends with biotinylated nucleotides before ligation (Lieberman-Aiden *et al.*, 2009). The resulting ligation products are sheared and enriched for fragments containing ligation junctions by pull-down with streptavidin-coated beads (Lieberman-Aiden *et al.*, 2009). The resulting library is directly sequenced using paired-end high-throughput sequencing (Lieberman-Aiden *et al.*, 2009). For the

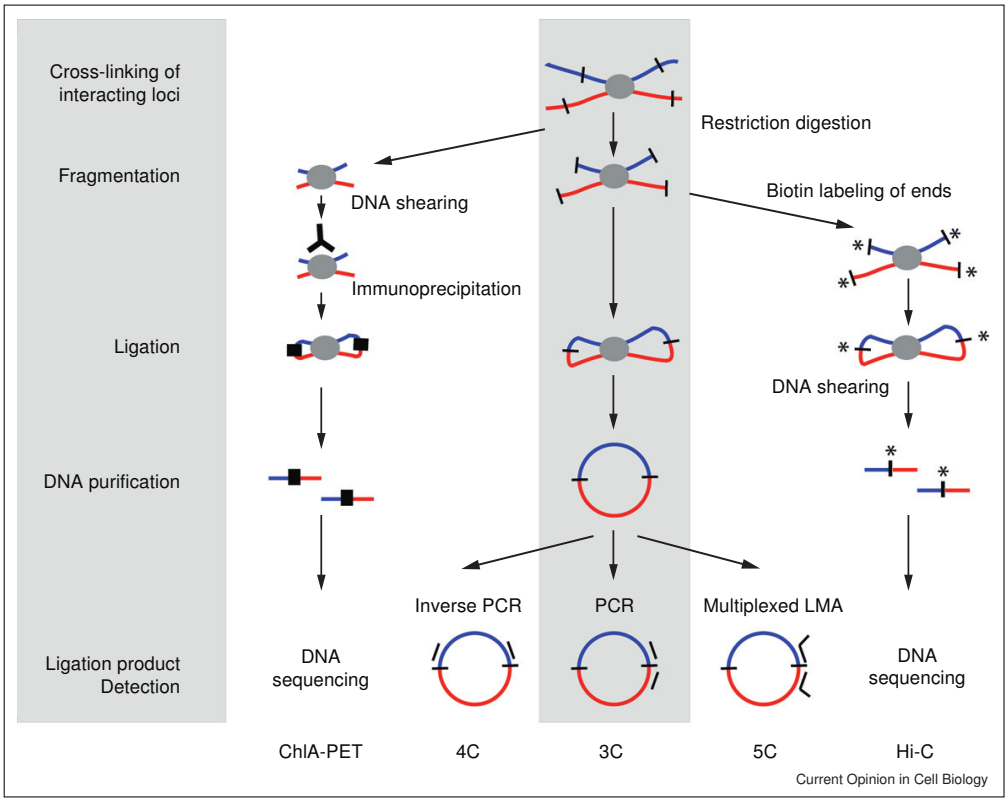


Figure 1.2: Schematic representation of the different 3C methods. Details on the different methods are described in the text. Reproduced with permision from Sanyal *et al.* (2011).

ChIA-PET protocol the cross-linked DNA is sheared, followed by an immunoprecipitation step to enrich for contacts mediated by a specific protein of interest (Fullwood *et al.*, 2009). Proximity ligation is performed with barcoded DNA linkers and the obtained paired-end tags are analyzed by high-throughput sequencing (Fullwood *et al.*, 2009).

The 4C sequencing technique was used for all projects presented in Chapter 2.

1.3 Synthetic genetic interactions

Synthetic genetic interactions were first described by Calvin Bridges in the form of synthetic lethality in 1922. Crossing certain homozygous Drosophila melanogaster strains, he noticed that some genes are lethal only in combination (Bridges, 1922). In general, the term synthetic genetic interaction describes the fact that only the combination of two genetic perturbations gives rise to a distinct phenotype, while each genetic perturbation alone does not. This principle can be readily transferred to different types of perturbations, e.g., over-expression of genes or drug treatments. An schematic example for a synthetic lethal interaction is shown in Figure 1.3.

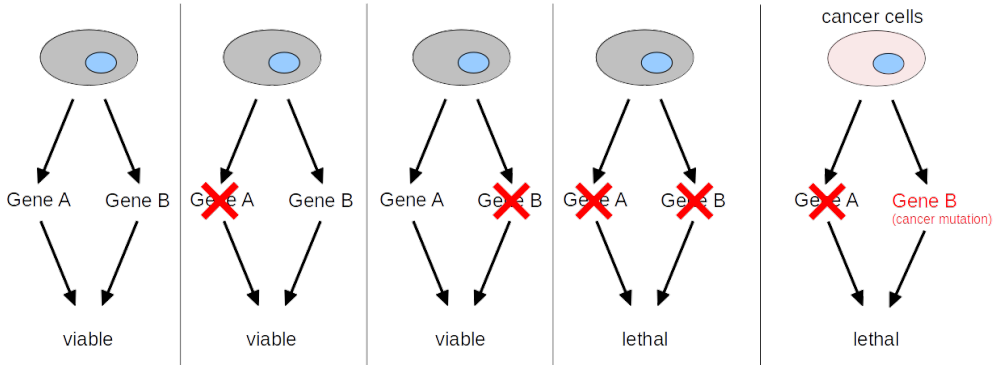


Figure 1.3: Schematic representation of a genetic interaction. Inhibition of gene A or gene B does not affect cell fitness, while the combined inhibition of A and B is lethal. In cancer cells, which carry a mutation in gene B, the inhibition of gene A alone is lethal.

The reason that such synthetic genetic interactions can be observed lies in the ability of cells and organisms to buffer genetic and environmental perturbations (Hartman *et al.*, 2001). With the possibility of high-throughput methods, genetic interactions were systematically studied in yeast by looking at the fitness of yeast double mutant colonies (Costanzo *et al.*, 2010). The result of this study is a comprehensive genetic interaction network for yeast. Furthermore, a functional mapping of genes was possible, because the interaction profiles of genes involved in the same or similar processes tended to cluster together. Using combinatorial RNAi, genetic interactions were recently studied in higher organisms and mammalian systems (Horn *et al.*, 2011; Roguev *et al.*, 2013; Laufer *et al.*, 2013). With

the use of the identified genetic interactions, these studies were able to reconstruct protein complexes and signaling pathways.

1.3.1 Implications for cancer therapy and drug discovery

As it is often difficult to target the known oncogenes driving cancer development, the use of synthetic genetic interactions, especially synthetic lethality, provides a new strategy for cancer therapy (Kaelin, 2005). The specific killing of BRCA1/2 mutated cancer cells with PARP inhibitors (Bryant *et al.*, 2005) represents a major milestone for this strategy of drug discovery. In recent years, several screens for synthetic lethal interactions of oncogenic mutations were performed. Examples are screens for synthetic lethal interactions of oncogenic RAS with different approaches in isogenic cell lines (Barbie *et al.*, 2009; Luo *et al.*, 2009) that identified different potential targets.

While compound screens for synthetic lethal interactions provide a direct lead compound, the target of the compound might not be known and secondary screens to detect the compound targets are required. On the other hand, RNAi screens directly reveal synthetic lethal interactions between targeted genes. Although an inhibitor for the identified target might not be available, RNAi screens reveal the connections of cellular pathways which might provide the possibility to inhibit other targets within the same pathway for which inhibitors are available. Furthermore, understanding the connections of pathways has become more and more important to understand the development of resistance mechanisms upon targeted therapies. For a recent review on using synthetic lethality for drug discovery see (Chan and Giaccia, 2011).

1.4 High-throughput microscopy screening

Automated fluorescence microscopy screening has become one of the most powerful tools to investigate cellular processes (see the recent review by (Conrad and Gerlich, 2010)). By using different fluorescent dyes or antibodies to label cellular components, high-content microscopy images allow one to identify sub-cellular structures in cells and track their changes upon perturbations. This technology was used successfully to characterize drugs and delineate drug targets based on the induced phenotypic changes (Perlman *et al.*, 2004; Young *et al.*, 2007). Combined with RNAi technology, high-throughput microscopy was used to investigate cellular processes on a genome wide level (Fuchs *et al.*, 2010; Neumann *et al.*, 2010). In recent years, high-throughput microscopy has been employed to study genetic interactions across multiple phenotypic features (Horn *et al.*, 2011; Laufer *et al.*, 2013).

Chapter 2

Analysis of 4C sequencing data

This chapter comprises three projects which used 4C sequencing to study the chromatin 3D structure during *Drosophila* and mouse embryogenesis. The content of this chapter is also described in the following manuscripts: Ghavi-Helm *et al.* (2014); Klein *et al.* (submitted for publication); Tsujimura *et al.* (submitted for publication); Symmons *et al.* (in preperation).

Several analysis analysis tools have been developed for 4C sequencing data already (Splinter *et al.*, 2012; van de Werken *et al.*, 2012; Thongjuea *et al.*, 2013). However, none of these methods provide a statistical method to perform differential 4C sequencing analysis between different tissue types or developmental stages. To address this need I generated the **R** package `FourCSeq` to process and analyze the obtained data. The processing steps and functionality of the package are described together with other developed methods in Section 2.1.

The results of the three projects are the content of Section 2.2. In the first project we studied 4C sequencing data from *Drosophila* embryos obtained at two developmental stages and from two types of tissues. The project aimed to investigate how interactions between enhancers and promoters are established and how they change between developmental time points and tissue types. Therefore, it was necessary to detect specific enhancer promoter interactions and to quantitatively compare interaction profiles. To get a genomewide overview, we used over 100 viewpoints spaced throughout the *Drosophila* genome. The wet lab experiments in this project were all performed by Yad Ghavi-Helm from the Furlong group at EMBL. Starting from my interaction calls and differential interaction analysis, follow-up analysis were done by Tibor Pakozdi from the Furlong group.

Both, the second and third project, studied the influence of genomic rearrangements on gene expression and chromatin structure in mouse embryos. The aim of these projects was to understand how 4C interactions profiles change when the genomic structure is reshuffled. This required tools to handle the genomic rearrangements and to visualize global interaction profile changes.

In the second project we focused on the genetic locus of the *Shh* gene, which is important for proper limb development. In this project the experiments were performed by Orsolya Symmons and Silvia Remeseiro in the Spitz group at EMBL.

The mouse *Ap2-γ - Bmp7* locus was studied in the third project. This locus contains two genes which are both important during embryogenesis. Taro Tsujimura performed the experiments for this project in the Spitz group at EMBL.

2.1 Materials and Methods

2.1.1 Chromosome conformation capture assays

For the projects presented in this chapter, my collaborators used 4C sequencing protocols (see Section 1.2) which all included a second round of restriction enzyme digestion in order to generate smaller fragments that can be more efficiently circularized and amplified by PCR. An overview of the workflow of the `FourCSeq` package used to analyze the obtained 4C sequencing data is shown in Figure 2.1.

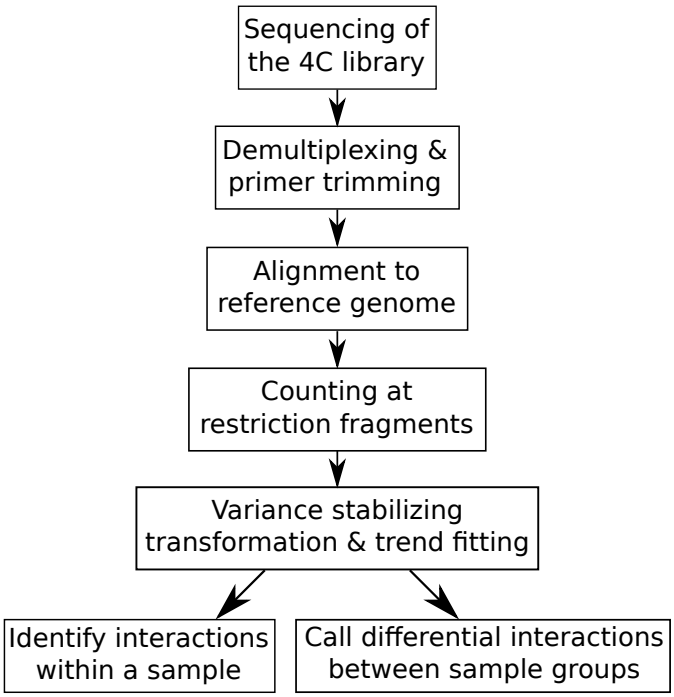


Figure 2.1: Workflow of the 4C sequencing analysis.

2.1.2 Data processing

2.1.2.1 Demultiplexing and trimming of primer sequences

The starting point for my analysis were the sequencing files in FASTQ format from Illumina HiSeq sequencing runs. In each run several viewpoints had been multiplexed. I demultiplexed the libraries using the corresponding viewpoint primer

sequences and, if present, additional barcodes that were used. In this step the barcodes and viewpoint primer sequences were trimmed, keeping the sequence of the restriction enzyme at the start of the read.

After demultiplexing, I aligned the remaining sequences to the corresponding full reference genome using the alignment tool Bowtie (Langmead *et al.*, 2009) for the mouse data and Novoalign (http://www.novocraft.com) for the *Drosophila* data.

The alignment output files in the binary alignment/map (BAM) format are the starting point for the analysis in **R** with the `FourCSeq` package.

2.1.2.2 Generating a fragment reference

For the statistical analysis, I generated a count table, with one row for each restriction fragment, and one column for each sample, with the table entries indicating how many reads have been assigned to each restriction fragment in each sample. By restriction fragment I mean the sequence in between two cutting sites of the first restriction enzyme used in the protocol. To assign the reads to the restriction fragment where they originate from I cut the reference genome *in-silico* using the recognition sequences of the restriction enzymes used in the protocol. The second restriction enzyme cuts these fragments again, creating smaller fragment ends, that can be more efficiently circularized and amplified by PCR. Correspondingly, fragment ends were defined as the sequence between the start or end position of a restriction fragment and the closest cutting site of the second restriction enzyme within the fragment sequence (Figure 2.2 a). The size of the fragment ends were defined by the length of the corresponding sequences. In the case of 100 % restriction enzyme cutting efficiency, only fragments that contained a cutting site of the second restriction enzyme would be amplified. Therefore, the fragments defined by the first restriction enzyme were categorized into two classes:

1. visible fragments, that contain at least one cutting site of the second cutter.

2. blind fragments, that do not contain a cutting site of the second cutter.

For further analysis blind and small fragments can be filtered out. In the current implementation I used the following classifications of fragments, which is visualized in Figure 2.2:

1. valid fragments that contain at least one cutting site of the second cutter.

2. invalid fragments that do not contain a cutting site of the second cutter or for which the length of both fragment ends are smaller than a threshold.

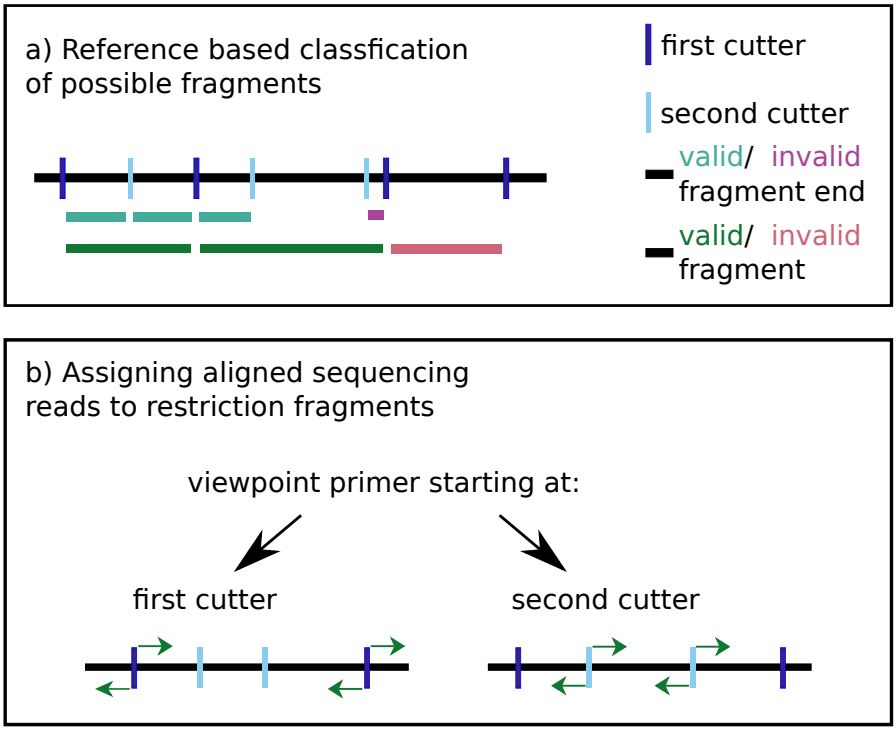


Figure 2.2: Schematic of the rules to define valid fragments that are used subsequently in the analysis.
a) The pink fragment end is smaller than the defined threshold, but since the other fragment end is valid, the fragment is kept for analysis. The red fragment is invalid because it does not contain a cutting site of the second restriction enzyme.
b) For viewpoint sequencing primers starting at the first restriction enzyme cutting site, reads (green arrows) that start at the fragment ends and are oriented towards the fragment middle are kept for analysis.
For viewpoint sequencing primers starting at the second restriction enzyme cutting site, reads (green arrows) that start right next to the cutting site of the second restriction enzyme and are directed towards the ends of the fragment are kept for analysis.

2.1.2.3 Mapping of primer sequences

In order to find the viewpoint fragment, I mapped the primer sequences to the reference genome. This also served as a sanity check that the used primer sequence in the protocol was unique in the reference genome. The corresponding fragment was used to calculate the genomic distance to all fragments on the same chromosome. The distance between the viewpoint and a fragment was calculated as the genomic distance between the middle of the viewpoint fragment and the middle of the other fragment. These distances to the viewpoint were used for further analysis.

2.1.2.4 Assigning aligned reads to the fragment reference

In order to generate a count table with the number of counts observed at each fragment, I assigned the aligned reads to the generated fragment reference. To filter out non-informative reads, I used the following criteria, which are motivated by the 4C sequencing protocol.

First, the reads must start directly at a restriction enzyme cutting site of the first or second restriction enzyme, depending on whether the sequencing primer started at the first or second restriction enzyme cutting site of the viewpoint fragment respectively. Second, the orientation of the read at the fragment end is defined by the sequencing primer (Figure 2.2 b). The reads must be directed towards the middle of the fragment, if the library was prepared with a sequencing primer starting at the first restriction enzyme cutting site of the viewpoint fragment. For libraries prepared with primers starting at the second restriction enzyme cutting site of the viewpoint fragment, the reads must be directed towards the fragment ends. For subsequent analysis I only counted reads at fragments ends that fulfilled these criteria. Additionally, I summed up the reads counted at each fragment end to obtain one count value per fragment.

2.1.2.5 Correcting the genomic distances for modified genomic regions

The reads of samples from mice carrying large genomic rearrangements were aligned to the normal reference genome and assigned to the generated fragment reference as described in the two previous sections. However, for these samples the genomic distance to the viewpoint had changed for the modified genomic regions. In order to correct this for an inversion, I *in-silico* inverted the modified region by linearly transforming the genomic coordinates of the fragments in the regions between the two break points of the inversion.

$$x_{\text{inv}} = p_2 - (x - p_1), \tag{2.1}$$

where x_{inv} is the new genomic coordinate after *in-silico* inversion, p_1 and p_2 are the genomic positions of the break points ($p_1 < p_2$), and x is the original genomic coordinate.

After this *in-silico* inversion, the 4C profiles showed the expected signal decay with genomic distance from the viewpoint. The two fragments that contained the breakpoints of the inversion were removed because the distances were ambiguous with respect to the corresponding fragment ends.

For the samples with deletions, I used a similar approach. Here the deleted part was removed *in-silico* including the fragments that contained the two breakpoints. The genomic coordinates after the deletion break points were shifted according to the genomic size of the deletion.

$$x_{\text{del}} = \begin{cases} x & \text{if } x < p_1 \\ x - (p_2 - p_1 + 1) & \text{if } x > p_2 \end{cases}, \tag{2.2}$$

where x_{del} is the new genomic coordinate after *in-silico* deletion, p_1 and p_2 are the genomic position of the break points ($p_1 < p_2$), and x is the original genomic coordinate.

The 4C profiles showed the expected signal after this *in-silico* deletion. In some cases I shifted the whole 4C profile for visualization so that the viewpoint positions of samples with inverted or deleted genomic regions aligned with the viewpoint position of the wild-type samples.

2.1.2.6 Quality control

To check the quality of a 4C library, I calculated several statistics when the aligned reads were assigned to the fragment reference. For each library the following numbers were calculated:

1. the total number of reads in the library
2. the number of aligned reads in the library
3. the number of low quality reads, if a threshold on the alignment quality was used
4. the number of reads that assigned to the fragment reference
5. the ratio of reads assigned to the fragment reference and number of aligned reads after removal of low quality reads

For the data I worked on, the percentage of aligned reads that could be assigned to the fragment reference was typically around 70-95 %. A value in that range should be a reasonable target for further 4C sequencing libraries.

2.1.2.7 RPM normalization

The easiest way to normalize the 4C signal between different libraries was the reads-per-million (RPM) normalization. In this method all read counts were divided by

the number of all reads mapped to either the reference genome or the viewpoint chromosome and then multiplied by one million.

$$RPM_i = \frac{c_i}{n} \cdot 1 \times 10^6, \tag{2.3}$$

where i is the index of the fragments, c_i is the count value observed for fragment i, and n is the total number of reads assigned to either the whole reference genome or the viewpoint chromosome.

2.1.3 Detecting interactions

2.1.3.1 Variance stabilizing transformation

The count values per fragment usually spanned several orders of magnitude. If the raw data is used for analysis, the standard deviation across samples are very large for fragments with a high number of counts. However, if the count values are transformed with a simple logarithmic transformation, fragments with a low number of counts show large standard deviations across samples. Both approaches are unsatisfactory because the analysis would skew the analysis towards fragments very close or very far from the viewpoint. Therefore, I used the variance stabilizing transformation introduced by Anders and Huber (2010) and implemented in the `DESeq2` package (Love *et al.*, 2014), to transform the count k_{ij} of fragment i in sample j to $v(k_{ij})$. The standard deviations (SDs) showed less dependence on the fragment abundance after this transformation as shown in Figure 2.3.

2.1.3.2 Trend fitting

With genomic distance from the viewpoint, the 4C signal decays towards a constant background level. This decay trend $f_j(d_i)$ is fitted using the transformed count values $v(k_{ij})$ as a function of the logarithm of the genomic distance d_i from each fragment i to the viewpoint.

Because the signal should monotonously decrease with increasing distance to the viewpoint, I used the smooth monotone fit implemented in the `fda` package (Ramsay *et al.*, 2014). Further, I assumed that the profile decay is symmetric around the viewpoint. Therefore, I combined the transformed count values of both sides to calculate the distance dependence fit. An example for the symmetric monotone fit is shown in Figure 2.4.

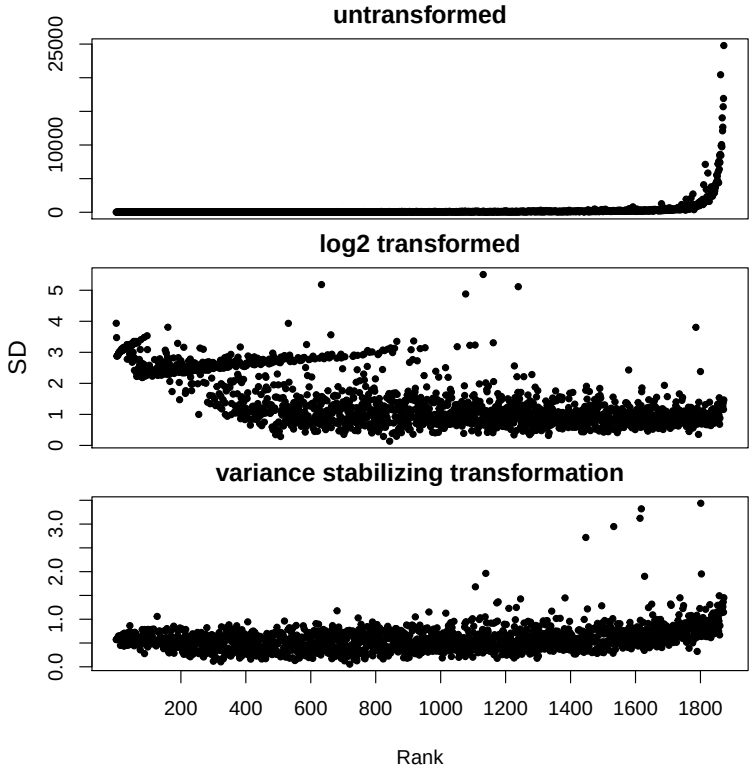


Figure 2.3: Variance-stabilizing transformation. For each fragment, the standard deviation of its count data was computed across all samples for the *apterous* CRM viewpoint. The plots visualize the distributions of these values for all fragments. Fragments close to the viewpoint are on the right side with higher count values. When the untransformed count data are considered (upper panel), the standard deviations are very large for high abundance fragments (close to the viewpoint). When the count data are considered on the logarithmic scale (middle panel), the standard deviations are very large for low abundance fragments (far from the viewpoint). Both effects would make the analysis highly susceptible to noise either close or far from the viewpoint respectively. When the data are transformed using a variance stabilizing transformation, the standard deviations show less dependence on the fragment abundance, allowing for a more consistent statistical treatment across the whole dynamic range of the data.

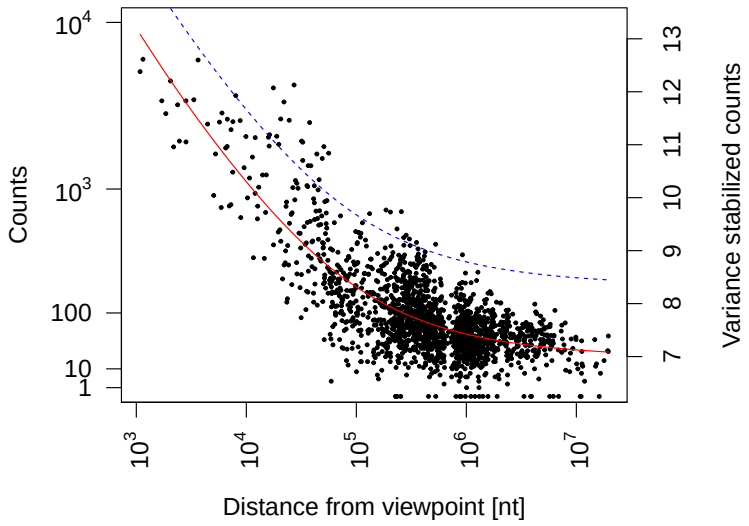


Figure 2.4: An example symmetric monotonous fit of the variance stabilized count data over log10 distance from the viewpoint. The red line shows the fit and the blue dashed line is the fit + 3σ of the fit residuals.

2.1.3.3 z-scores of residuals

To detect specific interactions that have a higher interaction frequency than expected at a given distance from the viewpoint, I calculated z-scores from the fit residuals and looked for large positive z-scores. I calculated the z-scores in the following way:

$$z_{ij} = \frac{v(k_{ij}) - f_j(d_i)}{\sigma_j}, \tag{2.4}$$

where $\sigma_j = \mathrm{MAD}_i(v(k_{ij}) - f_j(d_i))$, the median absolute deviation, is a robust estimator of scale, i runs over all fragments and j over all samples.

Assuming that the calculated z-scores follow Normal distribution under the null hypothesis, I calculated one sided p-values. To adjust for multiple testing, I used the method of Benjamini-Hochberg (Benjamini and Hochberg, 1995) to control for the false-discovery rate.

Specific interactions could then be detected by looking for fragments with large positive z-scores and low adjusted p-values.

2.1.4 Detecting changes

I observed that the distance dependence of the 4C signal was variable between samples. These differences were most likely due to technical differences in the cross-linking, restriction enzyme digestion, and PCR amplification steps. For comparisons between different samples of different conditions this had to be taken into account.

To address this problem I calculated a matrix of normalization factors n_{ij}, such that the scaled read counts $n_{ij}k_{ij}$ for fragment i became comparable across the samples j. For this, the normalization factors had to present the distance dependence on the scale of raw counts. Hence, I back transformed the fitted values f_{ij} to the scale of raw counts, using the inverse of the variance stabilizing transformation, and scaled these values to the geometric mean across samples to obtain the normalization factors.

$$n_{ij} = \frac{v^{-1}(f_j(d_i))}{\sqrt[J]{\prod_{j=1}^{J} v^{-1}(f_j(d_i))}}, \tag{2.5}$$

where n_{ij} is the normalization factor, $v^{-1}(f_j(d_i))$ is the back transformed fitted value at the genomic distance d_i. The index i runs over all fragments and j over all samples.

With these normalization factors I used the model implemented in the `DESeq2` package to detect differences between conditions (Love *et al.*, 2014). In this approach, the normalized fragment counts for each single fragment were quantitatively compared between conditions. Using the Wald-test statistic, the estimated fold-changes between conditions were compared to the variability observed between biological replicates. A change in interaction frequency for a fragment was only called significant if the observed fold-change between conditions was significantly higher than expected from the size of the changes seen between replicates.

2.1.5 Quantifying asymmetries

2.1.5.1 Segmentation of the 4C signal

For analyzing changes in the global interaction patterns upon genomic rearrangements, it is important estimate the primary interaction domain of a viewpoint. This is the genomic domain where the 4C signal is clearly above the background signal. To estimate the primary interaction domain for each viewpoint, I used the implementation by Huber *et al.* (2006) of a well established segmentation approach for the analysis of microarray data. This approach fits piecewise constant functions to the input data. The resulting change points of these fits define the segment boundaries. I used this algorithm on the 4C signal of each viewpoint, segmenting the signal into 3 segments (primary interaction domain, left and right region outside the primary interaction domain). For each viewpoint I removed a 10 kb

window around the viewpoint and all fragments that contained 0 counts across all experiments before applying the algorithm to each experiment individually.

2.1.5.2 Using the cumulative 4C signal

To quantifiy asymmetries in the interaction frequencies of viewpoints I calculated cumulative count distributions to each side of the viewpoint. Fragments with a genomic distance of less than 10 kb to the viewpoint, which have a high number of counts, were removed to reduce their strong influence on the cumulative distribution. To make the different libraries comparable, I normalized the cumulative signal to the total counts obtained for each library in the selected window size. With this normalization the cumulative signal in both direction sums up to 1.

2.1.6 Visualization of the 4C signal

2.1.6.1 Smoothing

The raw 4C signal of read counts observed at fragments can show spikes at certain fragment positions. In order to visualize the interaction profile more robustly, the signal can be smoothed in fragment windows. For our data I implemented this smoothing in such a way that a window with an odd number of fragments was used. The average read count observed for the fragments in this windows was assigned to the middle fragment. The result was a smoother interaction profile from the 4C data.

2.1.6.2 Hit fraction

When only a small starting amount of cells is available for the 4C protocol, the large number of PCR cycles can result in PCR artifacts. This will be visible as strong peaks along the genome, which are not reproducible between replicates. In such cases, it is not possible to directly use the 4C signal for the analysis of chromatin interactions. To address this problem I used an adaption of the transformation to unique coverage, which meant to collapse the whole coverage to 1 if at least one read was observed per fragment (Splinter *et al.*, 2012). Instead of using only 1 as a single threshold, I used several threshold values. Fragments for which the observed number of reads exceeded these thresholds were called *hits*. To estimate the contact propensity of a region with the viewpoint, I used different window sizes of fragments to calculate the fractions of hits observed in a given window. This number was then assigned to the middle fragment of each window as a *hit fraction*. A schematic of this procedure is shown in Figure 2.5.

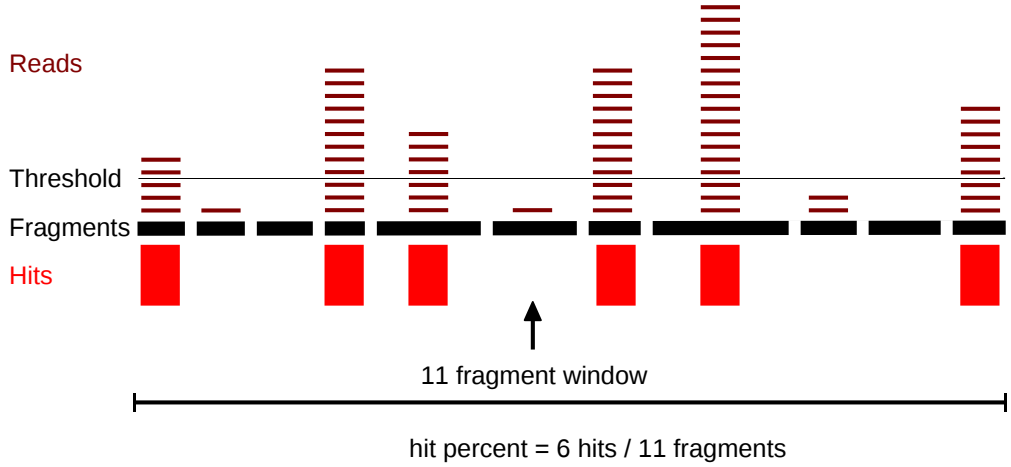


Figure 2.5: Schematic representation of the hit fraction calculation. Fragments for which the number of observed reads exceed the threshold are called hits. For the shown window of 11 fragments the hit fraction is calculated as $\frac{6}{11}$.

2.2 Results

2.2.1 Investigation of the chromatin interactome in developing *Drosophila* embryos

Drosophila embryogenesis is a well studied process, especially in the field of transcriptional regulation (Gregor *et al.*, 2014). After egg lay, the process proceeds very rapidly and the first instar larvae hatches after 21-22 hours (Campos-Ortega and Hartenstein, 1985). Gene transcription is very dynamic during embryogenesis (Graveley *et al.*, 2011). The robust process requires tight regulation of the dynamic gene transcription by enhancer and insulator elements (Levine and Davidson, 2005). To which extent the 3D chromatin structure, especially in enhancer looping, is involved in the dynamic regulation of transcription is unclear. To address this point, my collaborators generated 4C sequencing data for 103 viewpoints, spaced throughout the *Drosophila* genome. As viewpoints my collaborators chose developmental enhancers which showed diverse dynamics during embryogenesis (Zinzen *et al.*, 2009), priming our study to detect dynamic changes in chromatin 3D structure. Samples were taken from embryos at 2-4 h and 6-8 h after fertilization, at which time the embryos undergo marked morphological and transcriptional changes. In addition to whole-embryo samples from the two developmental time points, my collaborators also adapted the BiTS-ChIP method by Bonn *et al.* (2012) to perform mesoderm-specific 4C experiments. For the 4C sequencing experiment my collaborators used two restriction enzymes that cut double stranded DNA at specific 4 bp long recognition sequences. The first restriction enzyme used was *DpnII* and the second *NlaIII*.

After sequencing, my analysis started with demultiplexing the 4C sequencing

libraries as described in Section 2.1.2.1. Because the obtained sequencing reads were 105 bp long, I observed reads of short fragments that included the religation site and additional sequences from the ligated fragments. Because of the ambiguity of the two sequences in these reads, they could not be aligned unambiguously to the reference genome. Therefore, I scanned unaligned reads for the restriction fragment sequence and if such a site was present, the read was trimmed after this position, leaving the cutting site. After this trimming step a new alignment was attempted with the remaining sequence. With this strategy, on average approximately 10 % of the reads for each library could be aligned in a second step.

To assign the aligned reads to a fragment reference, I generated a reference of restriction fragments, as described in Section 2.1.2.2. The aligned reads were assigned to this reference as described in Section 2.1.2.4. For most libraries the percentage of aligned reads that could be assigned to a fragment is between 70 and 90 %. For further analysis I only kept valid fragments and removed all fragments that did not contain a cutting site of the second restriction enzyme and for which both fragment ends were smaller than 20 bp (Section 2.1.2.2).

To test if biological replicates were consistent, I generated scatter plots. For the 103 viewpoints, these plots showed good agreement for high count values. Higher variation was observed at lower count values, where the data are dominated by noise. An example plot for the *apterous* CRM viewpoint (*ap* viewpoint) is shown in Figure 2.6.

2.2.1.1 Detecting interactions

To find strong interactions that stand out from the general decay trend I proceeded as described in Section 2.1.3. As a first step I removed all fragments for each viewpoint that either contained a low number of reads or that were too close to the viewpoint. The latter fragments showed a high signal due to ligation caused by close linear proximity. In the first case, I excluded fragments from further analysis which had less than 40 counts on average across samples. For the second case, I automatically defined the first valid fragments as those that occurred after the initial signal decrease, starting from the viewpoint, where the signal began to increase again.

The parameters of the variance stabilizing transformation (see Section 2.1.3.1) were estimated on the count data of the fragments that passed these filter steps. Using the method described in Section 2.1.3.2, I fitted a smooth monotone symmetric curve to the data to estimate the general decay trend. A fit example is shown in Figure 2.4. From the fit residuals I calculated z-scores and associated p-values as described in Section 2.1.3.3.

The distribution of z-scores obtained for two libraries of the *ap* CRM viewpoint are shown in Figure 2.7. For the second replicate *MESO 68h 2* a second peak is observed in the histogram. This is due to fragments that contain 0 counts in this library which has a lower coverage. Since we are interested in finding strong interactions on the positive side of the distribution, we can continue with our approach

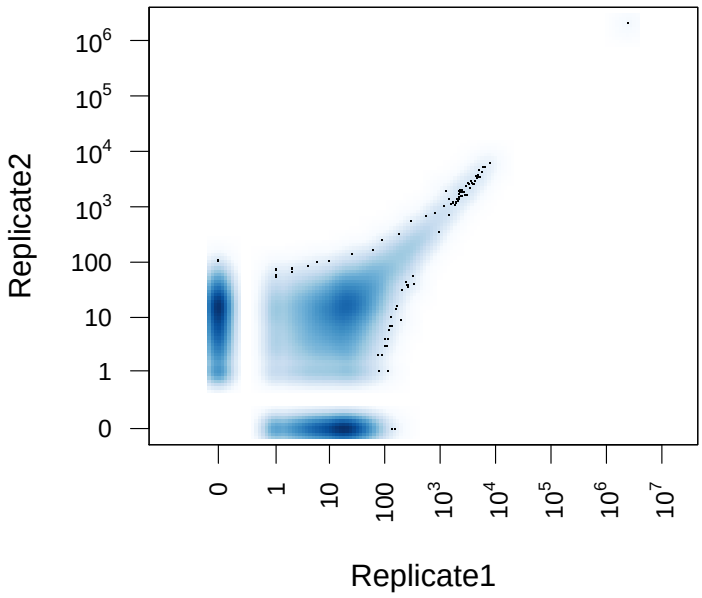


Figure 2.6: Scatter plot between two biological replicates of the *apterous* CRM viewpoint for whole-embryo tissue at 6-8 h after fertilization. In the plot a density estimate of the pairwise distribution of count values per fragment is shown. The x- and y-axes (drawn in logarithmic scale, with zero) correspond to the counts for the fragment in two biological replicate libraries for the same viewpoint and biological condition. The replicates show good concordance for higher count values. Fragments with 0 counts for both replicates are removed.

and capture the strongest contacts. However, if the shift of the distribution towards smaller values gets more extreme this might lead to an overestimation of the median absolute deviation and hence an underestimation of z-scores. It is therefore important to check the distribution of the calculated z-scores.

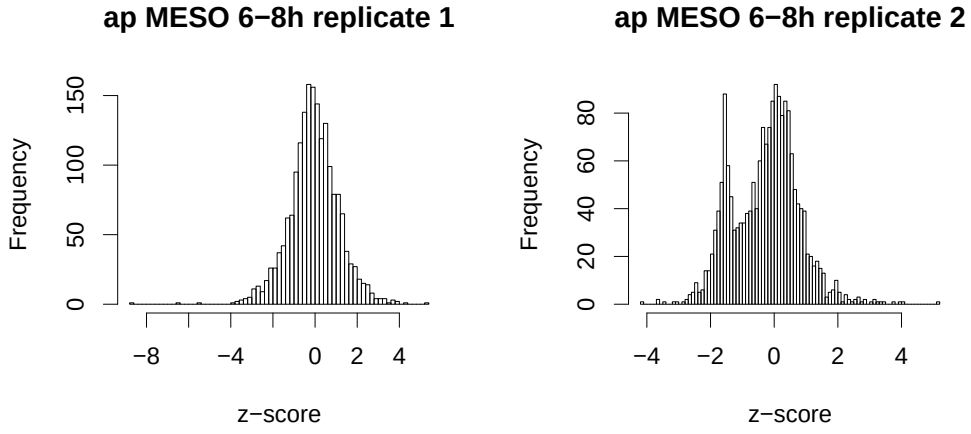


Figure 2.7: Distribution of z-scores calculated for two libraries of the *ap* CRM viewpoint.

For the follow-up analysis, we defined interacting regions using the following thresholds. For each fragment both replicates z-scores must be larger than 3 and in at least one replicate the adjusted p-value must be smaller than 0.01. Figure 2.8 shows the interacting fragments for the two replicates of the whole embryo 6-8 h condition of the *ap* CRM viewpoint. The interaction between the viewpoint and the *apterous* (*ap*) gene promoter on the right side of the viewpoint is captured in both replicates. Additional interactions that could not be directly linked to specific genomic elements are captured as well.

2.2.1.2 Long-range interactions in the *Drosophila* genome

For further analysis my collaborators merged interactions into interacting regions if interacting fragments were less than 1 kb apart. With these parameters we identified 1036 interacting regions based on 4247 interacting fragments (Ghavi-Helm *et al.*, 2014). On average, each viewpoint interacted with ten distinct genomic regions, of which 41% were annotated enhancers or promoters (Ghavi-Helm *et al.*, 2014). We observed, that 73% of the interactions span a distance of more than 50 kb to the viewpoint (Ghavi-Helm *et al.*, 2014). This shows that there are extensive long-range interactions throughout the compact *Drosophila* genome.

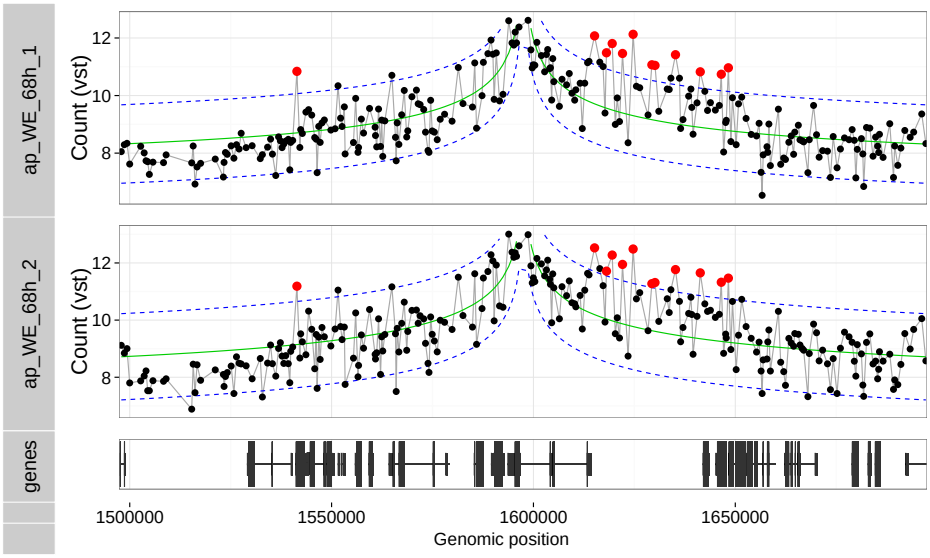


Figure 2.8: 4C signal of the *ap* CRM viewpoint on the variance stabilized scale. The green line shows the fitted values and the dashed blue lines show the interval of $\pm 3\sigma$ of the fit residuals. Interacting fragments are highlighted by red dots.

2.2.1.3 Detecting changes between conditions

To detect differences between conditions, I used the method described in Section 2.1.4. For the following discussion I will focus on the changes between conditions for the *ap* CRM viewpoint.

To illustrate the effect of the distance dependent normalization factors, I first calculated the differences between conditions using a normalization that only accounted for the different library sizes. The resulting MA plot is shown in Figure 2.9. It shows the log2-fold change between mesoderm tissue and whole-embryo for *Drosophila* embryos 6-8 h after fertilization as a function of the logarithmic transformed mean value. The distribution of log2-fold changes is skewed towards the mesoderm condition for high count values.

This influence of the distance dependence in the different libraries was captured by the normalization factors (see Section 2.1.4). In Figure 2.10 the values are distributed more symmetrically after normalization for the distance dependence.

An overview plot showing the 4C signal on the variance stabilized scale as well as the calculated log2-fold changes between the conditions is shown in Figure 2.11. Fragments that have an adjusted p-value of less than 0.01 for the Wald test statistic are highlighted by blue points, or orange points if they additionally are called as an interaction.

For the strong interaction at the *ap* promoter my method estimated a fold change of 2.25 between the conditions. Stronger contacts in the mesoderm tissue could be due to the fact that the *ap* gene is only expressed in the mesoderm (Ghavi-Helm *et al.*, 2014).

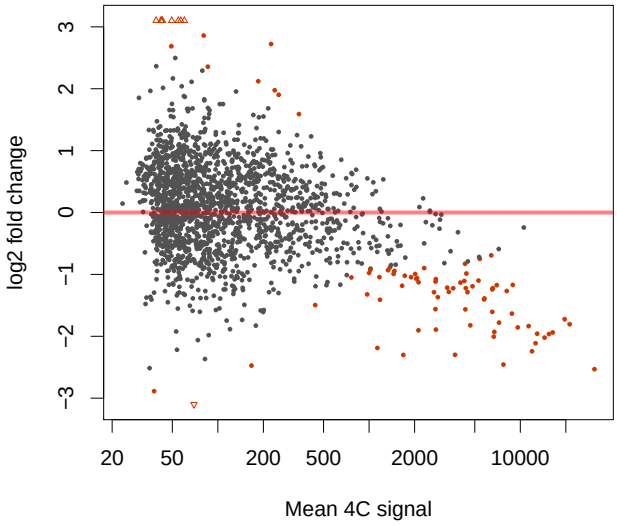


Figure 2.9: MA plot of the comparison between two conditions for the *ap* CRM viewpoint from our data set (Ghavi-Helm *et al.*, 2014) without normalizing for the distance dependence.The y axis shows the difference between log interaction counts for a given fragment plotted against the average log interaction per fragment on the x-axis. Red dots represent fragments that show differential interactions (p-adjusted < 0.01, Wald test)

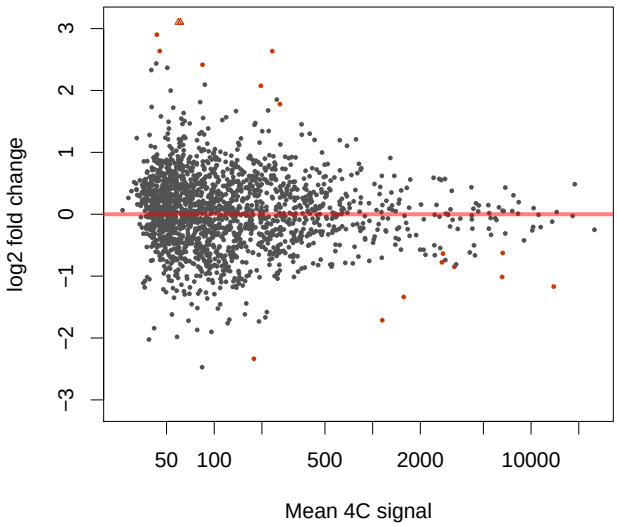


Figure 2.10: MA plot of the comparison between two conditions for the *ap* CRM viewpoint from our data set (Ghavi-Helm *et al.*, 2014) normalized for the distance dependence. The y axis shows the difference between log interaction counts for a given fragment plotted against the average log interaction per fragment on the x-axis. Red dots represent fragments that show differential interactions (p-adjusted < 0.01, Wald test)

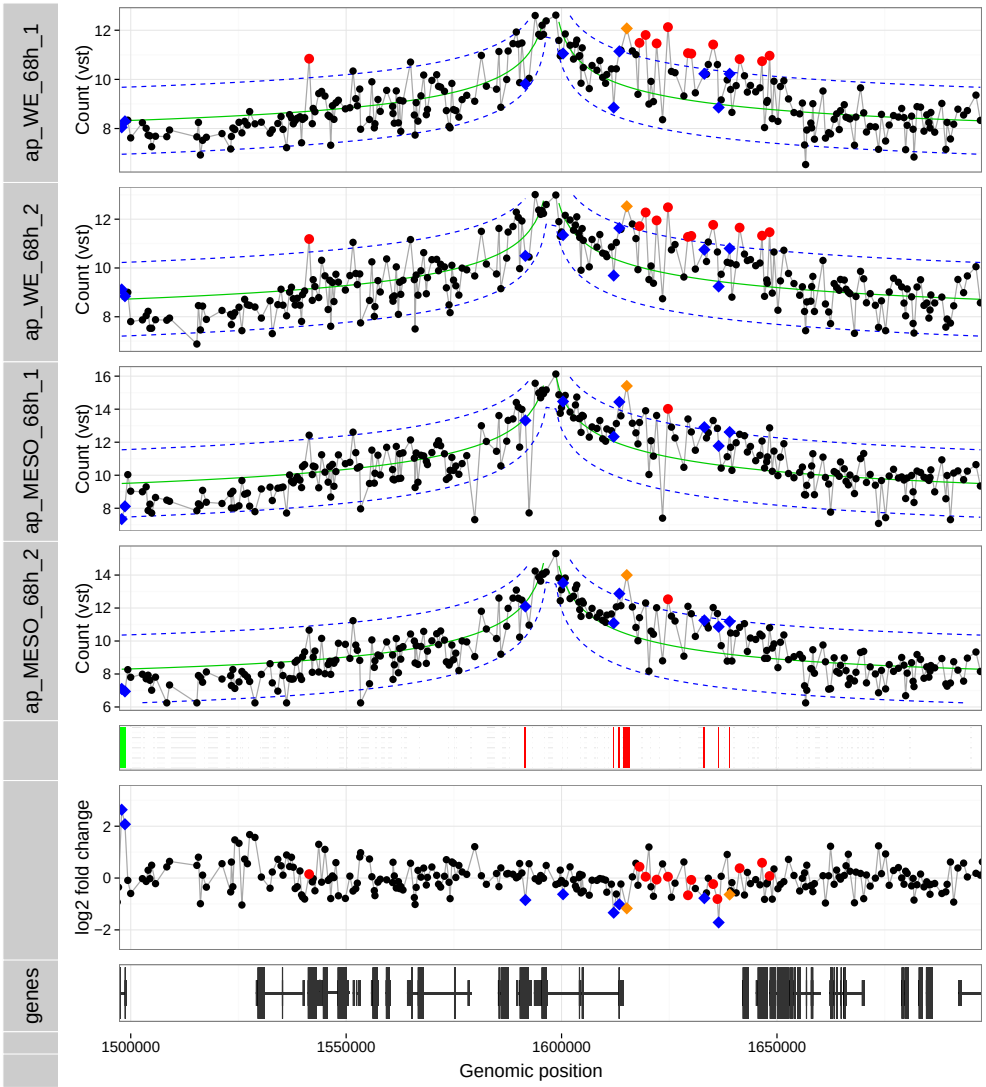


Figure 2.11: Detection of interactions and differences for the *ap* CRM viewpoint: The upper four, wide tracks show the variance stabilized counts for 2 biological replicates in two different conditions. The fit is shown as green solid line and the dashed blue lines represent the fit $\pm$ 3σ. Interactions detected by z-score > 3 in both replicates and p-adjusted < 0.01 for one replicate are shown as red or orange points per condition. Fragments represented by orange points additionally show differential interactions (p-adjusted < 0.01, Wald test). Differential changes in the contact profile that are not called as interactions are shown as blue points (p-adjusted < 0.01, differential Wald test). The color bar in the middle shows whether the upper condition (green) or the lower condition (red) has a higher signal for the detected differences (p-adjusted < 0.01, Wald test). The calculated log2 fold-change of the differential testing per fragment are shown as a track. The track at the bottom shows the gene model for the region.

2.2.1.4 Changes of chromatin 3D structure during embryogenesis

In general, we observed that the effect sizes for differential changes are very small and the overall pattern of the interaction profiles remains largely unchanged (Ghavi-Helm *et al.*, 2014). This was also the case for enhancer-promoter interactions of genes that switched from on to off or vice versa between the two investigated developmental time points at 3-4 h and 6-8 h after fertilization (Ghavi-Helm *et al.*, 2014). For example, this was the case for the *ap* CRM viewpoint already mentioned in the previous section. The *ap* gene is not expressed in the 3-4 h condition, but it is highly expressed at 6-8 h as shown in Figure 2.12 (Ghavi-Helm *et al.*, 2014). The interaction between the viewpoint and the *ap* promoter in the whole-embryo tissue is observed at both time points and does not significantly change between the 3-4h and 6-8h time point (Figure 2.12). The estimated log2-fold change is -0.175 with a standard error of 0.117. Although the gene is not expressed at 2-4 h, my collaborators observed a polII signal at the *ap* promoter (Ghavi-Helm *et al.*, 2014) and paused polII defined by genomic run-on sequencing (GRO-seq) (Saunders *et al.*, 2013).

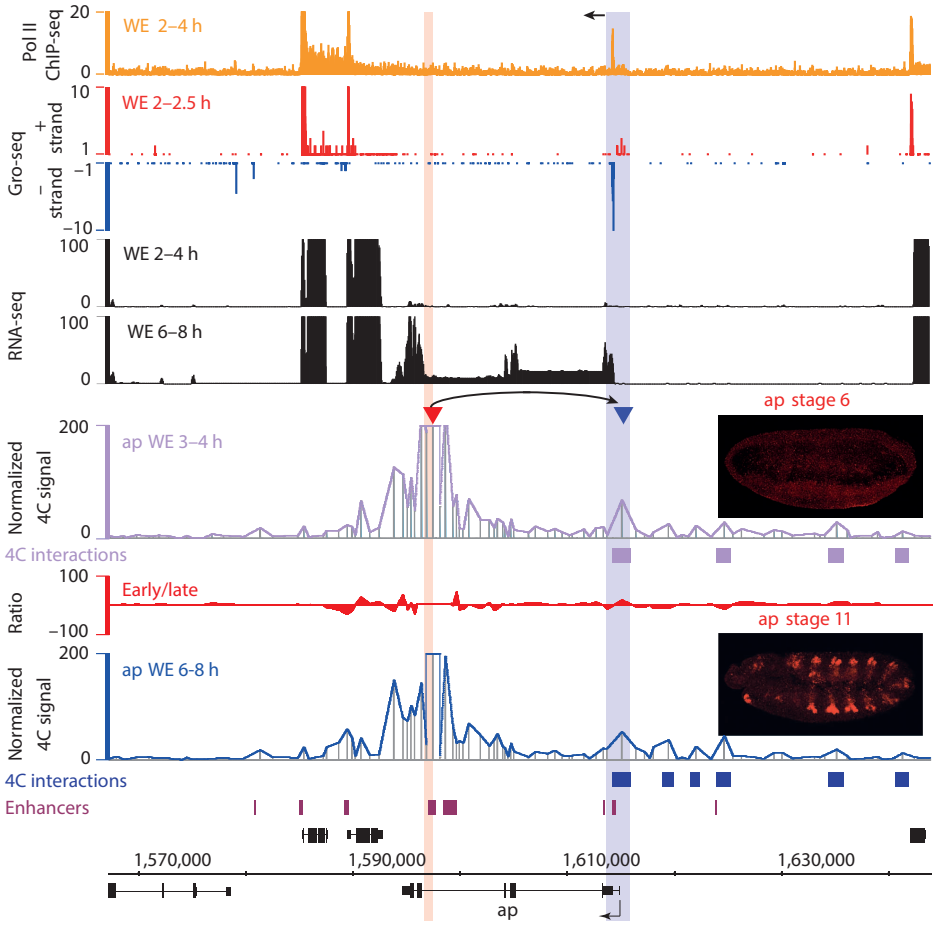


Figure 2.12: Normalized 4C interaction profile of the *ap* locus at 3-4 and 6-8 h in whole-embryo tissue. The tracks are listed top to bottom. polII CHiP-Seq signal from 2-4 h embryos (reads per genomic content (RPGC)), GRO-seq signal from 2-2.5 h embryos (plus strand in red, minus strand in blue), RNA-Seq signal (reads per kilobase per million mapped reads (RPKM)) from 2-4 and 6-8 h embryos, 4C interaction profile (back transformed to count scale and normalized) from 3-4 h (mauve) and 6-8 h (blue) embryos, the viewpoint is indicated by the red triangle. Differential 4C signal is shown in red. Detected 4C interactions and known enhancers are indicated at the bottom. Taken from Ghavi-Helm *et al.* (2014).

2.2.2 Long-range chromatin interactions within the *Shh* locus

The expression of *Shh* is important for proper limb development in the developing mouse embryo (Riddle *et al.*, 1993). It is expressed in the Zone of Polarising Activity (ZPA), which resides at the posterior part of the developing limb (Riddle *et al.*, 1993). *Shh* is activated by the conserved ZPA Regulatory Sequence (ZRS), a very distant limb-specific enhancer (Lettice *et al.*, 2003). A schematic of the genomic locus around the *Shh* gene is shown in Figure 2.13.

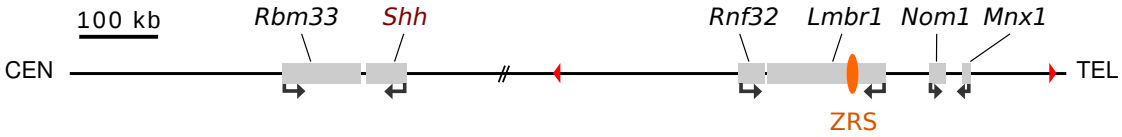


Figure 2.13: Schematic representation of the *Shh* locus. The *Shh* gene and other genes in the region are represented as gray boxes. The ZRS is highlighted as orange ellipse.

Because animals with limb malformations are viable and limb development is sensitive to quantitative and qualitative gene expression changes, the limb formation in mice embryos is a good model system to study parameters which can influence the efficiency of promoter-enhancer interactions. We therefore used the limb bud of the developing mouse embryo to study the effects of genomic rearrangements on gene regulation.

To get an overview of the regulatory landscape of *Shh* locus, my collaborators used several mouse lines that contain individual regulatory sensors (GROMIT system, Ruf *et al.* (2011)) in the genomic region around the *Shh* gene. The GROMIT system is based on a sleeping beauty transposon (SB transposon) that contains a regulatory sensor in form of a *LacZ* reporter gene under the control of the minimal promoter of the human β-globin gene and a *loxP* site. The mouse lines were generated by making use of the local hopping property of the SB transposon system. This allowed a systematic analysis of tissue specific regulatory input captured in the *Shh* locus.

2.2.2.1 Interaction of *Shh* and ZRS

The interaction of *Shh* and the ZRS has been shown in limb tissue using 3C and 3D-FISH (Amano *et al.*, 2009).

Published Hi-C data from mouse ESCs provides a coarse view of the chromatin structure for the whole region and is shown in Figure 2.14. There is a large topological domain or topologically associating domain (TAD), which represents a domain of preferred chromatin interactions identified by HiC data (Dixon *et al.*, 2012), with *Shh* at the left and the *Lmbr1* promoter at right boundary. On each side the TAD is flanked by another TAD. In this Figure the inserted reporter genes

are highlighted. They show activity throughout a region that overlaps with the *Shh* topological domain observed in the Hi-C data (Symmons *et al.*, in preperation).

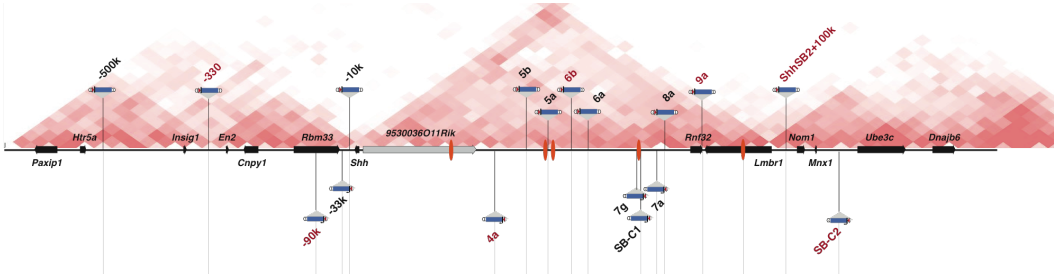


Figure 2.14: Hi-C map for the *Shh* locus generated from published mouse ESC data (Dixon *et al.*, 2012). The inserted regulatory sensors (blue and labeled constructs) are highlighted. Courtesy of Orsolya Symmons.

My collaborators generated 4C sequencing data for 5 viewpoints spread throughout the *Shh* locus using a 4bp-cutter to investigate the chromatin structure at higher resolution. The first restriction enzyme used in the experiment was *NlaIII* and the second restriction enzyme was *DpnII*. I assigned the aligned reads to a reference of *NlaIII* fragments as described in Sections 2.1.2.2 and 2.1.2.4. For all libraries the percentage of aligned reads that could be assigned to a fragment was above 90 %.

Unfortunately, the amount of starting material was very limited in the case of mouse embryo limbs. Because of this I observed spikes in the 4C signal for fragments that were not reproducible between replicates. This was most likely due to sampling effects in the PCR amplification cycles of the protocol. In order to prevent the influence of PCR artifacts on the analysis, I transformed the raw 4C signal into hit fractions as described in Section 2.1.6.2. As bin width I used 51 fragments, and as thresholds 10 and 100 counts per fragment. The resulting 4C hit fraction profiles are shown in Figure 2.15 for the 5 viewpoints investigated in the *Shh* locus. In order to have consistent hit fraction values for the different replicates of a viewpoint, I sub-sampled the aligned reads in each library to match the number of reads obtained for the smallest library. This is no optimal strategy, however the loss of data could be tolerated because the largest fold-change between libraries was 3.25 and for all viewpoints the smallest libraries contained a sufficient number of reads.

The *Rbm33* viewpoint had an 4C hit fraction interaction profile that is asymmetric. It extended primarily towards the centromere and the signal over the *Shh* TAD was reduced. The *Shh* viewpoint showed a hit fraction profile that extended over the whole *Shh* TAD and showed interactions in the region of the ZRS. It also slightly extended into the neighboring TAD on the centromeric site. The 4C hit fraction profile of the *Rnf32* and ZRS viewpoint were very similar. They both showed a broad extension over the *Shh* TAD with a peak at the position of the

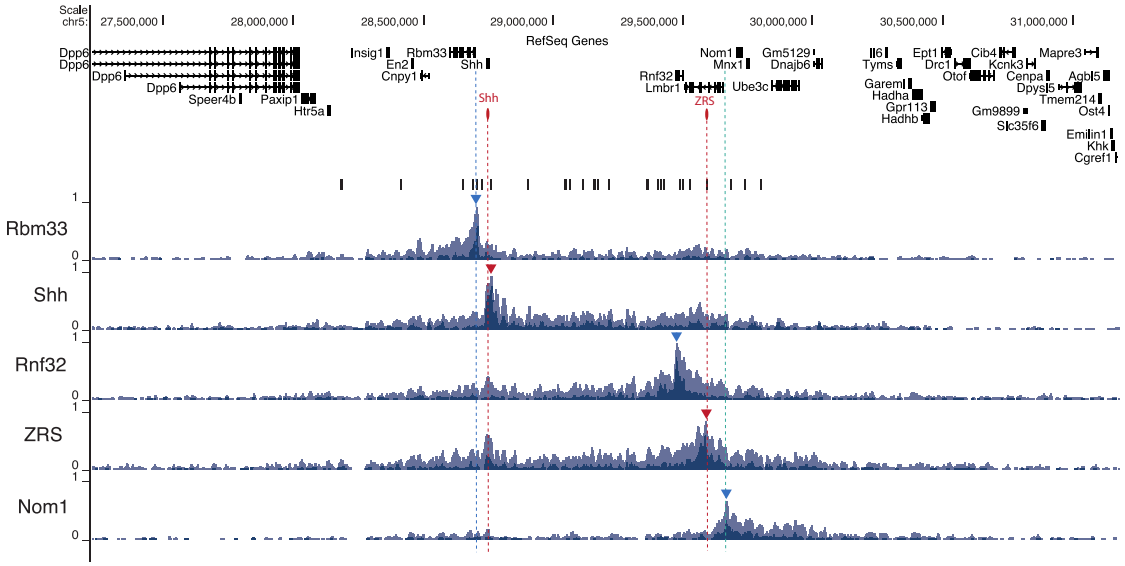


Figure 2.15: 4C hit fraction profiles of the viewpoints in the *Shh* locus. The legend at the top shows the gene model of the genomic region. The small bars below represent the positions of the regulatory sensor insertions. The *Shh* promoter and ZRS enhancer are highlighted. Small triangles mark the position of the viewpoints. The genes marked by the blue dashed vertical lines are outside the *Shh* TAD. Taken from Symmons *et al.* (in preperation).

Shh promoter. This peak was especially pronounced for the ZRS viewpoint. Both also extended into the neighboring TADs on the centromeric and telomeric side. For the *Nom1* viewpoint an asymmetric interaction profile was observed. It was strongly reduced over the *Shh* TAD and extended towards the telomere.

The profiles therefore recapitulated the TAD structure observed in the Hi-C data shown in Figure 2.14.

2.2.2.2 Influences of genomic inversion in the *Shh* locus

In order to investigate the influence of genomic rearrangements on the regulation of *Shh* by the ZRS, my collaborators generated mice with a genomic inversion in the region of the ZRS. For this they made use of the *loxP* site that is inserted with the transposon and used CRE-mediated recombination (Hérault *et al.*, 1998; Spitz *et al.*, 2005). This resulted in alleles carrying either large deletions, duplications or inversions in the region of interest. A schematic of the investigated Inv inversion is shown in Figure 2.16.

As a result of this inversion, the ZRS is moved approximately 160 kb closer to the *Shh* gene. The regulatory sensor which moved along with the inversion still captured *LacZ* expression in the posterior limb, suggesting that the ZRS was still active (Symmons *et al.*, in preperation). However, the ZRS was not able to activate *Shh* any more and *Shh* expression was lost in the Inv mouse embryos (Symmons

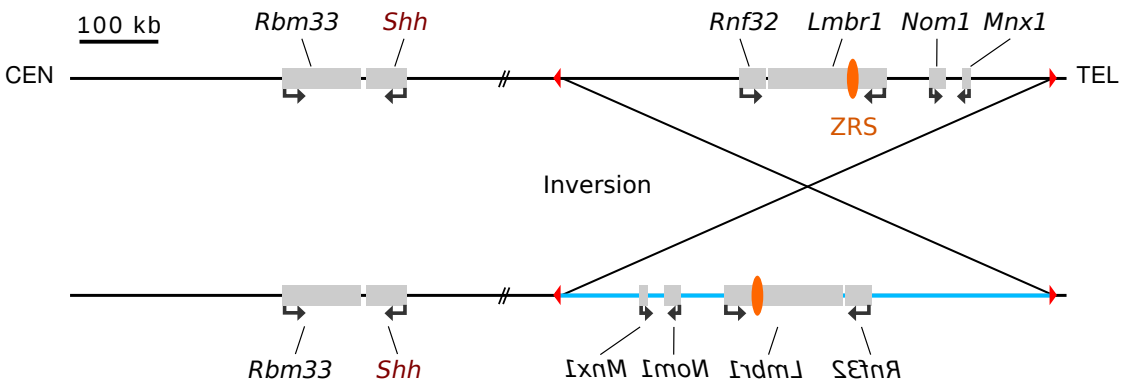


Figure 2.16: Schematic representation of the Inv inversion in *Shh* locus. The inverted region is marked by a blue line. The *Shh* gene and other genes in the region are represented as gray boxes. The ZRS is highlighted as orange ellipse.

et al., in preperation). The lost expression of *Shh* lead to truncated limbs in the Inv animals (Symmons *et al.*, in preperation).

To test how the inversion affects the chromatin structure, my collaborators performed 4C experiments from limbs of Inv mouse embryos. For visualizing the 4C signals of the inversion, I corrected the genomic coordinates of the inversion for these libraries as described in Section 2.1.2.5. The obtained 4C hit fraction profiles are shown in Figure 2.17 together with the wild type profiles.

Although the ZRS was closer to the *Shh* promoter, the contacts of the ZRS and *Shh* promoter were reduced in Inv samples. The 4C hit fraction profile of the of the ZRS is increased in the telomeric direction compared to the wild type (WT) samples. The same was true for the *Shh* viewpoint. The interactions with the region of the ZRS were lost. Instead, three new interacting regions were observed that fall in the region of the three gene promoters of *Mnx1*, *Nom1* and *Lmbr1*. The hit fraction profile of the *Rbm33* viewpoint did not change dramatically and was still asymmetric towards the centromeric direction. Only the interactions in the region around the ZRS are slightly reduced. For the *Nom1* viewpoint the direction of the 4C hit fraction profile was inverted. The extension towards the telomere was lost and it was asymmetric in the direction of the centromere, with slightly increased interactions at the *Shh* promoter.

To further quantify the changes of the ZRS and *Nom1* viewpoint upon genomic inversion, I calculated the cumulative 4C signal as described in Section 2.1.5.2 using a 2 Mb window to each side of the viewpoint. The resulting distributions are shown in Figure 2.18.

For both viewpoints a shift in the primary direction of interaction was observed. For the ZRS viewpoint the primary contacts towards the *Shh* gene were flipped towards the telomere as a consequence of the genomic inversion. The primary direction of interaction changed in the opposite way for the *Nom1* viewpoint. In the wild-type case the viewpoint primarily interacted towards the telomere, whereas

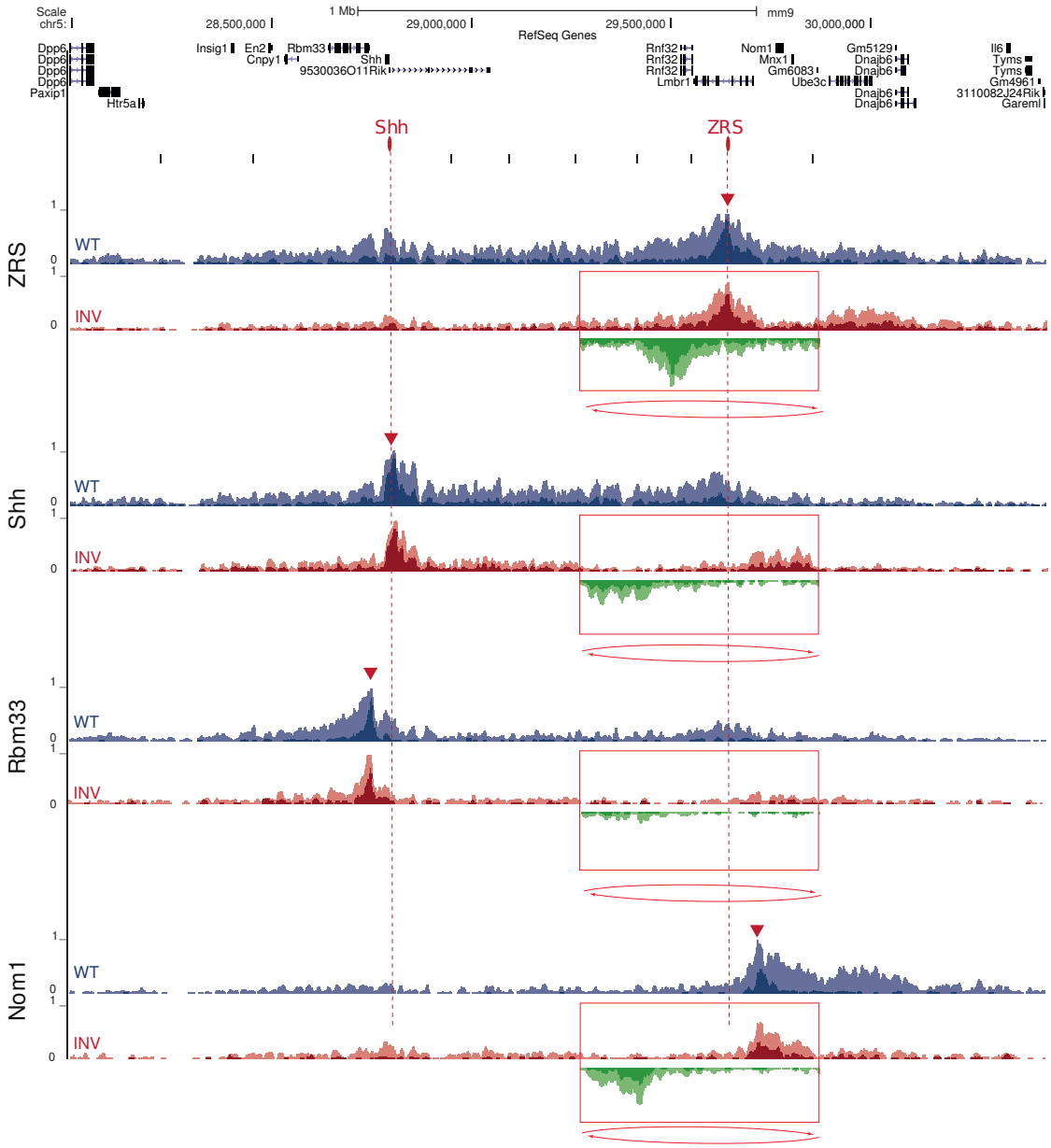


Figure 2.17: 4C hit fraction profiles of the viewpoints in the *Shh* locus with inversion. Wild-type profiles are shown in blue and Inv profiles are shown in red for the reference genome. The green Inv profile corresponds to the Inv allele after *in-silico* inversion. The inverted region is highlighted by a box. The legend at the top shows the wild-type gene model of the genomic region. The small bars below represent the positions of the regulatory sensor insertions. The *Shh* promoter and ZRS enhancer are highlighted. Small triangles mark the position of the viewpoints. Taken from Symmons *et al.* (in preperation).

its interactions were primarily directed towards the *Shh* gene and centromere in the Inv configuration.

In summary, the results indicated that the inversion moved the ZRS to a position where it could not contact *Shh* anymore, which was consistent with the limb phenotype and loss of *Shh* expression in the ZPA. *Shh* instead contacted the three genes *Mnx1*, *Nom1* and *Lmbr1*. The broad interactions, especially between ZRS and *Shh*, that previously defined the TAD structure were strongly reduced in the inversion genotype, even for the regions that were not directly affected by the inversion. This suggests that the compact wild-type TAD structure in general collapsed to a less compact chromatin structure.

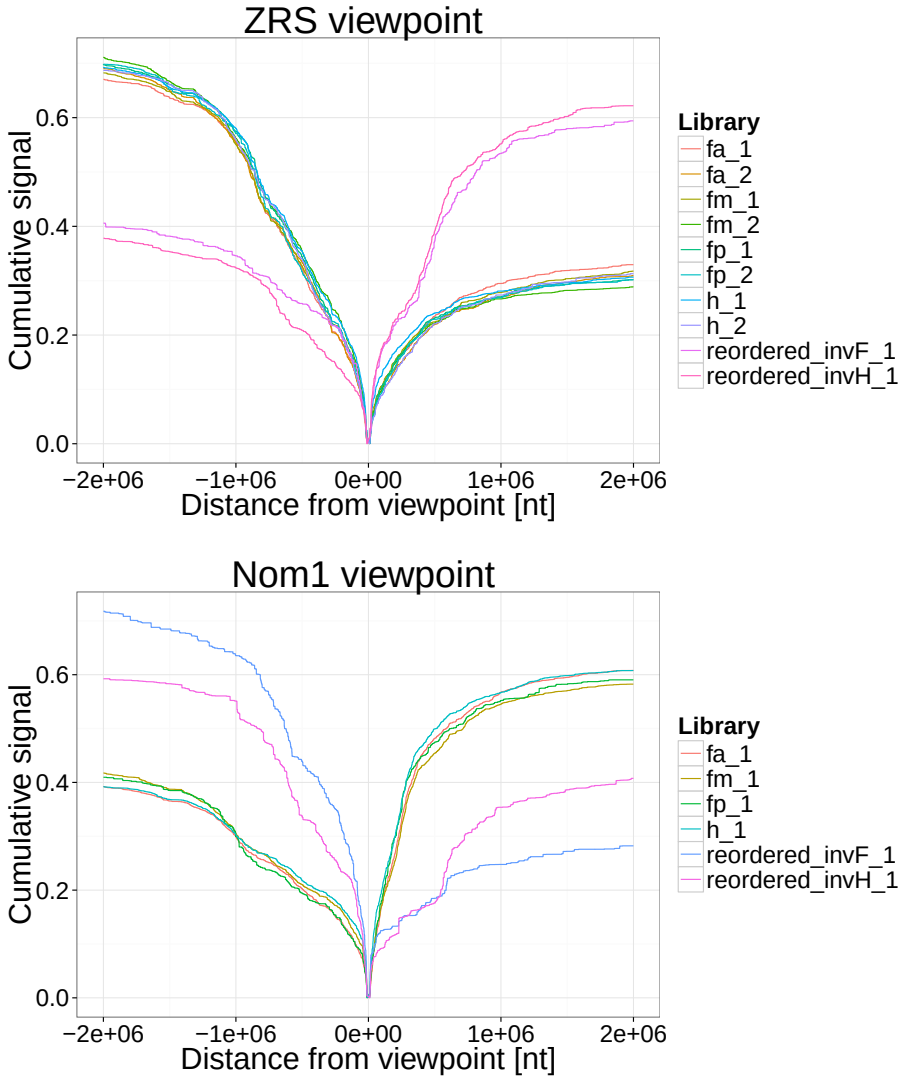


Figure 2.18: 4C cumulative signal of the ZRS and *Nom1* viewpoint in a 2 Mb window to each side of the viewpoint.

2.2.3 The two domain structure of the *Ap2-γ - Bmp7* locus

The *Ap2-γ- Bmp7* locus spans a genomic region of approximately 0.5 Mb. It contains two different genes, *Ap2-γ* and *Bmp7*, which are expressed during mouse development and show overlapping expression patterns for some tissues (Chazaud *et al.*, 1996; Danesh *et al.*, 2009), but also unique expression patterns in others. The locus therefore constitutes a nice model to study the specificity of gene regulation and expression in different tissues.

2.2.3.1 Regulatory landscape of the *Ap2-γ - Bmp7* locus

To investigate the regulatory landscape of the *Ap2-γ - Bmp7* locus, my collaborators employed the GROMIT system (Ruf *et al.*, 2011) of regulatory sensors described already in Section 2.2.2. The positions of the regulatory sensor insertions is shown schematically in Figure 2.19.

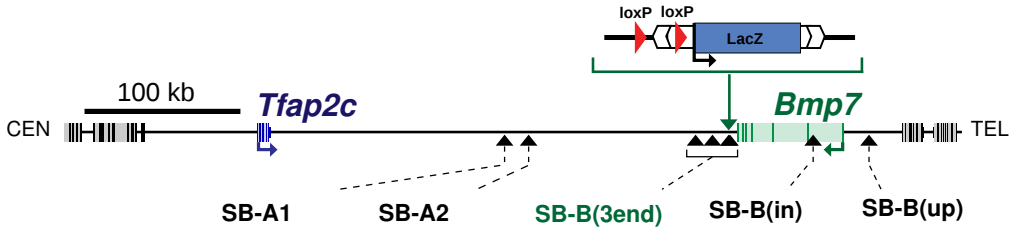


Figure 2.19: Schematic representation of the *Ap2-γ - Bmp7* locus. The positions of the regulatory sensors insertion sites are shown as black triangles. Taken from Tsujimura *et al.* (submitted for publication).

Scanning the interval with the regulatory sensor defined two distinct, non-overlapping regulatory domains. The first, with SB-A1/A2 overlaps with the expression patterns of *Ap2-γ* (Tsujimura *et al.*, submitted for publication). The second, from SB-3end to SB-5up, shared several, but not all, expression specificities with *Bmp7* (Tsujimura *et al.*, submitted for publication; Helder *et al.*, 1995; Adams *et al.*, 2007).

A combination of chromatin profiling for enhancer-associated marks (H3K27Ac and EP300 binding) in forebrain and heart tissue, as well as direct transgenic assays, identified two enhancers in the intergenic interval that drove expression in the forebrain (enhancer FB1) and in the heart (enhancer mm75), respectively (Tsujimura *et al.*, submitted for publication; Visel *et al.*, 2007). mm75 was annotated according to the VISTA Enhancer Browser (Visel *et al.*, 2007). These enhancers could be assigned to the two separated regulatory domains.

As a next step, we wanted to test whether this separation is reflected in the chromatin structure of the locus. A coarse view on the chromatin 3D structure of this locus could be generated from published Hi-C data in mouse ESCs (Dixon *et al.*, 2012). It is shown in Figure 2.20.

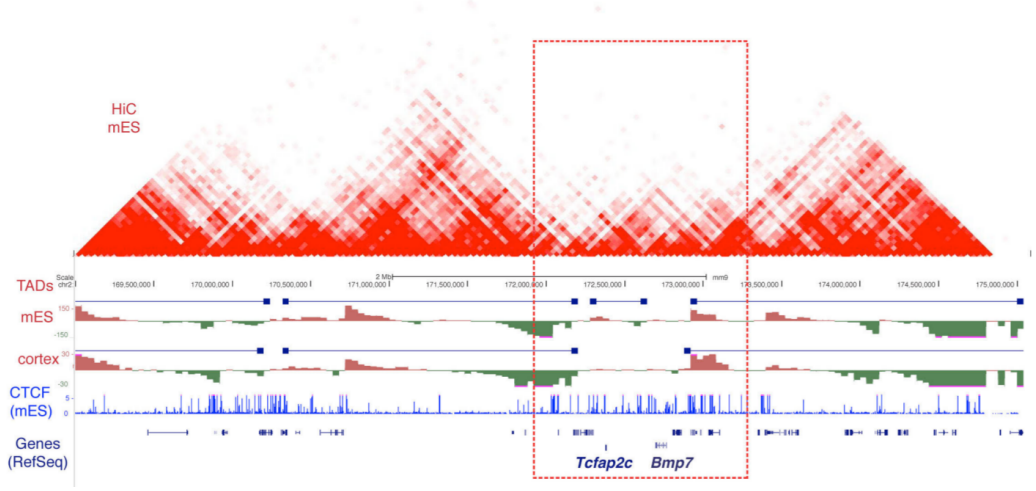


Figure 2.20: Hi-C map for the *Ap2-γ* - *Bmp7* locus generated from published mouse ESC data (Dixon *et al.*, 2012). Taken from Tsujimura *et al.* (submitted for publication)

While there are prominent TADs to each side of the *Ap2-γ* - *Bmp7* locus, for the TAD structure of the locus itself it is not clear, whether there is one larger TAD or two small TADs. This might be due to the limited resolution of 10 kb that cannot resolve fine structures.

In order to investigate the locus at a higher resolution, my collaborators generated 4C sequencing data with viewpoints spread throughout this locus. The *Ap2-γ* and *Bmp7* promoter were chosen as viewpoints and two additional viewpoints located in the region between the two genes. In this experiment the *NlaIII* restriction enzyme was used as the first cutter and *DpnII* as the second. Using their cutting sequences the reference genome was cut *in-silico* to generate the corresponding fragment reference as described in Section 2.1.2.2. To generate a fragment count table, I assigned the aligned reads to this fragment reference, as described in Section 2.1.2.4. The percentage of aligned reads that could be assigned to a fragment was above 90 % for all libraries. I corrected the genomic distance to the viewpoint for modified genomic regions in libraries generated from mice carrying large genomic rearrangements as described in Section 2.1.2.5.

To normalize for the different library sizes, I used the RPM normalization method for the counts on the viewpoint chromosome (chr2). For better visualization, I smoothed the count values with a smoothing window of 11 fragments, assigning the average within the window to the middle fragment. The resulting 4C profiles are shown in Figure 2.21

The 4C interaction profiles observed in the different tissues were very similar for each of the viewpoints. This suggested, that the overall chromatin structure did not change dramatically between different tissues for this locus. Only for the *Ap2-γ* viewpoint in the limb bud the 4C signal in the region of the known brain enhancer seemed to be reduced.

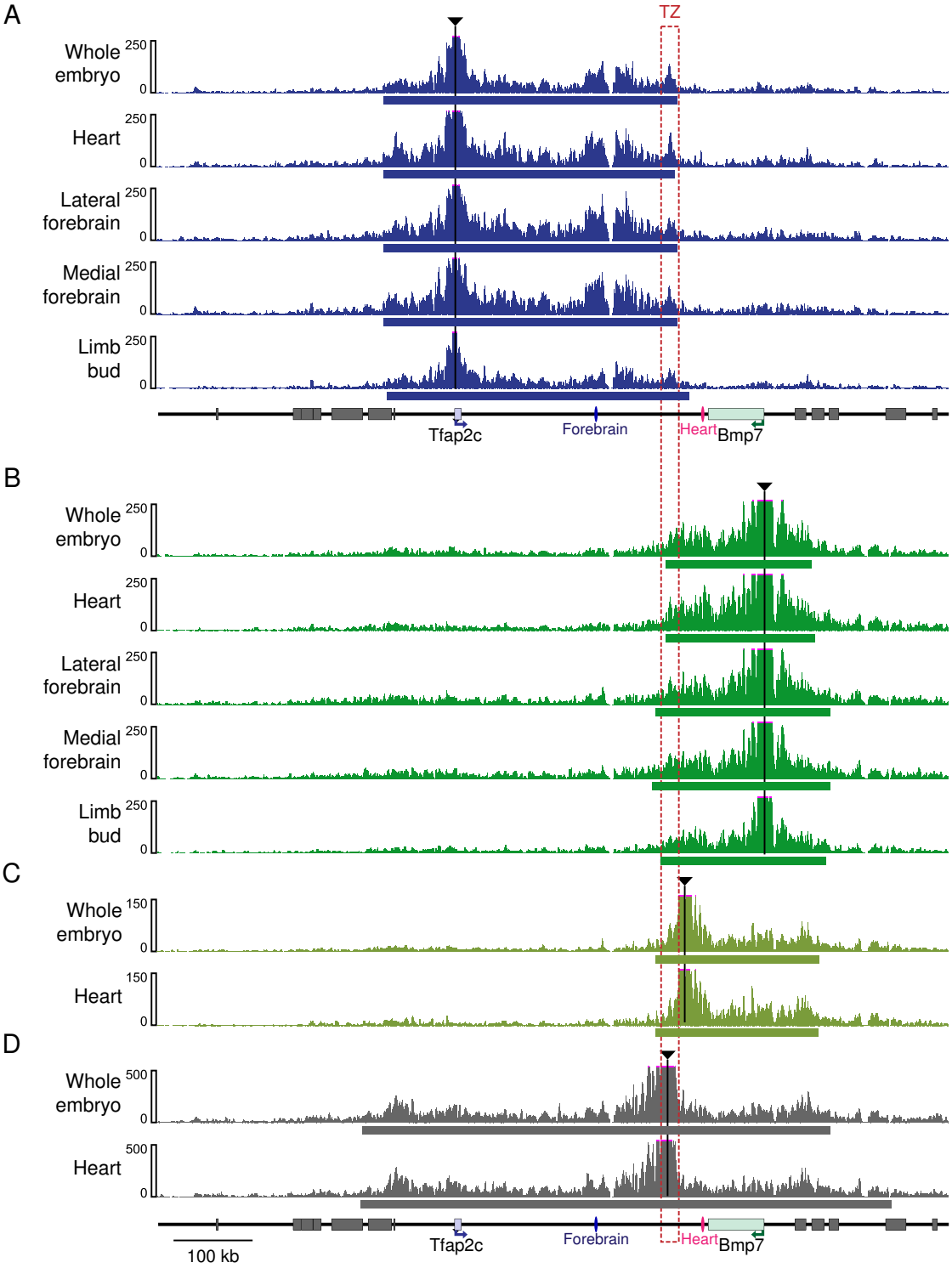


Figure 2.21: 4C profiles of the viewpoints at the *Ap2-γ* promoter (A, blue), the *Bmp7* promoter (B, green), and two viewpoints in between the two genes (C, light green and D, grey) in different tissues. The legend at the bottom shows the gene model of the genomic region. The known forebrain and heart enhancer are highlighted. The bar at the bottom of each profile the shows the estimated primary interaction domain. The transition zone (TZ) of the two regulatory domains is highlighted by dashed lines. Taken from Tsujimura *et al.* (submitted for publication)

The *Ap2-γ* viewpoints had asymmetric interaction profiles. They only extended approximately 100 kb towards the centromere, whereas in the direction of the telomere they extended approximately 300 kb followed by a abrupt decrease of the signal (Figure 2.21 A).

For the *Bmp7* viewpoints the 4C signal was more symmetric. The signals were strong over the *Bmp7* gene body and extend towards the centromere until a sudden decrease (Figure 2.21 B).

The profiles in between the two genes were dramatically different. The viewpoint that was closer to the *Bmp7* gene showed strong asymmetry in the 4C signal with no centromeric extension, but broadly extended over the *Bmp7* gene (Figure 2.21 C). The other viewpoint on the other hand showed symmetric contacts towards both genes (Figure 2.21 D).

In general, the 4C profiles recapitulated the two regulatory domains that were observed for the expression profiles. In order to estimate the primary interaction domains of the viewpoints, I used the segmentation approach described in Section 2.1.5.1. An example for segmentation result of *Ap2-γ* is shown in Figure 2.22.

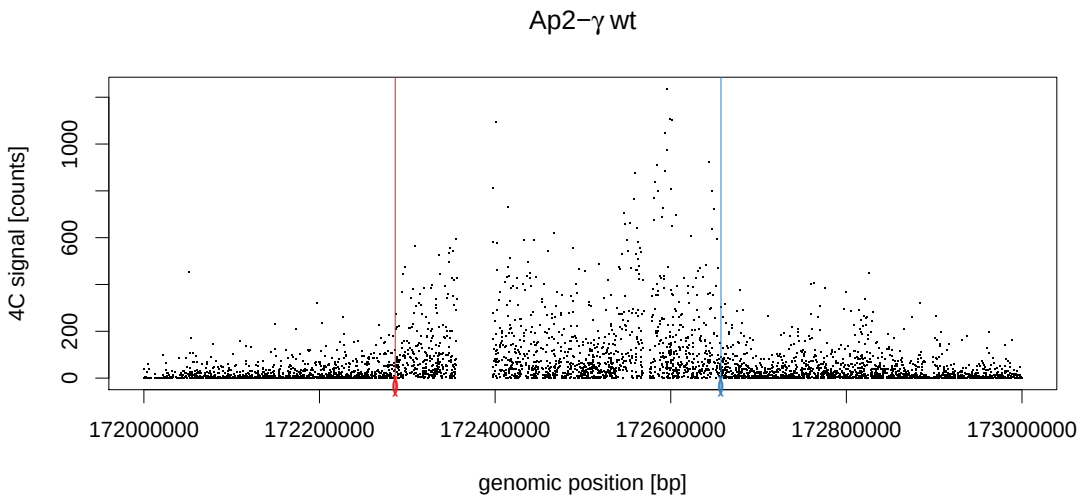


Figure 2.22: Segmentation of the 4C profile in whole-embryo of the *Ap2-γ* viewpoint into three segments using a piecewise constant function. The upper and lower breakpoints of the estimated piecewise function are shown with uncertainty estimates as blue and red bars.

Despite the main separation of the primary interaction domains from *Ap2-γ* and *Bmp7*, there seemed to be a small overlap (Figure 2.21). We defined this overlap as transition zone (TZ) between the two domains. The genomic coordinates of the TZ are approximately 172635000 $\pm$ 10000 to 172655000 $\pm$ 10000.

2.2.3.2 Influence of the transition zone on enhancer contacts and gene expression

In order to investigate how this TZ contributed to the regulatory and structural separation of the two gene domains, my collaborators generated additional alleles with deletions and inversions containing the TZ.

Using CRE-mediated recombination (Hérault *et al.*, 1998; Spitz *et al.*, 2005) between the *loxP* site of the SB-A1 insertion and a static *loxP* site at the end of *Bmp7*, my collaborators generated the del1 deletion. A schematic of this deletion is shown in Figure 2.23. The deleted region included the TZ and the known forebrain and heart enhancers. As a result of the deletion, *LacZ* expression was lost in heart and forebrain, however expression in the limb and jaw was contained (Tsujimura *et al.*, submitted for publication). RT-qPCR on del1 embryos showed that *Ap2-γ* expression is lost in the forebrain and *Bmp7* expression is lost in the heart (Tsujimura *et al.*, submitted for publication). This was in concordance with the loss of the known enhancers driving forebrain and heart expression of the respective gene.

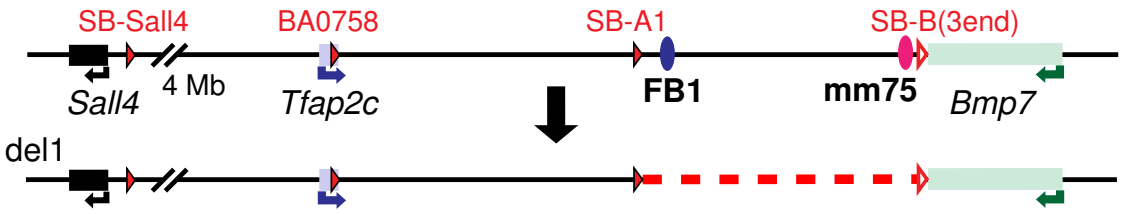


Figure 2.23: Schematic of the deletion del1 in the *Ap2-γ-Bmp7* locus. The deleted region contained the transition zone and annotated enhancers that drive forebrain and heart expression. Taken from Tsujimura *et al.* (submitted for publication).

The 4C signals obtained after the deletion of the TZ are shown in Figure 2.24. Upon deletion of the TZ we observed an extension of the chromatin contacts for both genes in the direction of the deleted region. The 4C signal of the *Ap2-γ* viewpoint extended towards the telomere, whereas the signal of the *Bmp7* viewpoint extended towards the centromere over the *Bmp7* gene body and other genes. For both viewpoints the profiles on the unaffected side remained highly similar to WT profiles. In this configuration the 4C profiles of *Ap2-γ* and *Bmp7* showed a large overlap. This overlap could be the result of merging the two domains into one larger domain by deletion of the boundary element.

To further assess how the observed TZ contributed to the regulation and structure of the *Ap2-γ* and *Bmp7* locus my collaborators generated additional alleles carrying inversions that contained the TZ.

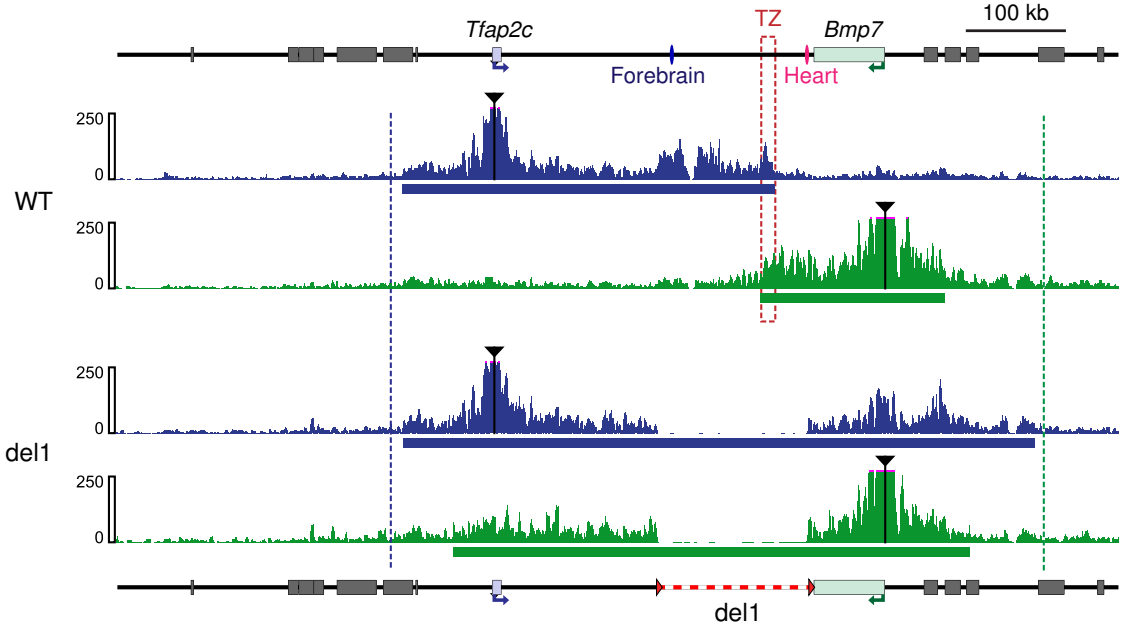


Figure 2.24: 4C signal of the *Ap2-γ* and *Bmp7* viewpoints after deletion of a region containing the TZ. The upper two tracks show the signal for WT embryos (*Ap2-γ* blue, *Bmp7* green). At the top the TZ and known enhancers are highlighted. The lower two tracks show the signal of embryos with the deletion. The deleted region is highlighted in the schematic at the bottom. The bar at the bottom of each profile shows the estimated primary interaction domain. Taken from Tsujimura *et al.* (submitted for publication).

Inv-M was a balanced inversion between the *loxP* sites of the SB-A2 and SB-B(3end) insertion. A schematic of this inversion is shown in Figure 2.25. The regulatory sensor stayed next to the heart enhancer and showed the corresponding expression (Tsujimura *et al.*, submitted for publication). In this configuration the mm75 heart enhancer was approximately equidistant from the two genes (187 kb versus 207 kb, compared to 80kb and 312kb in the wild-type allele).

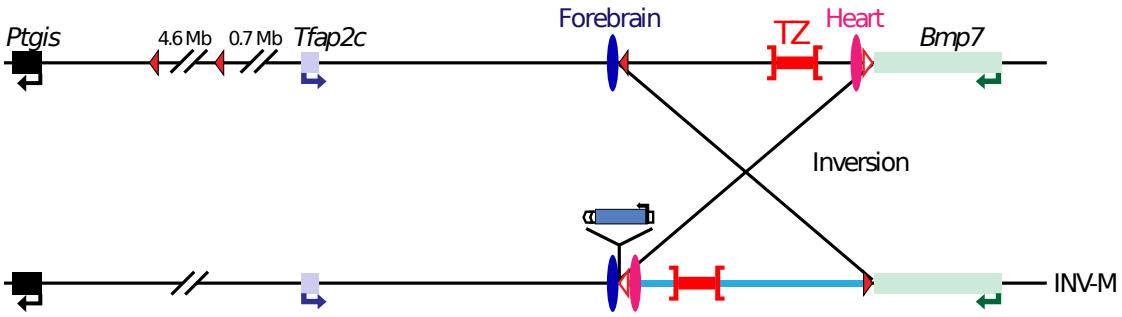


Figure 2.25: Schematic of the inversion Inv-M in the *Ap2-γ-Bmp7* locus. The inverted region contained the transition zone and the annotated mm75 enhancer that drove expression of *Bmp7* in the heart. Taken from Tsujimura *et al.* (submitted for publication).

Inv-L2 was a large balanced inversion between the *loxP* sites of the SB-L2 and SB-B(3end) insertion. A schematic of this inversion is shown in Figure 2.26. As for the Inv-M inversion, the regulatory sensor stayed next to the heart enhancer and showed the corresponding expression (Tsujimura *et al.*, submitted for publication). The relative distance of *Ap2-γ* to the FB1 and mm75 enhancers and the TZ did not change in this configuration, whereas the distance between the mm75 heart enhancer and *Bmp7* was increased to 1.1 Mb.

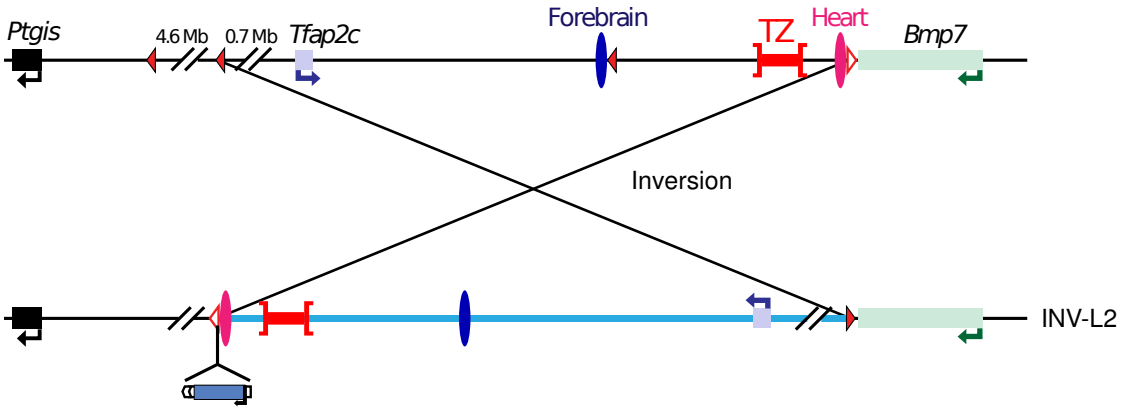


Figure 2.26: Schematic of the inversion Inv-L2 in the *Ap2-γ-Bmp7* locus. In the inverted region, the distance of *Ap2-γ* to the two enhancers and the transition zone was preserved. Adapted from Tsujimura *et al.* (submitted for publication).

To investigate how the inversion influenced the chromatin structure of the locus, my collaborators generated 4C profiles from Inv-M and Inv-L2 embryos. The 4C profiles of Inv-M (B) and Inv-L1 (C) are shown in Figure 2.27 along with the WT (A) profiles. Because the Inv-L1 inversion spanned several Megabases the profiles are shown at a large scale that allows one to investigate the global changes of the interaction profiles. The primary interaction domains are highlighted in the figure for the respective viewpoints. The change in the direction of the interactions for the viewpoint located between the mm75 enhancer and the TZ in Inv-M and Inv-L1 embryos is nicely visualized by the flip of the direction of the primary interaction domain compared to WT.

First I focused on the 4C profiles of the Inv-L2 embryos. The 4C profile of the *Ap2-γ* viewpoint in Inv-L2 embryos was flipped compared to WT, but in general resembled the WT profile. It extended until the TZ and only base line contacts were observed for the region of the mm75 heart enhancer across the TZ. In the direction away from the TZ the contacts only showed a short extension. This was consistent with the observation that *Ap2-γ* was expressed at WT levels in the heart (Tsujimura *et al.*, submitted for publication). This suggested, that the TZ located between the mm75 enhancer and *Ap2-γ*, prevented the enhancer from regulating *Ap2-γ* in this configuration, as in the WT case.

For the *Bmp7* viewpoint the relocation of the TZ lead to a large extension of the interaction domain towards the centromere in Inv-L2 embryos. The mm75 enhancer was moved over 1 Mb away from *Bmp7* and no chromatin contacts were established in this configuration. This was in agreement with the observation that *Bmp7* expression was strongly reduced in the heart of Inv-L2 embryos (Tsujimura *et al.*, submitted for publication).

In Inv-L2 embryos the chromatin interactions of the other two viewpoints showed a strong extension of more than 1 Mb in the direction of the centromere. A peak close to the end of the *Dok5* gene was observed for both viewpoints. In the telomeric direction the profiles resembled the WT profiles. The profile of the viewpoint between the mm75 heart enhancer and the TZ did not cross the TZ, whereas the profile of the viewpoint in the TZ extended over the *Ap2-γ* gene. In the Inv-L2 configuration *Dok5* was the closest gene to mm75 away from the TZ. However, my collaborators could not detect increased expression of *Dok5* in the heart (Tsujimura *et al.*, submitted for publication).

The interaction profiles of the four viewpoints in Inv-M embryos are shown in a smaller window of the locus in Figure 2.28. For direct comparison, the WT profiles are plotted on top and the Inv-M profiles below.

Ap2-γ (blue, A) showed strong contacts with the region of the FB1 and mm75 enhancer, with the latter being closer to the gene because of the inversion. In contrast to that, *Bmp7* (B, green) only showed baseline contacts with the new position of the heart enhancer. This agreed with the observed ectopic expression of *Ap2-γ* in the heart, which was driven by the mm75 enhancer in the Inv-M configuration (Tsujimura *et al.*, submitted for publication). Accordingly *Bmp7* expression was

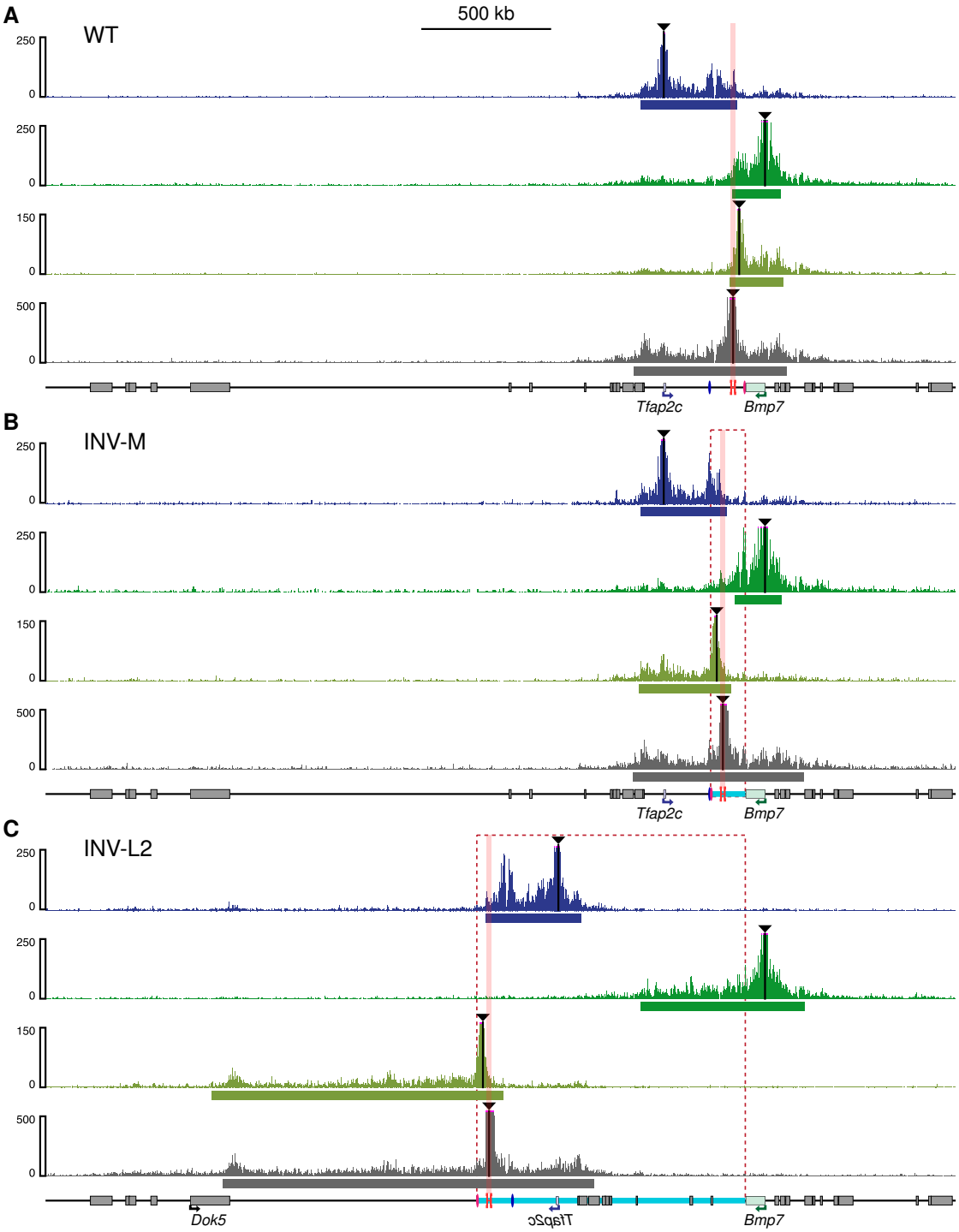


Figure 2.27: 4C signal of the four viewpoints in WT (A), Inv-M (B) and Inv-L2 (C) embryos. The 4C signal of *Ap2-γ* is colored in blue, *Bmp7* in green, viewpoint outside the TZ in light green and viewpoint inside the TZ in grey. The estimated primary interaction domain for each profile is shown as a bar at the bottom. The genomic coordinates are shown for the each allele at the bottom of the corresponding profiles. The TZ is highlighted by a red bar and the red dashed boxes represent the inverted region. Taken from Tsujimura *et al.* (submitted for publication).

dramatically reduced in the heart and it seemed likely that the regulation by the mm75 enhancer was lost, although the genes were approximately at an equivalent genomic distance to the heart enhancer (Tsujimura *et al.*, submitted for publication). These observations suggested that the relative position of the enhancer and genes in respect to the transition zone were important for chromatin interactions and regulatory action.

The viewpoint located between the mm75 enhancer and the TZ (C, light green) still had an asymmetric interaction profile, but it was mirrored at the TZ compared to WT. The interaction profile of the viewpoint in the TZ (D, grey) stayed symmetric and did not show strong changes.

To compare the changes in the read distribution for the two viewpoints in the TZ in a more quantitative way, I calculated counts in different windows. For this, I used RPM normalized data and excluded a 10 kb window around the viewpoint. The inverted region on chromosome 2 was used as window and two adjacent 400 kb windows to the centromeric and telomeric side. For each window I calculated the fraction between the number of counts within the window and the sum of counts in all 3 windows. The fractions are shown, as percentages, in Figure 2.28. The fractions reflect the observation that the interaction profile of the viewpoint between the mm75 enhancer and the TZ was mirrored at the TZ, keeping approximately the same ratios in the read distribution. For the viewpoint within the TZ the profile stayed symmetric and the numbers were nearly unchanged.

Taken together, in all configurations the position of the TZ was important for the distribution of the 4C contacts. Especially the asymmetric profile of the viewpoint located between the mm75 heart enhancer and the TZ in the direction opposite of the TZ was conserved.

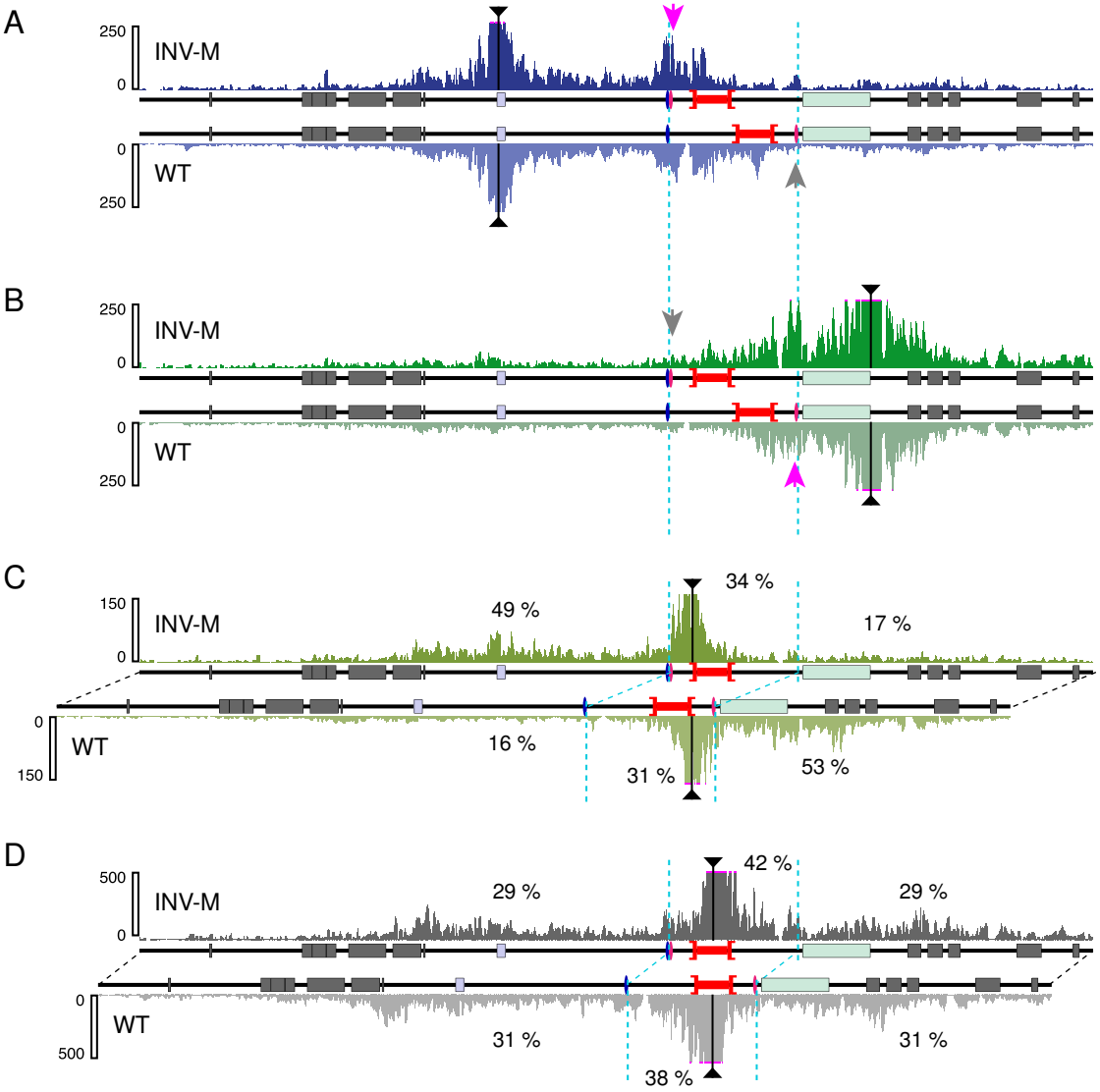


Figure 2.28: 4C signal of the four viewpoints in Inv-M embryos. The signal of WT embryos is shown as well. The genomic coordinates are shown for the Inv-M and WT case. 4C signal of *Ap2-γ* in blue (A), *Bmp7* in green (B), viewpoint outside the TZ in light green (C) and viewpoint inside the TZ in grey (D). For C and D the fraction of reads in the TZ and adjacent 400 kb windows is calculated. Taken from Tsujimura *et al.* (submitted for publication).

2.3 Discussion

2.3.1 Comparison of `FourCSeq` with other methods

The `FourCSeq` package, in combination with general sequence alignment software, contains the functionality to perform a full end-to-end analysis of 4C sequencing data (Figure 2.1). The included Python program can be used to demultiplex and trim the primer sequences from the FASTQ output of the sequencing machine. After alignment to the corresponding reference genome and generation of BAM-files, the statistical analysis is performed in **R**.

As a reference genome any FASTA file or Bioconductor `BS.genome` package can be used by the `FourCSeq` package . For example, in Section 2.2.1 I used the dm3 *Drosophila* reference genome and the Sections 2.2.2 and 2.2.3 the mm9 mouse reference genome. The `r3cseq` package (Thongjuea *et al.*, 2013) offers only the mm9, hg18 and hg19 genomes as reference genomes.

In general, my approach to detect peaks is similar to the method implemented in the `r3cseq` package (Thongjuea *et al.*, 2013). However, instead of using the data on a raw count scale, I use a variance stabilizing transformation on the count data for a more consistent statistical treatment. The approach of Thongjuea *et al.* (2013) on the count scale has less power to detect interactions at large distances from the viewpoint region, where the signal is orders of magnitudes smaller compared to the viewpoint region. To find specific interactions in the 4C signal, I fitted the decay of the 4C signal on the variance stabilized data as function of the genomic distance to the viewpoint. z-scores for each sample were calculated by dividing the fit residuals by the median absolute deviation (MAD) observed for all residuals of the corresponding sample. Specific interactions were defined by setting thresholds on the z-scores and associated p-values. This approach allowed me to detect interactions at short and larger distances from the viewpoint in a consistent way.

To compare the 4C profiles of different conditions, the `r3cseq` package simply uses the log2 fold-changes between conditions. In the `FourCSeq` package I made use of the framework for differential expression analysis implemented in the `DESeq2` package (Love *et al.*, 2014). With this approach the variability of the data was taken into account for the comparison between conditions. Differential interactions were only called if the observed fold change between conditions was significantly larger than expected, given the level of noise in the data. This allowed for a quantitative comparison that takes the noise of the data into account.

In contrast to `FourCSeq` and `r3cseq`, a customized approach for the alignment of sequencing reads to a restriction fragment reference is used in the method by van de Werken *et al.* (2012). This count data is further normalized and visualized by the provided tool. The results are presented as 4C profile plots and domainograms, generated by analyzing the data in different bin sizes. For the comparison of different experimental conditions this approach only provides a qualitative comparison of the interaction profiles and domainogram patterns.

To summarize, the `FourCSeq` package provides tools for the analysis of 4C sequencing data that are statistically sound, were not available before and should be useful for other researchers interested in analysing this type of experimental data.

2.3.2 Implementation of `FourCSeq`

The `FourCSeq` package is implemented within the Bioconductor framework of *GenomicRanges* (Lawrence *et al.*, 2013). This makes the package available across different computer platforms and easily accessible to many users.

It is easy to integrate called interactions and differences in interaction frequencies between different conditions with other genomic data using the Bioconductor framework. Furthermore the package allows to export interaction profiles and downstream results as bigWig or bedGraph files for visualization in a genome browser, in which additional tracks can be inspected in parallel. Flexible and automatable visualizations are also possible in R, as shown in many of the plots generated for this dissertation.

In summary, the `FourCSeq` package provides the tools to perform an end-to-end analysis of 4C sequencing data and to easily integrate the results with other genomic features. Its potential to be widely used makes it a promising tool that will help to better understand what role the chromatin 3D structure plays in transcription and other biological processes.

2.3.3 Long-range interactions and interaction changes in the developing *Drosophila* embryo

Using the `FourCSeq` package, I could detect specific chromatin interactions in the 4C sequencing profiles of developing *Drosophila* embryos. With the approach described in Section 2.1.3, ten known enhancer-promoter interactions could be confirmed (Ghavi-Helm *et al.*, 2014). Furthermore, the analysis revealed that there is quite extensive 3D connectivity throughout the compact *Drosophila* genome and complex interaction patterns between several enhancers and promoters were observed (Ghavi-Helm *et al.*, 2014). This is in agreement with recent observations from Hi-C experiments in human fibroblast cells (Jin *et al.*, 2013). We found several pairs of co-regulated genes that contacted common enhancers, where in some cases the enhancer-promoter interactions spanned distances greater than 200 kb (Ghavi-Helm *et al.*, 2014). Such long range chromatin interactions are known in mammalian genomes (Lettice *et al.*, 2003). In general, these observations suggest that the 3D organization of the gene-dense 180 Mb *Drosophila* genome into TADs (Sexton *et al.*, 2012) is similar to the organization of mammalian genomes (Dixon *et al.*, 2012).

Looking at the fold-changes between the different developmental time points and tissue contexts, calculated with the `FourCSeq` package, we observed that en-

hancer interaction profiles remain largely unchanged (Ghavi-Helm *et al.*, 2014). Focusing on genes that changed their expression status from off in 3-4 h to on in 6-8 h embryos, we found that chromatin interactions were already established and polII was recruited to the promoters in a poised state (Ghavi-Helm *et al.*, 2014). Such preformed chromatin structures have already been observed in human cells and they are proposed to allow rapid transcription activation upon the binding of specific transcription factors (Jin *et al.*, 2013). Such poised configurations seem very reasonable in the fast developing *Drosophila* embryo as they provide a scaffold for fast and precise spatio-temporal gene regulation by specific transcription factor patterns.

2.3.4 TAD organization defines long-range enhancer interaction specificity

With the analysis of the 4C sequencing data of the *Shh* and *Ap2-γ - Bmp7* locus, I showed, that in both wild-type cases the interactions between enhancer and target genes were defined by the 3D chromatin structure of the respective locus. The ZRS enhancer interacted with *Shh* in a large Mb size TAD with both elements being close to the TAD boundaries (Dixon *et al.*, 2012). The *Ap2-γ - Bmp7* locus clearly was separated by a small region which we called the transition zone (TZ). It is unclear, whether this separation corresponded to two small TADs in an unstructured region or to adjacent sub-TADs (Dixon *et al.*, 2012; Phillips-Cremins *et al.*, 2013). Moreover, the true biological situation might be more complex than these two simple idealizations that might only reflect the chromatin folding at a certain scale Mirny (2011).

The ZRS contacts *Shh* in the whole limb bud of developing mouse embryos but is only expressed in the ZPA (Amano *et al.*, 2009). In the ZPA, the ZRS and *Shh* loop out of their chromosome territory only in a few cells (Amano *et al.*, 2009). Likewise *Shh* is only expressed in a subset of cells in the ZPA (Amano *et al.*, 2009). The ZPA is defined by specific transcription factors, and point mutations within the ZRS lead to different binding patterns causing limb malformation and polydactyly (Lettice *et al.*, 2012). After disrupting the interactions of the ZRS and *Shh* by genomic inversion, we also observed a phenotype of limb malformation (Symmons *et al.*, in preperation). Although the ZRS was moved closer to the *Shh* promoter, *Shh* expression was lost as a consequence of the inversion, although the ZRS was still active since the regulatory sensor still captured enhancer activity. The observed redistribution of contacts for the 4C interaction profiles of the viewpoints throughout the locus showed that the long-range interaction between the ZRS and *Shh* was lost upon inversion in the locus. The preferred direction of interaction was inverted for the ZRS and *Nom1* viewpoint, which showed that the sequence between ZRS and *Nom1* gene was important for defining the chromatin structure of the locus. These observations correspond to a permissive model of enhancer-gene interaction where the enhancer contacts its target promoter within

a preconfigured chromatin interaction domain (de Laat and Duboule, 2013).

It was recently shown that the genomic distance between an enhancer and a promoter does not have a strong influence on whether the enhancer can interact with the promoter, but rather the chromatin domain configuration in which they reside (Symmons *et al.*, 2014). New data from an inversion that moved the ZRS even closer to *Shh* promoter showed a partial rescue of the limb malformation phenotype observed in the presented Inv inversion (Symmons *et al.*, in preperation). This suggests, that random collisions can occur at short genomic distances between separated TADs, while for large genomic distances the TAD organization favors the possible chromatin interactions within the TADs.

The observed separation of the *Ap2-γ* - *Bmp7* locus into two regulatory domains also seemed to be defined the TZ. This TZ was important for the specificity of allocating the correct enhancer to the respective target gene, by defining two distinct domains of gene regulation (Tsujimura *et al.*, submitted for publication).

The 4C profiles of different tissues showed high similarities independent of whether the gene was expressed in the respective tissue or not. This indicated that the locus formed a stable 3D configuration which remained largely unchanged. Gene activity and the enhancer contacts were regulated within this predefined domains as recently reported for other loci and organisms (Montavon *et al.*, 2011; Jin *et al.*, 2013; Ghavi-Helm *et al.*, 2014).

The extension of the interaction profiles upon deletion of the TZ region shows that the TZ is important for the separation of the two domains. The strong overlap of the *Ap2-γ* and *Bmp7* interaction profiles indicate, that the two domains are fused into one larger domain. This fusion of neighboring TADs, upon deletion of the boundary element between them, has been observed for deletions in the *Xist* loci (Nora *et al.*, 2012).

The inversions in the *Ap2-γ-Bmp7* locus indicated, that the TZ acted as boundary irrespective of the surrounding sequences. For example, in the Inv-M configuration, the interaction profile of the viewpoint located between the TZ and the mm75 enhancer was mirrored at the TZ. Due to its new position, the mm75 enhancer could activate *Ap2-γ*, which resulted in strong ectopic expression of *Ap2-γ* in the heart (Tsujimura *et al.*, submitted for publication). In the Inv-L2 configuration the interaction profiles of the viewpoint in the TZ and between the TZ and mm75 enhancer extended towards the centromere until the end of the *Dok5* gene. The end of this extension coincided with a TAD boundary (Dixon *et al.*, 2012). Interestingly, the viewpoint sitting inside the TZ had an interaction profile that extended into both directions in all configurations and respected the boundaries of the two neighboring domains. This suggests that the TZ does not function as an interaction blocker, but rather as an interaction sink or decoy. This behavior of strong interactions into both directions has often been observed at TAD boundaries (Dixon *et al.*, 2012).

In summary, both loci suggest a fixed chromatin structure that restricts the range of enhancers into regulatory domains and thereby defines enhancer speci-

ficity. The transition between the neighboring domains seems to be defined by the sequence of a TZ. However, what exactly is responsible for establishing this configuration has yet to be studied in detail. CCCTC-binding factor (CTCF), cohesin and Mediator are proteins involved in the 3D organization of the nucleus and they are proposed to play an important role in organizing the chromatin 3D structure at different length scales (Phillips-Cremins *et al.*, 2013). However, a recent study showed that depletion of CTCF and cohesin had only a weak influence on TAD boundaries (Zuin *et al.*, 2014). To functionally dissect the role of each of the involved factors, more specific experiments will have to be performed.

It has been shown recently for the HoxD cluster (Noordermeer *et al.*, 2014) that these boundaries are not completely fixed and can change to some extent between tissue types and developmental stages. The observed changes are only local within the HoxD cluster and the long-range interactions with neighboring TADs remain stable, showing that local changes within a larger stable 3D configuration are possible (Noordermeer *et al.*, 2014).

In contrast to this permissive model of chromatin 3D conformation, some loci were found, where *de novo* formation of chromatin 3D structure is triggered by tissue specific factors. Examples for this are the α- and β-globin locus during the maturation of erythroid cells (Drissen *et al.*, 2004; Vernimmen *et al.*, 2007). A recent study showed that *de novo* looping of the LCR to the promoter of the β-globin gene is required and is sufficient for the initiation of transcription in erythroid cells (Deng *et al.*, 2012).

Taken together, the predefined chromatin structure enables the formation of robust systems for gene expression which are required for development. These stable configurations allow for tight spatial and temporal regulation of gene activity by TF gradients and patterns. On the other hand, *de novo* formation of chromatin contacts upon binding of tissue specific factors contributes to cell maturation and stable establishment of new cell identities.

2.3.5 Relevance in cancer and disease

Long-range gene regulation is also the underlying molecular mechanism for some diseases, causing both Mendelian and complex disease traits. The effect sizes of these regulatory mutations range from small to large and their frequencies range from rare to common (Hindorff *et al.*, 2009). Approximately 90 % of SNPs from genome-wide association studies (GWASs) are variants in non-coding regions, which attributes these variants an important role in common diseases (Hindorff *et al.*, 2009).

In our analysis of genomic rearrangements and gene regulation in mouse we focused on non-coding enhancer regions. We observed limb malformations upon inversion of the region containing the ZRS, which is a result of the disruption of *Shh* expression in the developing limb (Symmons *et al.*, in preperation). Similarly, point mutations in the ZRS region are linked to heritable pre-axial polydactyly in human families (Lettice *et al.*, 2003). Despite disrupting existing enhancer

sequences, it also has been shown that point mutations are capable of forming new cis-regulatory sequences which regulate neighboring genes. An example is the inherited blood disorder α-thalassemia, which is caused by *de novo* formation of a promoter-like element that interferes with the regular expression of the α-globin genes (De Gobbi *et al.*, 2006).

In addition to point mutations, larger genomic rearrangements can disrupt the regulation of long-range enhancers. For example, in the Inv-M configuration of the *Ap2-γ* - *Bmp7* locus, the mm75 heart enhancer is moved into the regulatory domain of *Ap2-γ* which results in ectopic expression of *Ap2-γ* and loss of *Bmp7* expression in the heart (Tsujimura *et al.*, submitted for publication). This phenomenon of enhancer reallocation has been observed in several cases (Hérault *et al.*, 1997; Lower *et al.*, 2009). The relevance to cancer has been shown by Gostissa *et al.* (2009) in the case of Igh–c-myc translocations in B-cell lymphomas. Along this line, a recent study showed that genomic structural variants in group 3 and 4 medulloblastomas, which account for most paediatric cases, moved oncogenes of the GFI1 family into domains of active enhancers, causing the activation of GFI1 family oncogenes (Northcott *et al.*, 2014).

In summary, point mutations and genomic rearrangements of non-coding sequences can have profound molecular consequences for long-range gene regulation. To better understand and functionally dissect the influence of genomic rearrangements and point mutations on long-range gene regulation, 4C profiles need to be generated from the respective tissues and compared to control samples. This might help to narrow down the target genes of the non-coding variant and understand the molecular impact of the altered regulatory sequences.

2.3.6 Outlook

Most of the current chromosome conformation capture experiments require several million cells as starting material and only measure chromatin structure as an ensemble average. The protocols have to be improved in terms of sensitivity, to be able to use the techniques on cell or tissue samples for which only a few cells are available. A recent approach in this direction was the generation of single-cell Hi-C data (Nagano *et al.*, 2013). In this study they showed that single cells show variable chromosome structures that, when averaged over all single cells, resemble the signal obtained from bulk Hi-C experiments. Additional RNA- or GRO-seq data on the same individual cells would be of great value to understand the link of chromatin structure and gene regulation.

The cell to cell variability in chromosome structure has also been observed in single-cell microscopy studies (Amano *et al.*, 2009; Noordermeer *et al.*, 2011). These observations have to be integrated with the data obtained from 3C based experiments to refine the role of chromatin 3D structure in gene regulation.

A problem with all methods mentioned before is the requirement of cell fixation by cross-linking. This only allows the investigation of snap-shots of 3D chromatin dynamics. Further approaches are necessary to uncover the dynamics of chromatin

structure in live cells. A recent study gave first insights into these dynamics by tagging distal regions with Tet-operator binding sites in pro-B cells and tracing their trajectories with Tet-repressor-EGFP (Lucas *et al.*, 2014). They showed that distal elements bounce back and forth in a spring like fashion until they meet each other and form stable chromatin contacts. Furthermore, the time until these contacts are established for the first time mainly depends on the 3D confinement of TADs (Lucas *et al.*, 2014). However, to obtain a more global view on chromatin structure dynamics, high-throughput methods are necessary that can be used on many cells and cell types. Additionally, visualization of nascent RNA molecules from the corresponding gene in such a system would be of great value to completely understand the dynamics of chromatin structure and gene expression.

Chapter 3

The Pharmacogenetic Phenome Compendium (PGPC) resource

3.1 Description

In this chapter I describe the analysis of a high-throughput microscopy project carried out in collaboration with Marco Breinig from the Boutros group at the German Cancer Research Center (DKFZ). Marco Breinig performed the screen and provided the microscopy data on which this analysis is based. Subsequent follow-up experiments were performed by Marco Breinig, and I performed the statistical analysis of data from the high-throughput screen and the follow-up experiments. The content of this chapter is based on the manuscript *Integration of pharmacogenetic and phenotypic profiling predicts drug synergism and compound mode-of-action*, which we currently prepare for journal publication (Breinig *et al.*, in preperation), and on the **R** package `PGPC` which will be submitted to Bioconductor once the paper is accepted.

The aim of the screen was to find chemical-genetic interactions by screening the LOPAC drug library (Sigma, www.sigmaaldrich.com) of 1,280 pharmacologically active compounds against a panel of isogenic cell lines. It included two parental HCT116 colon cancer cell lines (P1 and P2), that each have three heterozygous driver mutations in the KRAS, PI3KCA, and β-catenin (CTNNB1) genes. Three cell lines were generated for KRAS, PI3KCA and CTNNB1 with a knockout (KO) of a single oncogenic mutant allele, leaving only one copy of the respective WT allele. These cell lines are referred to as KRAS WT, PI3KCA WT and CTNNB1 WT in this chapter. All other cell lines contain additional knockouts of AKT1, AKT1/2 together, PTEN, MAP2K1 (MEK1), MAP2K2 (MEK2), TP53, and BAX. HCT116 P1 was obtained from ATCC (www.lgcstandards-atcc.org) and all other cell lines from Horizon Discovery Ltd. (www.horizondiscovery.com/). An overview of the cell line genotypes and the affected pathways is shown in Figure 3.1.

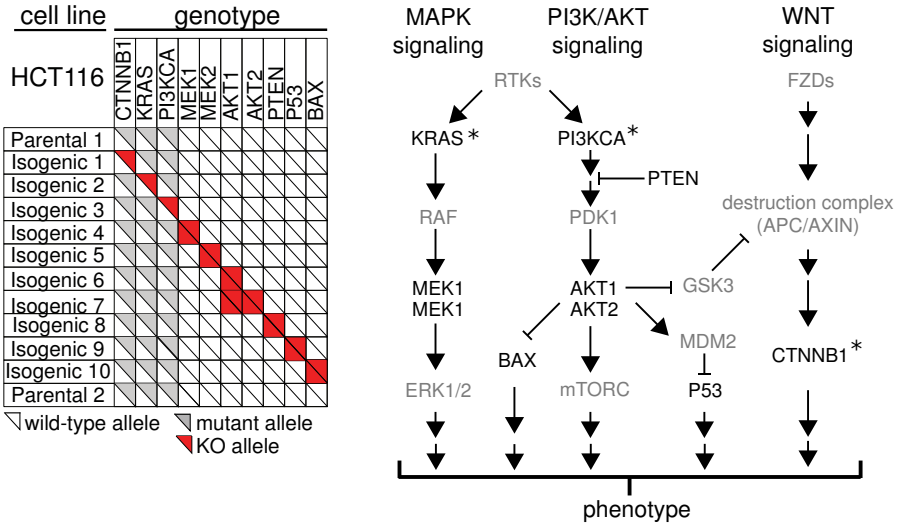


Figure 3.1: Overview of knockout alleles in the HCT116 colon cancer cell line panel on the left. The right panel shows the affected pathways. Taken from (Breinig *et al.*, in preperation).

3.2 Materials and Methods

3.2.1 Screening protocol

A workflow of the screening protocol is shown in Figure 3.2. The cells of each cell line were seeded and treated in 384 well plates. On average 1,250 cells were seeded in each well and incubated for 24 h. The compounds, dissolved in DMSO Dimethyl sulfoxide (DMSO), were added at a final concentration of 5 μM. After the compound treatment, cells were incubated for 48 h. Then, the cells were fixed and stained with Hoechst (DNA) and TRITC-Phalloidin (Actin). In the following step, the plates were imaged using an IN Cell Analyzer 2000 microscope (GE) with 10x magnification. For each well, we obtained four 12 bit gray scale images each of the Hoechst and TRITC channel. The size of the images was 2048 x 2048 px. In total, approximately 295,000 images were recorded.

3.2.2 Image and data processing

I processed the images using the Bioconductor packages EBImage (Pau *et al.*, 2010a) and imageHTS (Pau *et al.*, 2010b) following previously established approaches (Fuchs *et al.*, 2010; Horn *et al.*, 2011).

Due to low excitation and decreased signal at the image boarders, I cropped 150 px at each side of each image before processing. As a first step, nuclei in the DNA channel images were segmented by adaptive thresholding and subsequent morphological opening on the obtained segmentation mask. Next, a cell mask was generated from the Actin channel using a combination of a fixed and an adaptive

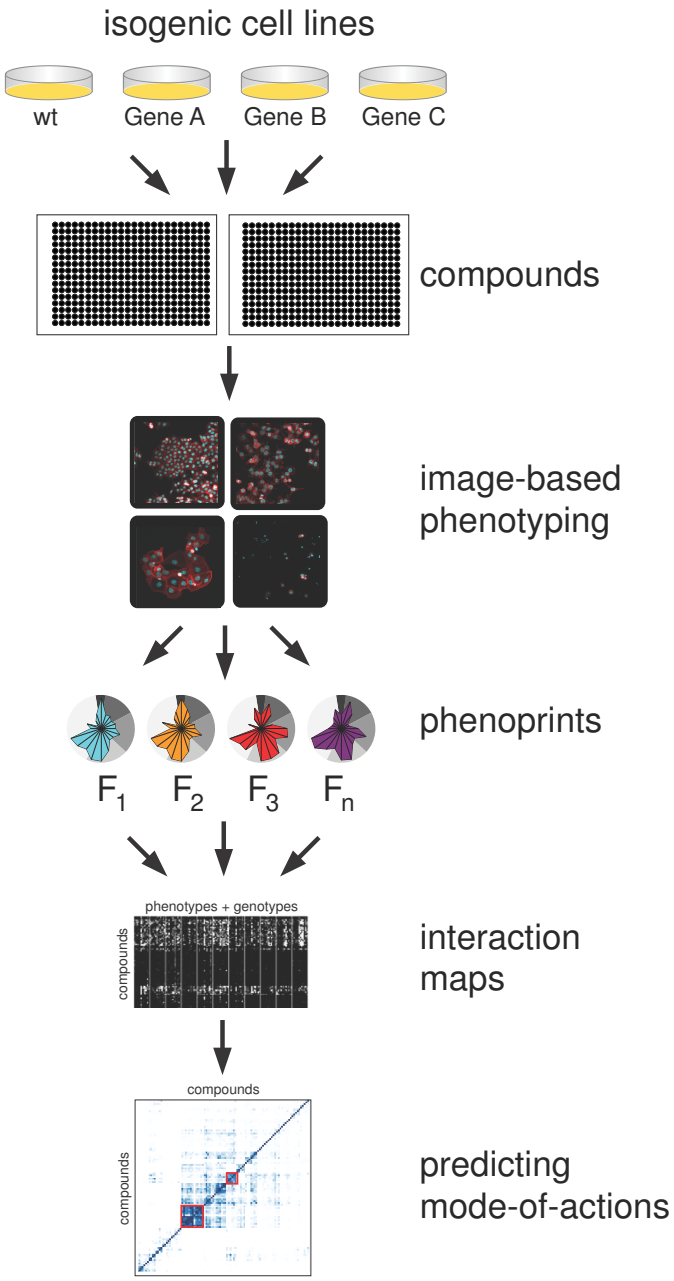


Figure 3.2: Workflow of the performed drug screen. Taken from (Breinig *et al.*, in preperation).

threshold. To segment cells, nuclei were used as seeds for a Voronoi tessellation and inflated into the cell mask (Jones *et al.*, 2005). To filter out small debris and large staining artifacts, candidate nuclei objects with a size of less than 100 px or more than 3000 px, and candidate cell objects with a size of more than 15000 px were removed from the segmentation masks.

After nuclei and cell segmentation, I extracted morphological and texture features from each single cell by using the nuclei and cell segmentation masks. The single cell features were summarized per well. The number of segmented cells per well was used as a surrogate for cell number or growth. I calculated the trimmed mean across all cells in a well for each feature, discarding the bottom and top 10 % of the values. Additionally, the standard deviation of each feature was calculated across the cell in each well. For some features additional quantiles at 1 %, 5%, 95% and 99% were calculated. In total 395 features were extracted for each well.

It is convenient to transform the feature data to a logarithmic scale for model fitting with a multiplicative interaction model, which was used in previous studies (Mani *et al.*, 2008; Costanzo *et al.*, 2010; Horn *et al.*, 2011). To avoid singularities with negative feature values, I used a so-called generalized logarithm transformation (Huber *et al.*, 2002):

$$f(x, c) = \log\left(\frac{x + \sqrt{x^2 + c^2}}{2}\right). \tag{3.1}$$

For $c = 0$, this corresponds to the regular logarithm. For $c > 0$, the function is smooth for all real values of x, avoiding singularities at zero and negative values. For $x \gg c$ the function is approximately equivalent to the regular logarithm. For c, I used the 5% quantile of the empirical distribution of each feature. In some cases, when the quantile was zero, I used 0.05 times the maximum as parameter c.

To assess the reproducibility for each feature, I calculated the correlation between the two replicates. All features that had a correlation coefficient of less than 0.7 between the replicates were removed from further analysis.

3.2.3 Feature selection and display of phenotypic profiles as *phenoprints*

The data set contained partially redundant features. Therefore I adapted an approach to select a subset of the feature containing the least redundant information, as described in a previous study (Laufer *et al.*, 2013). Starting from the preselected cell number feature, this approach iteratively selected the feature that had the highest residual information. For this, each feature was fitted by a linear model of the previously selected features and the correlation between the model residuals of the replicates was used as surrogate for the residual information. This process was stopped when the number of residual feature correlation coefficients with a positive value was smaller than the number of negative values. At this point the sampling of new information switched over to sampling from random noise for

which one expects to have approximately 50 % positively and 50 % negatively correlated features.

The isogenic cell lines had varying proliferation rates that gave rise to differences in the final number of cells detected an the endpoint of the protocol. To account for this effect, the cell number values were scaled, by affine transformation, for each cell line and replicate to the dynamic range defined by the median values of the negative (DMSO) and positive (Paclitaxel) controls. The value of the negative controls was set to 1 and the value of the positive controls was set to 0. With this approach scaled values below 0 and above 1 were possible.

For visualization of compound phenotypes from the extracted features, I scaled the features using the following formula:

$$y_{ij} = \frac{x_{ij}}{\mathrm{MAD}_j}, \tag{3.2}$$

where x_{ij} is the mean value of the two replicates for each well i and feature j, and MAD_j is the mean of the median absolute deviation calculated across all wells for each replicate. These scaled feature values y_{ij} were further transformed with an affine transformation to the interval from 0 to 1, so that, after the transformation, the minimum value for each scaled feature is 0 and the maximum value 1. These final values of selected features, which we termed *phenotypic profiles*, were shown in radar plots, which we also called *phenoprints* (Section 3.3.2).

3.2.4 Detection of gene-drug interactions

To detect chemical genetic interactions I fitted the data with a multiplicative model as previously described (Costanzo *et al.*, 2010; Horn *et al.*, 2011; Laufer *et al.*, 2013). The model was fitted using robust L1 regression implemented in the *medpolish* function of the **R** package stats. In this iterative process row and column medians were subtracted alternately until the relative change of the residuals falls below a predefined threshold. The drug and cell line effects were represented by the final row and column values respectively. The residual terms represented the chemical genetic interaction coefficients. The process was performed separately on both replicates for each individual feature.

To detect significant interactions, I used a moderated t-test, implemented in the *limma* Bioconductor **R** package (Smyth, 2005) on the interaction coefficients of both replicates with null hypothesis $\mu = 0$. The p-values of this test were multiple testing corrected using the method of Benjamini and Hochberg (Benjamini and Hochberg, 1995) to control the false discovery rate. All interactions with an adjusted p-value below the chosen threshold of 0.01 were selected as significant interactions.

3.2.5 Clustering of drugs and cell lines based on their interaction profiles

To display the interaction profiles, I calculated robust z-scores of the interaction coefficients, scaling them in the same way as described in Formula (3.2).

Cell lines or compounds were clustered using the correlation between robust z-score interactions calculated for cell lines and compound profiles respectively. The following distance metrics were used between cell line and compound profiles:

$$d_{\text{cell line}}(y_1, y_2) = exp(-cor(y_1, y_2)) - exp(-1). \tag{3.3}$$

$$d_{\text{compounds}}(y_1, y_2) = 1 - cor(y_1, y_2), \tag{3.4}$$

where y_1 and y_2 are the scaled interaction profiles of the respective cell lines or compounds.

The cell lines and drugs were clustered using hierarchical clustering with the calculated distances between cell line and drug profiles respectively.

3.2.5.1 Chemical similarity of compounds

To include information about the chemical similarity of compounds, I calculated the chemical distance between compounds using Tanimoto coefficients (Tanimoto, 1957). I used the implementation in the `ChemmineR` Bioconductor **R** package (Cao *et al.*, 2008) to perform this task. This chemical similarity information is displayed together with the drug interaction profile similarities.

3.2.5.2 Correlation between interaction profiles of shared drug targets

To test whether the combined approach of high-content imaging and using a panel of isogenic cell lines was superior to using each approach alone, I calculated the empirical distribution of correlation coefficients between scaled drug interaction profiles. As inputs I used the following three data sets. First, the whole data set, representing the combined approach, second, the data set of all features for the parental cell line *HCT116 P1*, representing a high-content approach with one single cell line, and third, the data set of only the cell number features for all 12 cell lines, representing a viability screen of a cell line panel. The correlation coefficients were categorized into the classes *shared* or *non-shared* target selectivity based on whether they are annotated to share the same drug target. The difference in the area under the curve (AUC) of the empirical cumulative distribution function between the classes was calculated. This value served as an surrogate for how well each approach could separate the two classes.

In a second categorization, correlation coefficients were classified into *similar* or *different* based on their chemical structure similarity. If the chemical distance, calculated with the `ChemmineR` **R** package, was below 0.6, compounds were assigned to the *similar* chemical structure class.

3.2.6 Follow-up: Quantification of synergistic compound combinations

To test predicted synergistic drug combinations, my collaborator measured the influence of pairwise compound combinations on cell viability. He mimicked the cell line genotype of a certain gene knockout, for which a chemical genetic interaction was observed with a compound, by treating the parental HCT116 P1 cell line with a combinatorial treatment of this compound and an inhibitor of the knocked out gene. As a readout my collaborator used a CellTiterGlo assay (Promega). Compounds were combined and then diluted in a 1:2 dilution series to cover 10 concentrations. In the final concentration of drug treatment they cover a concentration range of 10 μM to 0.0195 μM. For the assay 1000 cells were seeded in 384 well plates and incubated for 24 h. After the compound treatment cells were incubated for additional 72 h, followed by the CellTiterGlo assay readout using a Mithras LB940 plate reader (Berthold Technologies).

For data normalization, I used the *cellHTS2* Bioconductor **R** package. The plate reader data was first log transformed and then normalized with the normalized percent inhibition (NPI) method. This means that, for each plate, the value of each measurement is subtracted from the mean value of the positive controls and then divided by the difference between the mean values of the positive and negative controls.

To quantify synergistic drug combinations, I used the Bliss independence (BI) and the highest single agent (HSA) previously described in large-scale combinatorial compound screens for drug synergism (Borisy *et al.*, 2003; Tan *et al.*, 2012).

In the multiplicative BI model, the expected value of combined drug inhibition is the sum of the individual inhibitions on the log scale:

$$I_{\text{combination,BI}} = I_{\text{drug A}} + I_{\text{drug B}}. \tag{3.5}$$

For the HSA model the expected value is the maximum inhibition of the two individual drug inhibitions:

$$I_{\text{combination,HSA}} = \max(I_{\text{drug A}}, I_{\text{drug B}}) \tag{3.6}$$

The inhibition is calculated from the measured NPI values in the following way:

$$I_{\text{drug x}} = 1 - \text{NPI}_{\text{drug x}} \tag{3.7}$$

To detect statistically significant synergistic drug combinations, the measured inhibition values of the drug combinations were compared to the expected values using a one-sided two-sample Student t-test with the alternative that the mean of the measured values is smaller than the mean of the expected values.

3.2.7 Follow-up: Quantification of the proteasome inhibition of compounds

To test whether the compounds found in one of the drug clusters affect proteasome activity, my collaborator performed a cell-based proteasome activity assay. For this, HCT116 cells were seeded in a 384 well plate at a concentration of approximately 3000 cells per well and incubated for 24 h. Then the compounds ZPCK, Disulfiram, CAPE, tyrphostin AG555, AG1478, DAPH (all from Sigma), Bortezomib (from NEB) and MG132 (from MerckBioscience) were added at a final concentration of 5 μM and 0.1 % DMSO. After additional incubation for 24 h the caspase-like, trypsin-like and chymotrypsin-like proteasome activity was measured with the Proteasome-GloTM Cell-Based Assay Kit (Promega). The readout was done with a Mithras LB940 plate reader (Berthold Technologies). To normalize for compound effects on cell viability, a CellTiterGlo assay (Promega) was performed.

For the data analysis, I calculated the mean value of the two wells for each assay-drug combination on the 5 plates. These values were corrected for viability effects by dividing the corresponding value of the CellTiterGlo assay on the plate, thereby setting the values of the CellTiterGlo assay to 1 on each plate for each assay. The proteasome activity was then calculated relative to the DMSO controls by dividing each value by the corresponding DMSO value, setting the proteasome activity of all DMSO measurements to 1 on each plate for each assay. The inhibition I_{drug} for each drug was calculated as the difference from this value:

$$I_{\text{drug}} = 1 - \text{NPA}_{\text{drug}}\,, \tag{3.8}$$

with NPA_{drug} being the normalized proteasome activity for the given drug.

To test for statistically significant proteasome inhibition of the compounds, I performed a Student's t-test comparing inhibition values of each compound against the null hypothesis of zero effect.

3.3 Results

We established a high-throughput microscopy approach to quantitatively measure the genotype dependent phenotypes in response to compound treatment. The LOPAC library of 1280 pharmacologically active compounds was screened against a panel of isogenic HCT116 colon cancer cell lines (overview in Figure 3.1).

The images were processed as described in Section 3.2.2. For each well 395 features were extracted that describe the phenotype of cells treated with the compound in the respective well.

3.3.1 Data quality control and feature selection

In this section I describe three quality control steps that I performed on our data set. First, I checked the reproducibility of features between replicates, followed by

the selection of a non-redundant set of features. For the selected features I checked that no obvious technical biases were present.

The reproducibility between replicates was high for many features. In total 310 features exceeded the threshold of 0.7 for the Pearson correlation coefficient. The distribution of these correlation coefficients is shown in Figure 3.3. For three features individual scatter plots are shown in Figure 3.4. They represent the feature *cell number*, which is highly reproducible, the feature *1 % quantile of nuclei eccentricity*, which falls just below the chosen threshold of 0.7, and the feature *mean position of cell nuclei on the x-axis*, which, as expected, has essentially zero correlation.

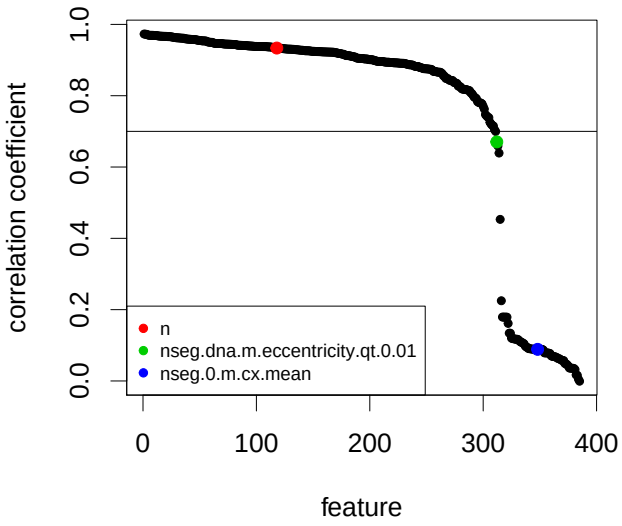
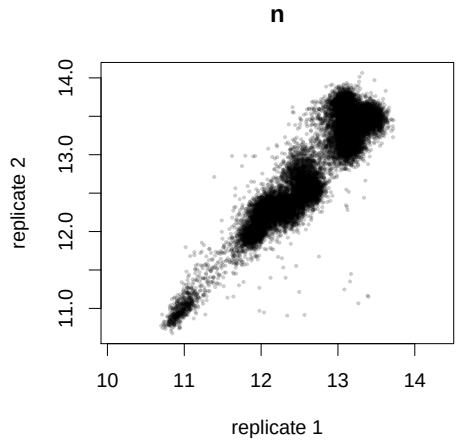


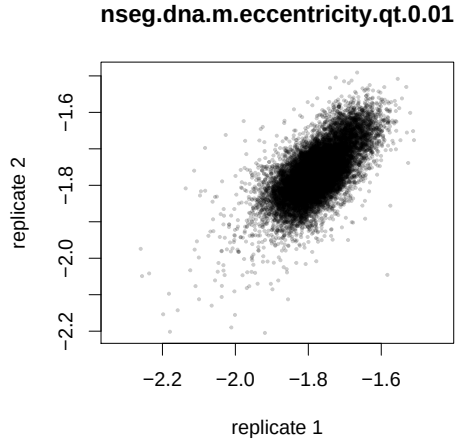
Figure 3.3: Distribution of correlation coefficients for all extracted features. The horizontal line represents the threshold of 0.7 used to filter out features that were not reproducible across replicates. The colored points represent the single features shown in Figure 3.4.

The 310 features that were left after the filtering step were partially redundant. In order to select a non-redundant subset of features, I used the iterative feature selection algorithm described in Section 3.2.3, using cell number as the first preselected feature. This algorithm selected a set of 20 features before the stop criterion was reached. The residual correlation of the selected feature at each iteration is shown in Figure 3.5 (a). Except for the second iteration, where the first free feature is chosen after the preselected cell number, the values of the residual correlations decreased with the number of iterations, representing a decrease of the residual information that was contained in the remaining features.

(a)



(b)



(c)

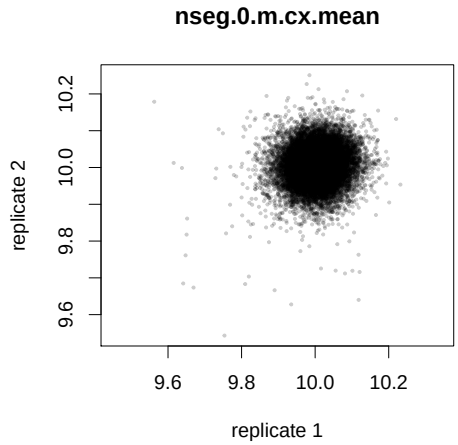


Figure 3.4: Scatter plots of individual features. Cell number, which was highly correlated (a), 1 % quantile of the nuclei eccentricity, which was just below the threshold of 0.7 (b), and the mean position of cell nuclei on the x-axis with a low correlation (c)

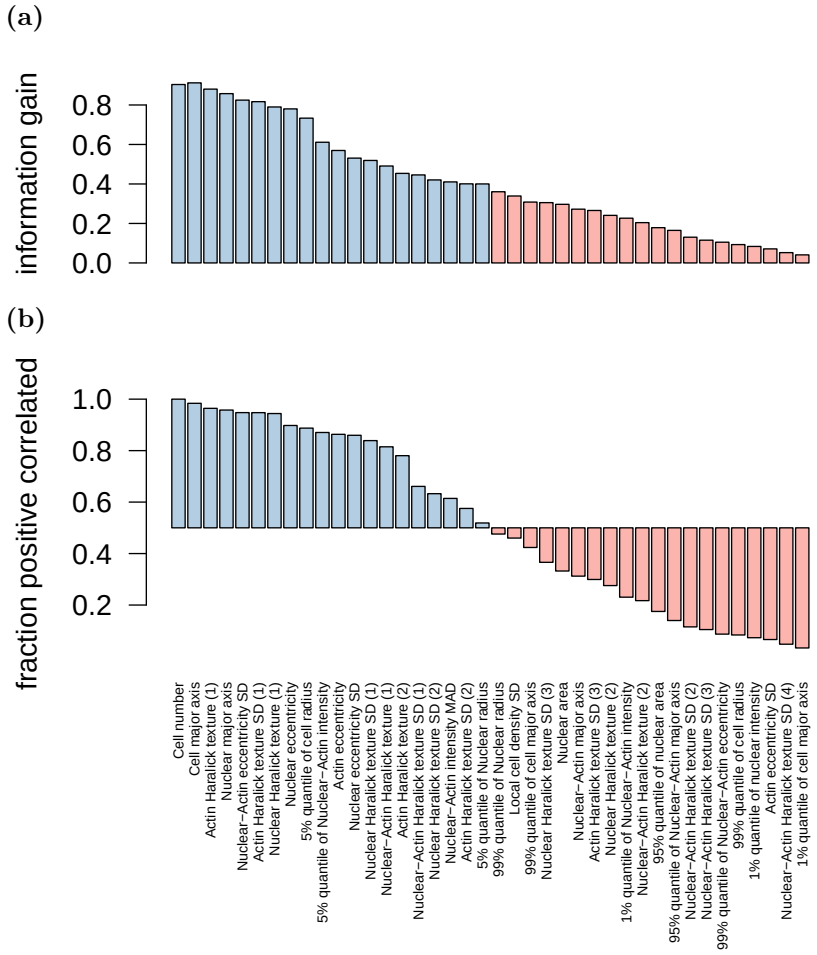


Figure 3.5: Result of the feature selection algorithm. Correlation of the residual feature values at each step of the iterative feature selection algorithm (a). Fraction of positively correlated residual features at each step of the iterative feature selection algorithm (b).

The fraction of positively correlated residual features, which was used for the stop criterion of the algorithm is shown in Figure 3.5 (b). All features for which the number of positively correlated residual features was greater than 0.5 were selected.

To check for potential batch effects and spatial plate effects I generated box plots and screen plots of the whole plate for the transformed data of all selected features. In Figure 3.6 the box plots for the cell number is shown. The pattern of 4 plates per cell line is clearly visible in this plot and the plates were very consistent per replicate.

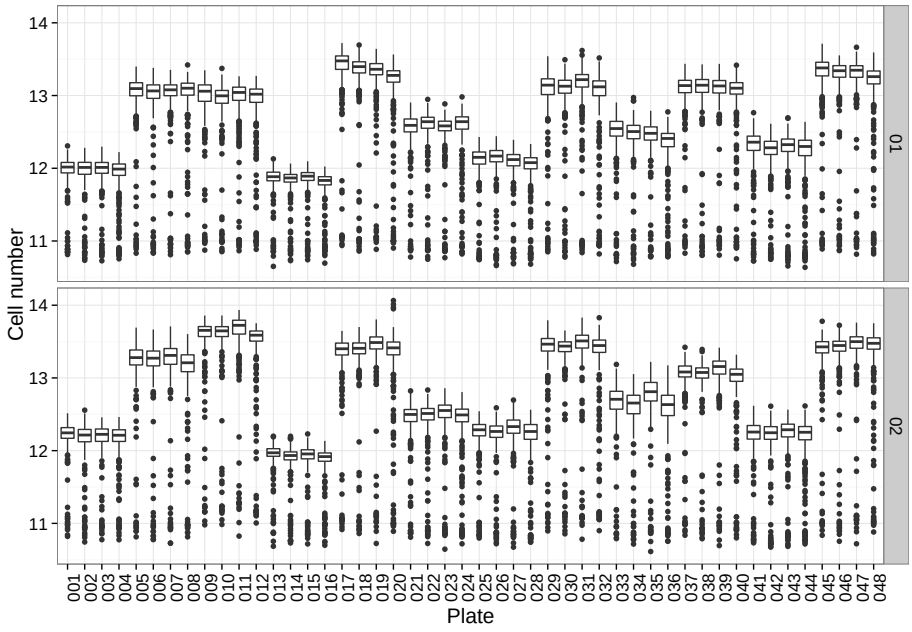


Figure 3.6: Box plots of the transformed cell number for each plate. The upper panel shows the values for replicate 1 and the lower for replicate 2.

A screen plot of the cell number is shown in Figure 3.7 of all plates for replicate 1. The controls on the left side of the plates could be seen, as well as the periodic pattern of 4 plates per cell line. No strong spatial effects over any plate could be observed in this plot. Similar plots were observed for the other features, showing that the analysis is not strongly influenced by possible plate or batch effects.

To account for the different proliferation rates of the isogenic cell lines (see Figure 3.6), I scaled the cell number feature using the median values of the positive and negative controls for each cell line as described in Section 3.2.3. A scatter plot of the scaled cell number is shown in Figure 3.8. By comparison to the unscaled cell number shown in Figure 3.4 (a), we can see that the effect of different proliferation rates has been removed.

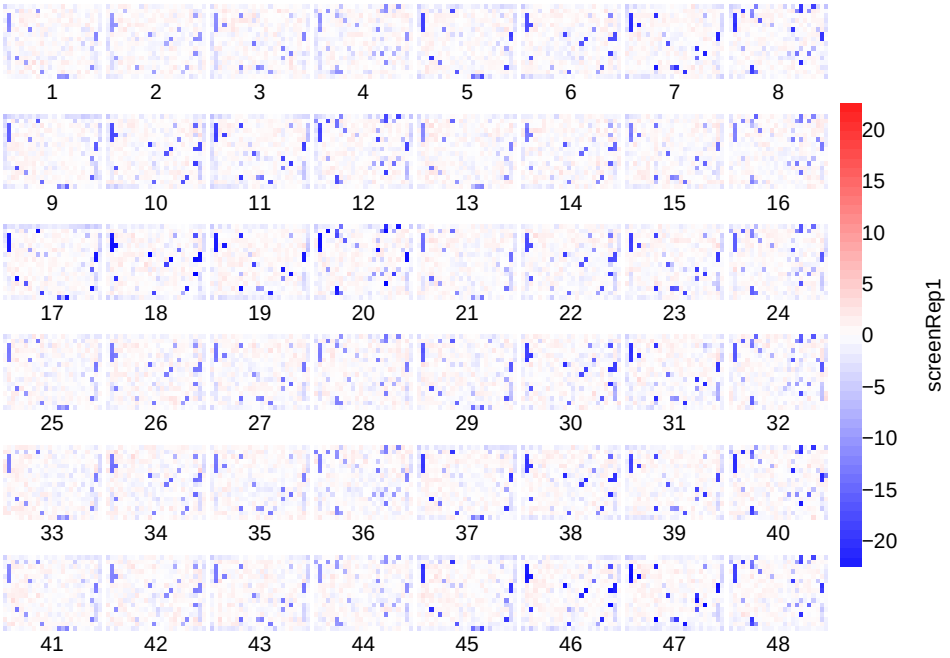


Figure 3.7: Screen plot showing the cell number feature on all plates of replicate 1 in the screen. The plot represents robust z-scores, calculated per plate by subtracting the plate median and scaling by the median absolute deviation of each plate.

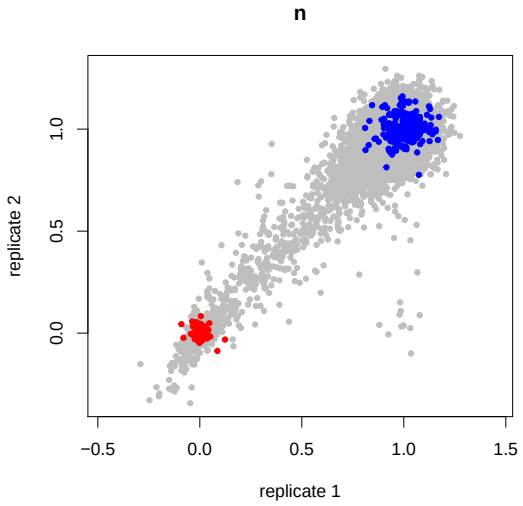


Figure 3.8: Scatter plot of cell number scaled by the median values of the positive and negative controls of each cell line. The positive controls are shown in red and the negative controls in blue.

3.3.2 Representation of phenotypes by phenoprints

To visualize the extracted features for different drugs, the values for each of the 20 features were scaled and transformed to the interval from 0 to 1 as described in section 3.2.3. The 20 features were grouped into the 5 phenotypic categories, namely cell number, DNA texture and intensity, nuclear shape, cell shape, and actin texture and intensity. The phenotypes were displayed as radar plots named *phenoprints* with the features ordered according to their phenotypic group (Figure 3.9).

We first focused on phenoprints of different drugs observed in the parental cell line HCT116 P1. The phenoprint for a DMSO control is shown in Figure 3.9 (b). This phenoprint was similar to many compound phenoprints that did not show a compound phenotype. An example for this was the phenoprint of the MEK-inhibitor PD98.059 shown in Figure 3.9 (c). For other compounds that induced a strong phenotype the phenoprint changed accordingly compared to the DMSO control phenoprint. These changes were consistent for compounds which gave rise to the same phenotype. For tubulin-targeting compounds the phenotype corresponded to cell death and condensed apoptotic nuclei. The phenoprints for the two tubulin-targeting compounds Vinblastine and Vincristine represented this phenotype and were highly similar (Figure 3.9 (d), (e)). The same was true for the topoisomerase-targeting compounds Etoposide and Amsacrine. For those, the phenoprint represented the phenotype of increased nuclear and cell size, attributed to mitotic catastrophe (Maskey *et al.*, 2013). Both compounds had similar phenoprints as shown in Figure 3.9 (f), (g). Other captured phenotypes were, for example, cell death with abnormal nuclei and strongly decreased actin signal, observed for the compound Ouabain (Figure 3.9 (h)), and elongated cells for the compound Rottlerin (Figure 3.9 (i)).

The different phenotypes inherent to the different genetic backgrounds were also captured by phenoprints as shown in Figure 3.10. However, the differences are only modest for most isogenic cell lines compared to the parental cell lines. The strongest differences were observed for the double AKT1/2 knockout cell line, which exhibited strongly reduced actin signals.

Taken together, this showed that our approach was capable of detecting different phenotypes induced upon compound treatment or based on the underlying genotype of the isogenic cell line and that these phenotypes could be represented by the selected features as phenoprints.

3.3.3 Cell line specific responses to compound treatment

As already mentioned, the phenoprints for the different isogenic cell lines represented some differences between the isogenic cell lines. For example, the phenoprint of the CTNNB1 wild-type cell line represented the protrusions of the cell bodies and aberrant nuclear shape compared to the parental cell line (Figure 3.11)

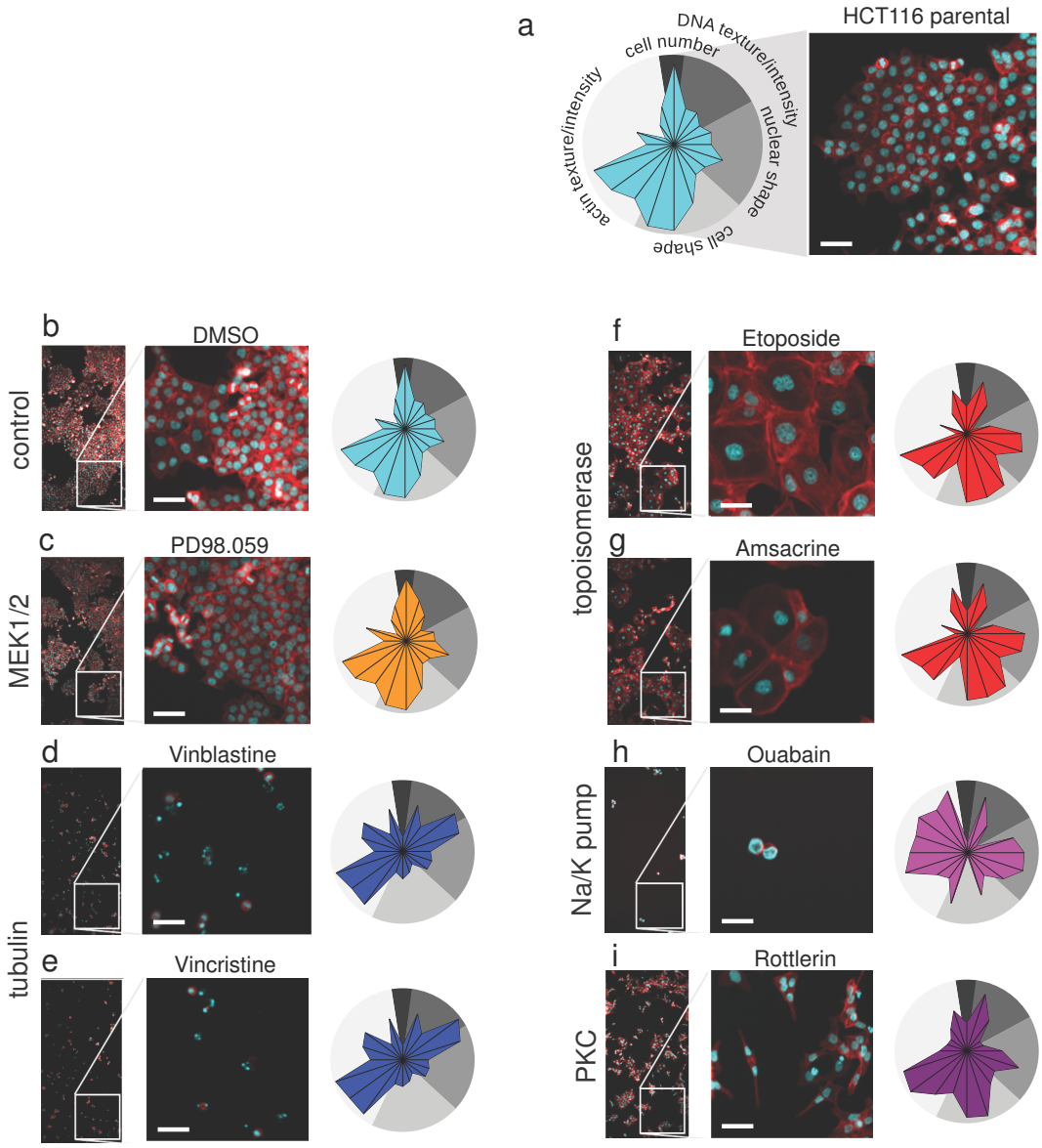


Figure 3.9: Phenotypes and phenoprints of different compounds. The features are ordered according to their phenotypic group, as shown in the legend (a). Phenoprint for DMSO control (b), MEK-inhibitor PD98.059 (c), tubulin-targeting compounds Vinblastine (f) and Vincristine (e), topoisomerase-targeting compounds Etoposide (f) and Amsacrine (g), Na/K pump inhibitor Ouabain (h) and PKC inhibitor Rottlerin (i). Scale bars, 20 μm. Taken from (Breinig *et al.*, in preperation).

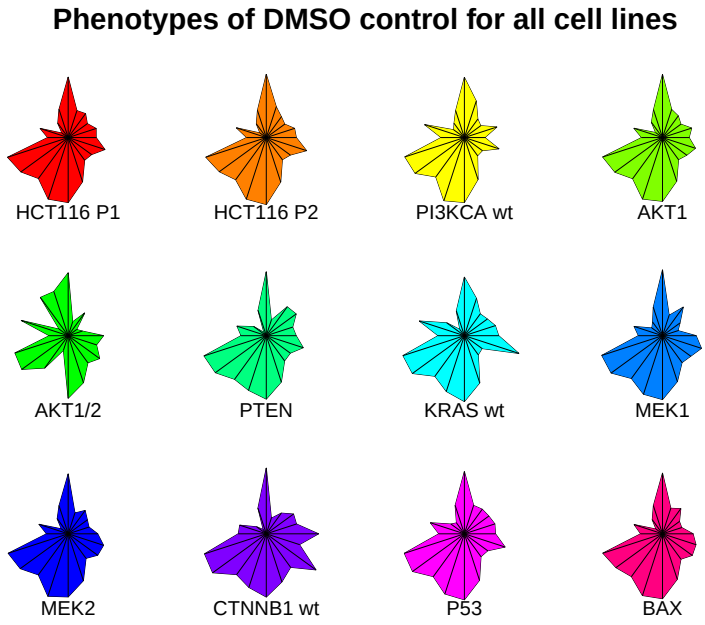


Figure 3.10: Phenoprints of the different isogenic cell lines for one DMSO control well.

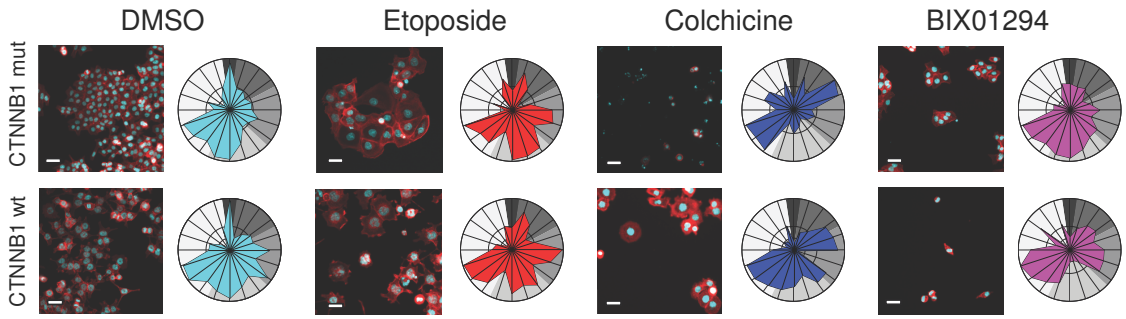


Figure 3.11: Phenotypes and phenoprints for DMSO, Etoposide, Colchicine and BIX01294 in the parental cell line (CTNNB1 mutant (mut)) and CTNNB1 mutant knockout cell line (CTNNB1 WT). Taken from (Breinig *et al.*, in preperation).

Compound treatments either induced genotype-dependent or genotype-independent phenotypic alterations in the different isogenic cell lines. For example, cellular and nuclear size increased in parental HCT116 cells and CTNNB1 WT cells upon Etoposide treatment. In both cases the corresponding phenoprints changed accordingly, displaying comparable changes (Figure 3.11). Colchicine, a spindle toxin, on the other hand, induced apoptosis in parental HCT116 cells and led to inflated cells in the CTNNB1 WT background (Figure 3.11). Furthermore, BIX01294, a histone methyltransferase inhibitor, induced strong cell condensation in CTNNB1 WT cells while only having a slight effect on parental HCT116 cells (Figure 3.11).

In both cases these phenotypic differences were observed as different changes in the shapes of the corresponding phenoprints shown in Figure 3.11.

In order to quantitatively assess such differences in the response to compound treatment between the genetic backgrounds, I calculated compound-cell line interaction coefficients for the 20 selected features as described in section 3.2.4.

As an example, the obtained interaction coefficients are shown in Figure 3.12 for cell number. The interaction terms clustered around the origin and only a few statistically significant interactions were observed, colored in blue if the adjusted p-value was below 0.05 and colored in red if the adjusted p-value was below 0.01.

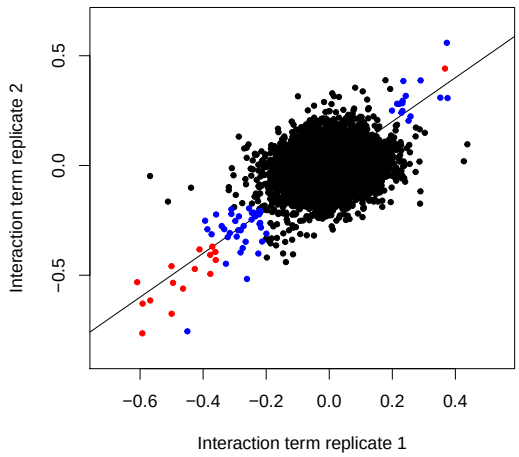


Figure 3.12: Interaction coefficients calculated for cell number. Points with an adjusted p-value below the threshold of 0.05 are colored in blue and below 0.01 in red.

After multiple testing correction and using a stringent threshold of 0.01 for the adjusted p-values, 2359 significant interactions were scored. In total, the 2359 significant interactions represented 0.8 % of all possible interactions. 193 compounds showed at least one significant interaction for one of the features, which represented 15.1 % of the 1280 compounds tested. The distribution of the number of drugs that had at least one interaction with one or up to twelve cell lines is shown in Figure 3.13 (a). The total number of significant interactions and number of drugs having at least one significant interaction varied strongly between the different isogenic cell lines (Figure 3.13 b and c). We observed that the cell lines that already showed different phenotypes, e.g., the AKT1/2 knockout and CTNNB1 WT cell line, showed the highest number of interactions, whereas the lowest number of interactions was observed for the two parental HCT116 cell lines. This is in line with previous studies in yeast, where the number of interactions per gene correlated with the strength of the effect of the single gene deletion (Costanzo *et al.*, 2010).

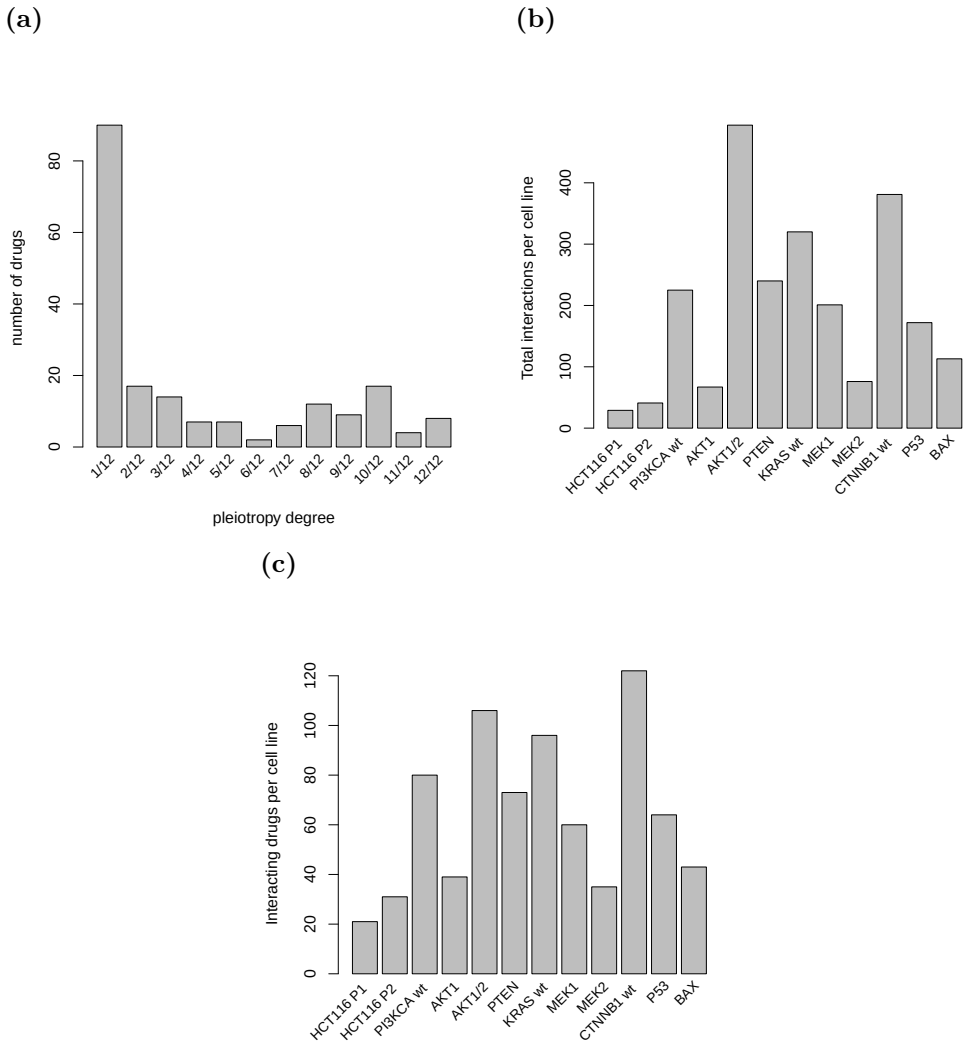


Figure 3.13: Distribution of the number of drugs that had at least one significant interaction in a certain number of cell lines out of the 12 possible cell lines (a). Distribution of the total number of significant interactions per cell line (b). Number of compounds that had at least one significant interaction per cell line (c).

Since we calculated interactions for all 20 selected features, we looked at the overlap of called interactions from the 5 phenotypic categories. As shown in Figure 3.14, there were many interactions specific to one of the 5 phenotypic categories. Thereby, the use of a multi-parametric feature space strongly increased the tested interactome space for compound-cell line interactions.

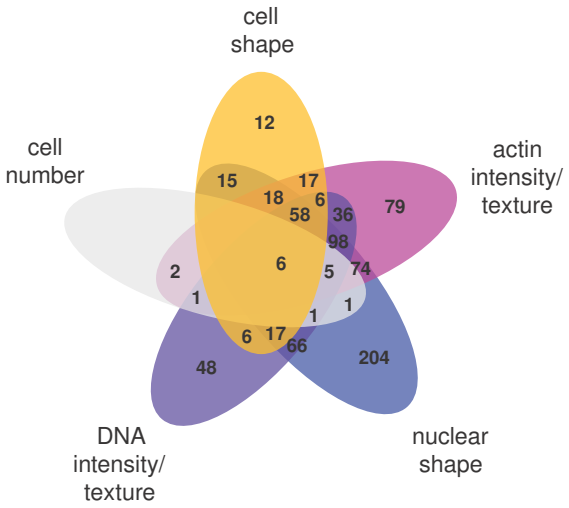


Figure 3.14: Venn diagram of interactions shared between the 5 phenotypic categories. Blank field represent zero values that have been removed for the visualization. Taken from (Breinig *et al.*, in preperation).

3.3.3.1 Specific interaction of the KRAS WT cell line and Ganciclovir

To benchmark the quality of our approach to call cell line specific interactions, I made use of a positive control that was contained in our data set. The KRAS WT cell line expresses a viral Thymidine Kinase (TK) that was used as selection marker during genetic engineering of the cell line and left as a fingerprint. The compound Ganciclovir is converted into a highly toxic compound upon TK activity (Crumpacker, 1996). Therefore Ganciclovir selectively impaired cell fitness for the KRAS WT cell line, affecting the cell number and causing phenotypic changes of increased cell nuclei and body (Figure 3.15).

The interaction profiles of Ganciclovir are shown in Figure 3.16 as starplot. In this representation the interaction coefficient of each feature is represented as segment from the origin (0) to the corresponding value. The scale is defined by the maximum value observed across the features. Specifically for the KRAS WT cell lines interactions were observed. For all other cell lines the calculated interaction coefficients were extremely small. Therefore, our approach clearly identified the specific interaction of Ganciclovir and the KRAS WT cell line. In this respect Ganciclovir served as a good positive control.

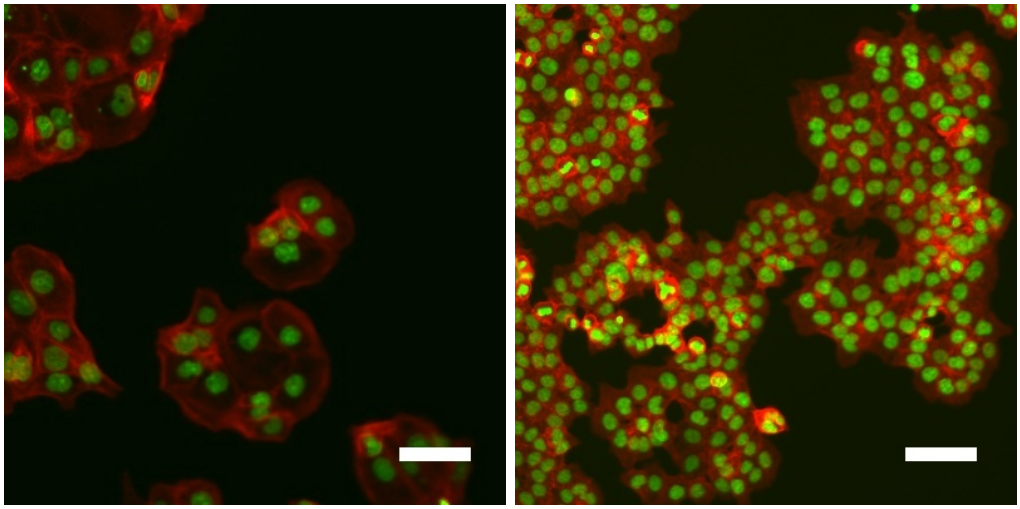


Figure 3.15: Phenotype of the KRAS WT cell line upon Ganciclovir treatment (left panel). Cell nuclei and body are increased in comparison to DMSO treated cells and the cell number is reduced (negative control, right panel). Scale bars, 20 μm.

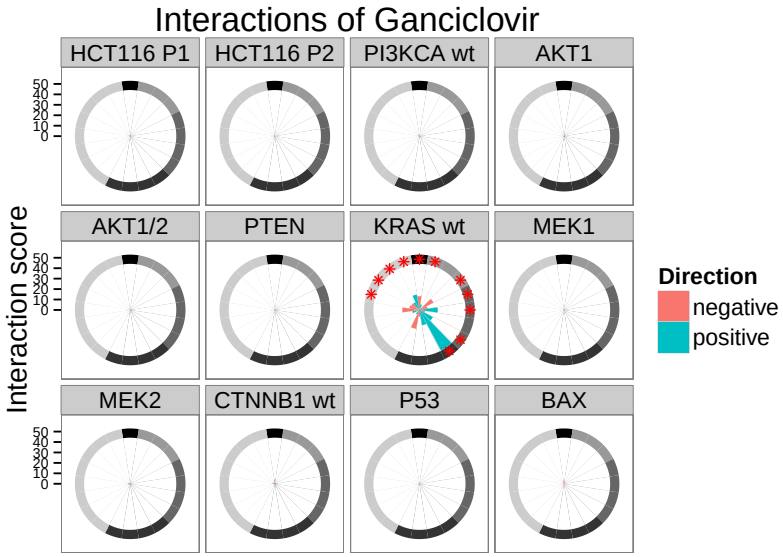


Figure 3.16: Starplots of the interaction coefficients observed for the different cell lines upon Ganciclovir treatment. The KRAS WT cell lines shows strong interactions for several of the selected features. (*) adjusted p-value < 0.01, red and turquoise define whether the sign of the interaction coefficient was positive or negative.

3.3.3.2 Specific interactions of the CTNNB1 WT cell line

Examples for cell line specific changes upon compound treatment were the interactions between the CTNNB1 WT cell line and the compounds Colchicine and BIX01294, already mentioned at the beginning of Section 3.3.3. For these compound cell line combinations the interaction profiles are shown in Figure 3.17.

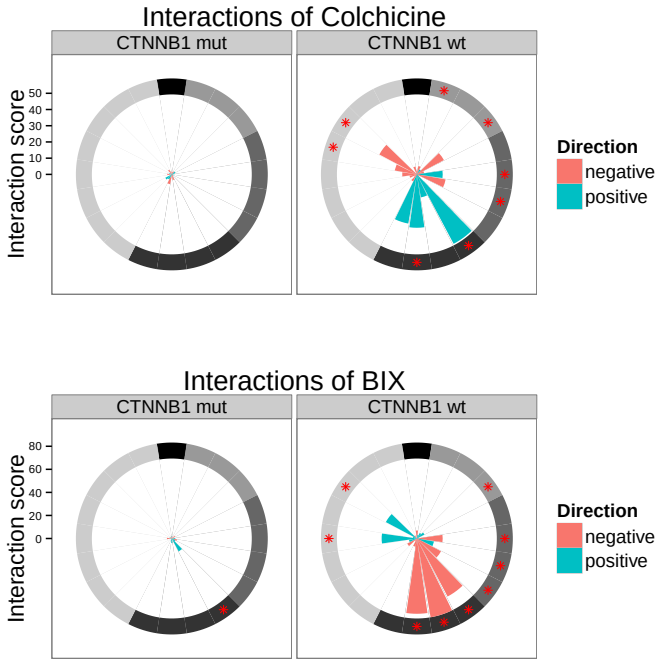


Figure 3.17: Starplots of the interaction coefficients observed for the parental cell line HCT116 P1 and the CTNNB1 WT cell lines and the compounds Colchicine and BIX01294. (*) adjusted p-value < 0.01, red and turquoise define whether the sign of the interaction coefficient was positive or negative.

The observed phenotypic changes were therefore captured by the selected set of features and detected as cell line specific interactions with our approach. Of importance is the fact that in both cases no chemical genetic interaction was observed for the cell number. Therefore, these interactions would not be detected, if only cellular fitness was used as a readout.

3.3.3.3 Compound cell line interaction network

Integrating all detected interactions, we created a phenotypic chemical-genetic interaction map. For this map, I filtered out compounds that showed an interaction for only one single feature or with more than 3 cell lines to reduce over-plotting. The resulting map is shown in Figure 3.18.

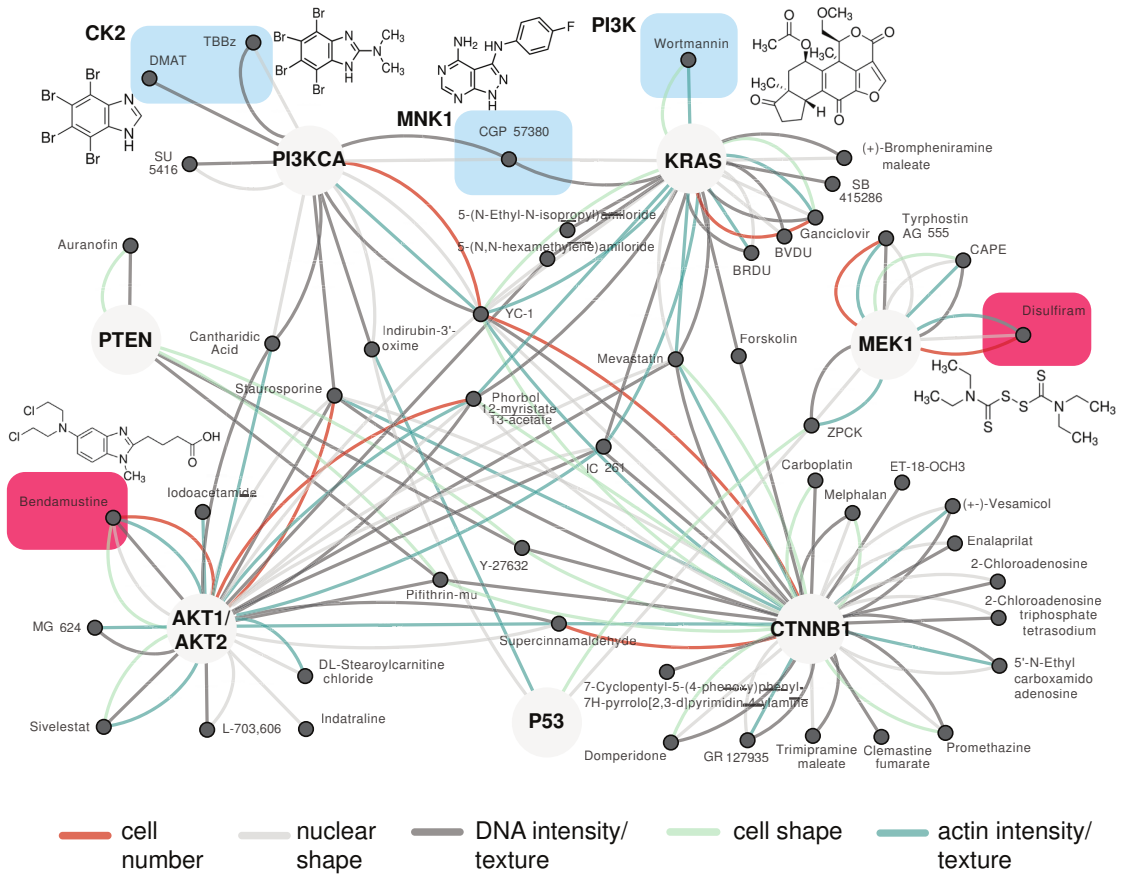


Figure 3.18: Interaction map of chemical-genetic interactions observed in our data set. The colored edges in the network represent the different phenotypic classes. Boxes highlight compounds for which the chemical structures are shown and that are discussed in more detail in the text. Taken from (Breinig *et al.*, in preperation).

Several examples of biologically relevant interactions are highlighted. The PI3K inhibitor Wortmannin was interacting with KRAS WT cells. This suggests that cells with WT KRAS signaling are more dependent on PI3K signaling and are particularly affected by the inhibition of PI3K.

The Casein Kinase 2 (CK2) inhibitors TBBz (4,5,6,7-Tetrabromobenzimidazole) and DMAT (2-Dimethylamino-4,5,6,7-tetrabromo-1H-benzimidazole) showed interactions with the PI3K WT cell line but not with the AKT1 and AKT1/2 KO cell lines, suggesting that CK2 could compensate perturbations at the level of PI3K and potentially further upstream in the PI3K-AKT pathway. This is in line with the known function of CK2 as a direct activator of AKT and its inhibitory function on PTEN (Torres and Pulido, 2001; Di Maira *et al.*, 2005). In addition, the recent finding that a combinatorial inhibition of CK2 and epidermal growth factor receptor (EGFR) is synergistic, agrees with our data (Bliesath *et al.*, 2012).

The MNK1 inhibitor CGP 57380 (N3-(4-fluorophenyl)-1h-pyrazolo[3,4-d]pyrimidine-3,4-diamine) showed interactions with the KRAS WT and PI3K WT cell line. This suggests that the MAPK- and PI3K-signaling pathway meet at the level of MNK1. MNK1 has been shown to be activated by MAPK signaling through Erk1 and Erk2 and to modulate cell growth and proliferation through the interaction with eIF4E (Waskiewicz *et al.*, 1997; Pyronnet *et al.*, 1999). The observed link between MAPK/MNK1 and PI3K/mTOR signaling is further supported by the discovery that mTOR inhibition induces eIF4E phosphorylation via PI3K and Mnk (Wang *et al.*, 2007).

In summary, these observations showed that our phenotypic chemical-genetic interaction map captured known connections of signaling pathways. It further revealed new potentially druggable interactions of signaling pathways relevant to cancer.

3.3.4 Prediction of synergistic drug combinations based on cell line specific interactions affecting cell number

As a direct application of harvesting the potentially druggable interactions that we identified, we looked for compound cell line interaction that showed a synthetic lethal interaction affecting the observed cell number. Focusing on drugs that are used in the clinic we found two of such interactions (highlighted as red blocks in Figure 3.18).

First, the DNA-alkylating agent Bendamustine showed a synthetic lethal interaction with the AKT1/2 KO cell line. Bendamustine is approved for the treatment of Chronic Lymphocytic Leukemia (CLL) and it does not show cross-resistance with other alkylating agents (Keating *et al.*, 2008).

Second, the Aldehyde Dehydrogenase inhibitor Disulfiram and the MEK1 KO cell line showed a synthetic lethal interaction. Disulfiram is used in clinic for the treatment of alcoholism (Lövborg *et al.*, 2006).

For both compounds several interactions affecting other features were observed

for the cell line with the corresponding synthetic lethal response. The complete interaction profiles of these drugs across all cell lines are shown in Figure 3.19 and 3.20.

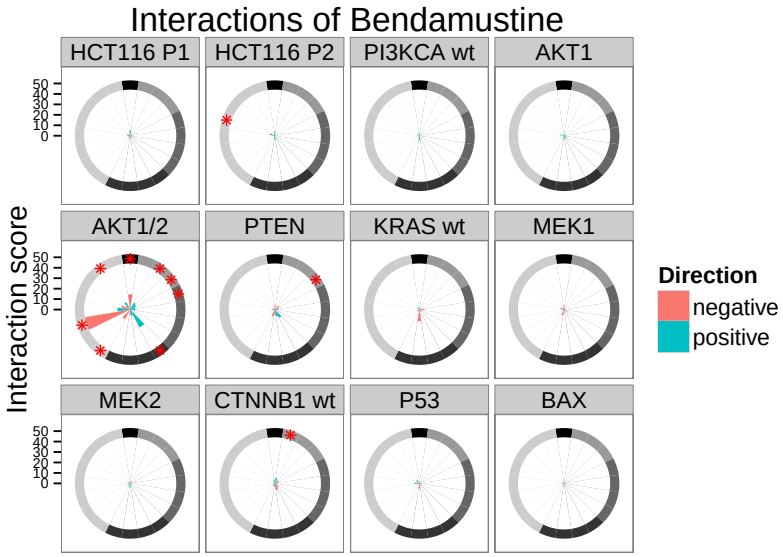


Figure 3.19: Interaction profiles of the drug Bendamustine. Bendamustine has a synthetic lethal interaction with the AKT1/2 KO cell line. (*) adjusted p-value < 0.01, red and turquoise define whether the sign of the interaction coefficient was positive or negative.

To test if these synthetic lethal interactions could be used for improved treatments, we tested for drug synergism of Bendamustine with AKT inhibitors and Disulfiram with MEK inhibitors as described in Section 3.2.6. Indeed, we could show that the combination of Bendamustine with AKT inhibitors significantly reduced cell viability of parental HCT116 cells as shown in Figure 3.21. The same was true for the combination of Disulfiram with MEK inhibitors as shown in Figure 3.22. This observation was consistent for the BI and HSA model used to estimate the synergistic effect. Additionally, we were able to show the synergism of these drug combinations in another colon cancer cell line (DLD-1).

In summary, this demonstrated that our chemical-genetic interaction map allowed us to predict synergistic drug combinations whose discovery would not have been immediately obvious by other means.

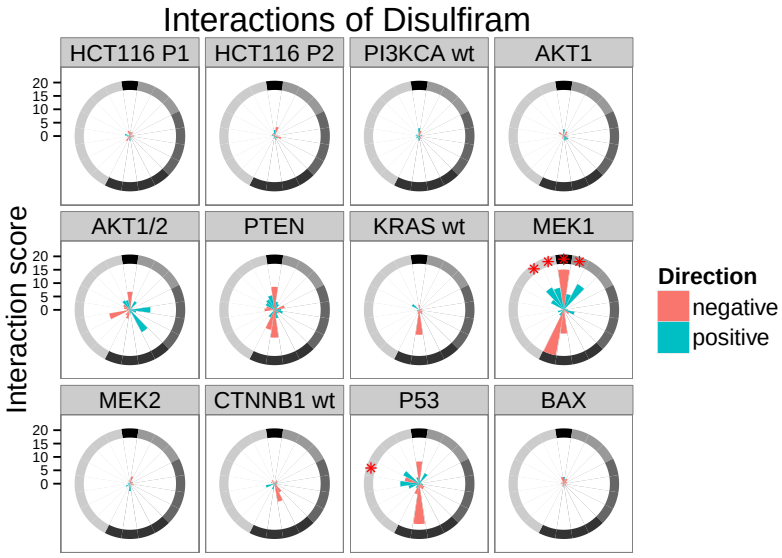


Figure 3.20: Interaction profiles of the drug Disulfiram. Disulfiram has a synthetic lethal interaction with the MEK1 KO cell line. (*) adjusted p-value < 0.01, red and turquoise define whether the sign of the interaction coefficient was positive or negative.

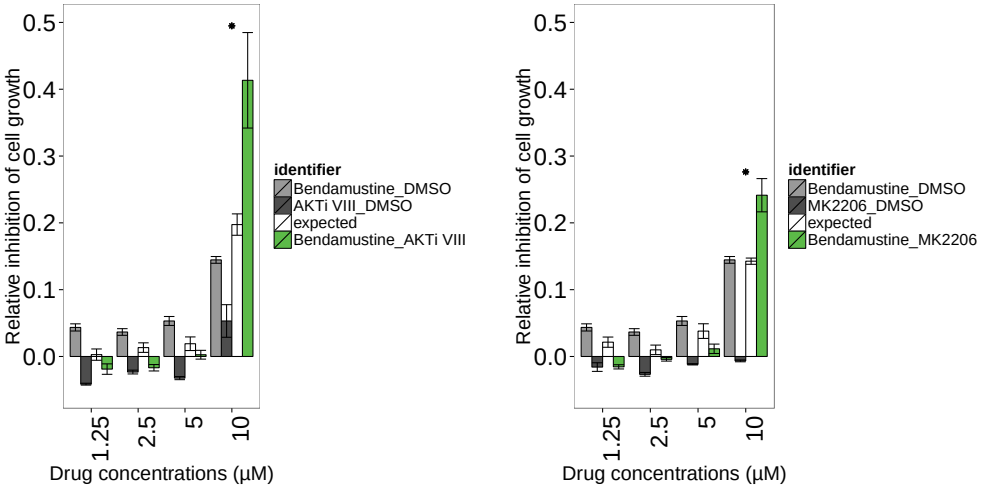


Figure 3.21: Synergistic drug combinations of Bendamustine and AKT inhibitors. Single drug effects, estimated combined effect (BI model) and measured combined effect are shown for several drug concentrations. (*) p-value < 0.05, one-sided t-test.

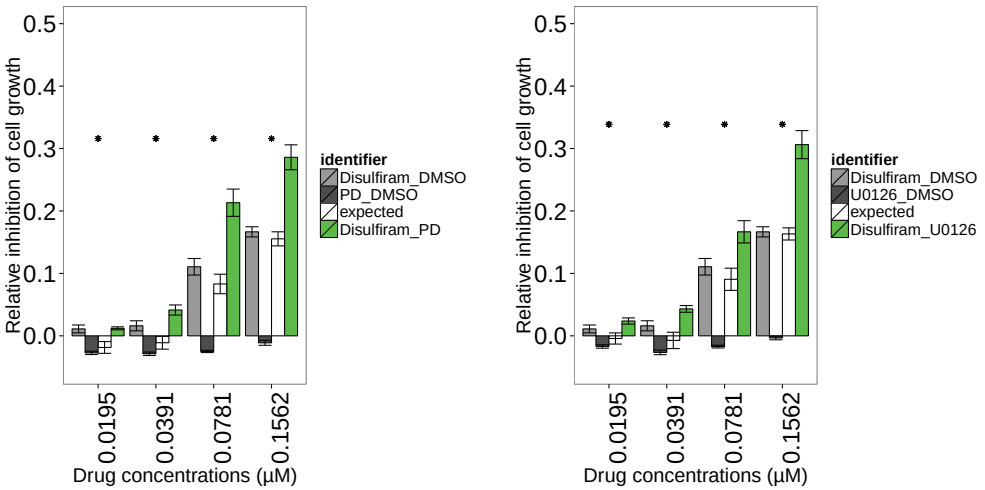


Figure 3.22: Synergistic drug combinations of Disulfiram and MEK inhibitors. Single drug effects, estimated combined effect (BI model) and measured combined effect are shown for several drug concentrations. (*) p-value < 0.05, one-sided t-test.

3.3.5 Discovery of connected biological processes, drug mode-of-action and off-target effects

Our data set could be used to assess information of connected biological processes, drug mode-of-action and off-target effects, by using a *guilt-by-association* approach. These approaches were used previously to characterize compounds, including chemical similarity of compounds (Young *et al.*, 2008), phenotypic similarity of compound responses (Perlman *et al.*, 2004), and compound-gene interaction similarity (Parsons *et al.*, 2006).

I used hierarchical clustering as described in Section 3.2.5 to cluster cell lines and compounds based on their interaction profiles.

3.3.5.1 Clustering cell lines based on their interaction profiles

The distance between cell line interaction profiles was calculated using Equation (3.3) and used for hierarchical clustering. The resulting cluster tree is shown in Figure 3.23.

The two parental cell lines clustered closely together as expected. The cluster tree split into three main branches. Interestingly, the KRAS WT and MEK1 KO cell line clustered together and away from the MEK2 KO cell line, which is consistent with the finding that MEK1 is the essential regulator in the MAPK-signaling cascade (Catalanotti *et al.*, 2009).

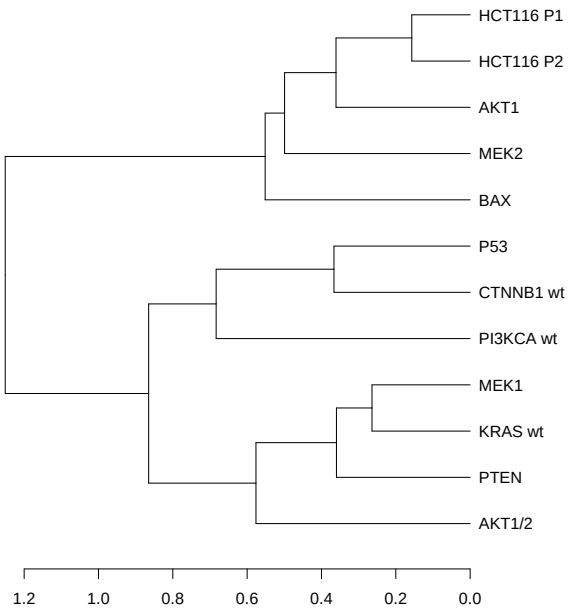


Figure 3.23: Cluster tree from hierarchical clustering of cell lines based on their interaction profiles.

3.3.5.2 Clustering of compounds based on their interaction profiles

To cluster the compounds based on their interaction profile, I calculated the distances between compound interaction profiles using Equation 3.4. Additionally, I calculated the chemical similarity of compounds as described in Section 3.2.5.1. To integrate both sources, the interaction profile similarity and chemical similarity were displayed together in Figure 3.24.

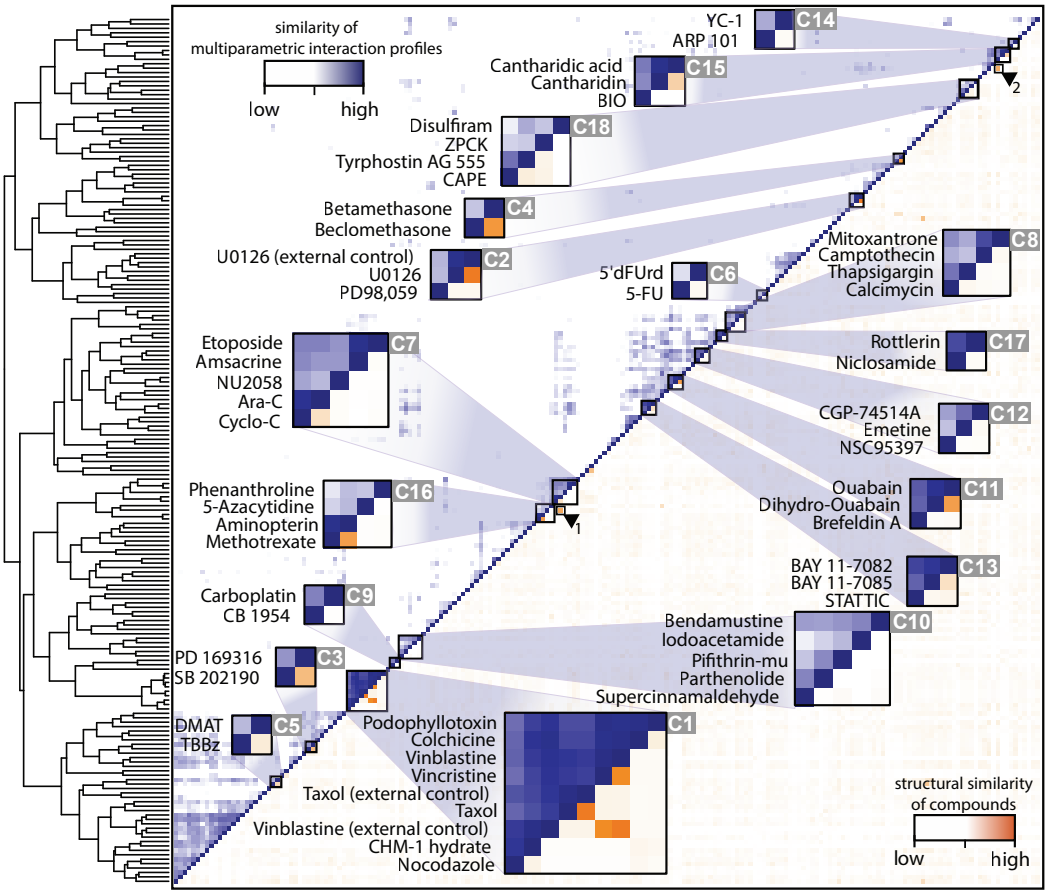


Figure 3.24: Interaction profile similarity and chemical similarity matrix of compounds. Compounds were clustered by hierarchical clustering of the interaction profiles. Above the diagonal the distance between compound interaction profiles is shown. Below the diagonal the distance between Tanimoto coefficients is shown as measure for chemical similarity. Several clusters which were linked to functionally related compounds or biological processes are highlighted. Taken from (Breinig *et al.*, in preperation).

From the exploration of this combined similarity matrix we could draw conclusions about drug mode of action and interlinked biological processes in compound sets that clustered together. The chemical similarity information showed that these sets contained chemically similar as well as divergent compounds. A complete list

of annotated clusters can be found in Table 1 in the Appendix.

For example, we found numerous tubulin targeting compounds to share highly similar interaction profiles (C1, Figure 3.24). This cluster included structurally similar as well as structurally unrelated compounds. Likewise, we observed a cluster of two different MEK inhibitors (C2, Figure 3.24). Hence, our approach identified drugs sharing the same target despite structural divergence.

We could further use the clustering result to predict off-target effects or the primary mode-of-action of compounds. Cluster C17 and C18 were examples for this.

In cluster C17 the compounds niclosamide and rottlerin clustered together. Niclosamide is an anti-helmitic compound, which uncouples oxidative phosphorylation (Weinbach and Garbus, 1969; MacDonald *et al.*, 2006). Rottlerin is classified as a PKC-δ inhibitor, however our data suggested that its primary mode-of-action is linked to uncoupling oxidative phosphorylation. This is in agreement with the results of a recent study (Soltoff, 2001).

Cluster 18 contained Disulfiram, the chymotrypsin inhibitor ZPCK, the EGFR inhibitor tyrphostin AG555 and the NF-κB inhibitor CAPE. A potential link between ZPCK and CAPE is the fact, that NF-κB is regulated via pathway components of the proteasome (Kisselev *et al.*, 2012). It is also known that Disulfiram impairs proteasome activity (Lövborg *et al.*, 2006). This suggests, that these compounds cluster together because they all affect proteasome activity. This has not yet been reported for tyrphostin AG555 which is also in the cluster. To test this hypothesis, we performed an assay in parental HCT116 cells, measuring the caspase-like, chymotrypsin-like, and trypsin-like proteasome activity as described in Section 3.2.7. The result of this assay is shown in Figure 3.25. We observed significant inhibition of the chymotrypsin-like and trypsin-like proteasome activity in cells treated with Disulfiram, ZPCK, tyrphostin AG555, and CAPE. Compared to the established proteasome inhibitors MG132 and bortezomib, the proteasome inhibition was smaller for all compounds of the cluster. Additionally, AG1478 and DAPH, two structurally different EGFR inhibitors did not affect proteasome function. In summary, our results indicate that the EGFR inhibitor tyrphostin AG555 shows an off-target effect by impairing proteasome function.

3.3.5.3 Correlation between interaction profiles of shared drug targets

To test the value of our integrated approach of high-content phenotypic and pharmacogenetic profiling, I generated interaction similarity matrices mimicking other screening methodologies. First, I used only the data set of all features for the parental cell line HCT116 P1, representing a high-content approach with one single cell line, and second the data set of only the cell number of all 12 cell lines, representing a viability screen of a cell line panel. These matrices are shown in Figure 3.26 ordered in the same way as the interaction similarity matrix for the whole data set in the right panel and in Figure 3.24. The delineation of clusters is best for the whole data set and gets noisier for the other approaches. Only the

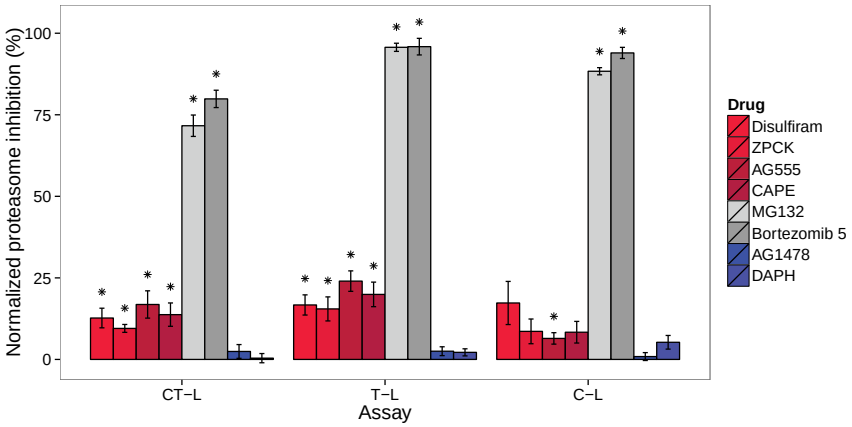


Figure 3.25: Proteasome inhibition upon compound treatment. The Chymotrypsin-like (CT-L), trypsin-like (T-L), and caspase-like (C-L) activity of different compounds was measured. The signals were normalized to DMSO control and corrected for cell viability measured by a CellTiterGlo assay. (*) p-value < 0.05.

tight cluster of tubulin targeting compounds was stable in all approaches.

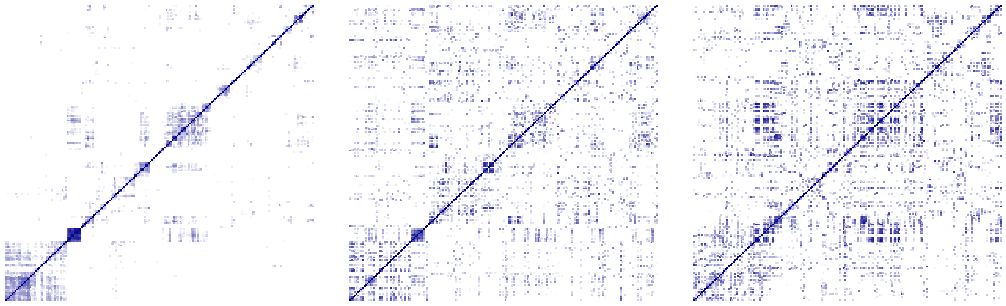


Figure 3.26: Interaction profile similarities for the whole data set (left panel), interaction profile similarity mimicking a high-content screen (all features of the parental HCT116 P1 cell line, middle panel) and a pharmacogenetic screen (cell number feature of all cell lines, right panel)

Next, I tested how well interaction profile similarities agreed with shared target selectivity and chemical similarity as described in Section 3.2.5.2. After classifying compound pairs into the classes of *shared* or *non-shared* selectivity and likewise *similar* and *different* structure, the distribution of correlation values between the compound pairs was investigated. In Figure 3.27 the correlation density is plotted for each class of the target annotation.

The differences in the AUC values calculated from the empirical cumulative distribution function are used to separate the two classes. In Figure 3.28 the result is shown for the separation of the two classes defined by either target selectivity

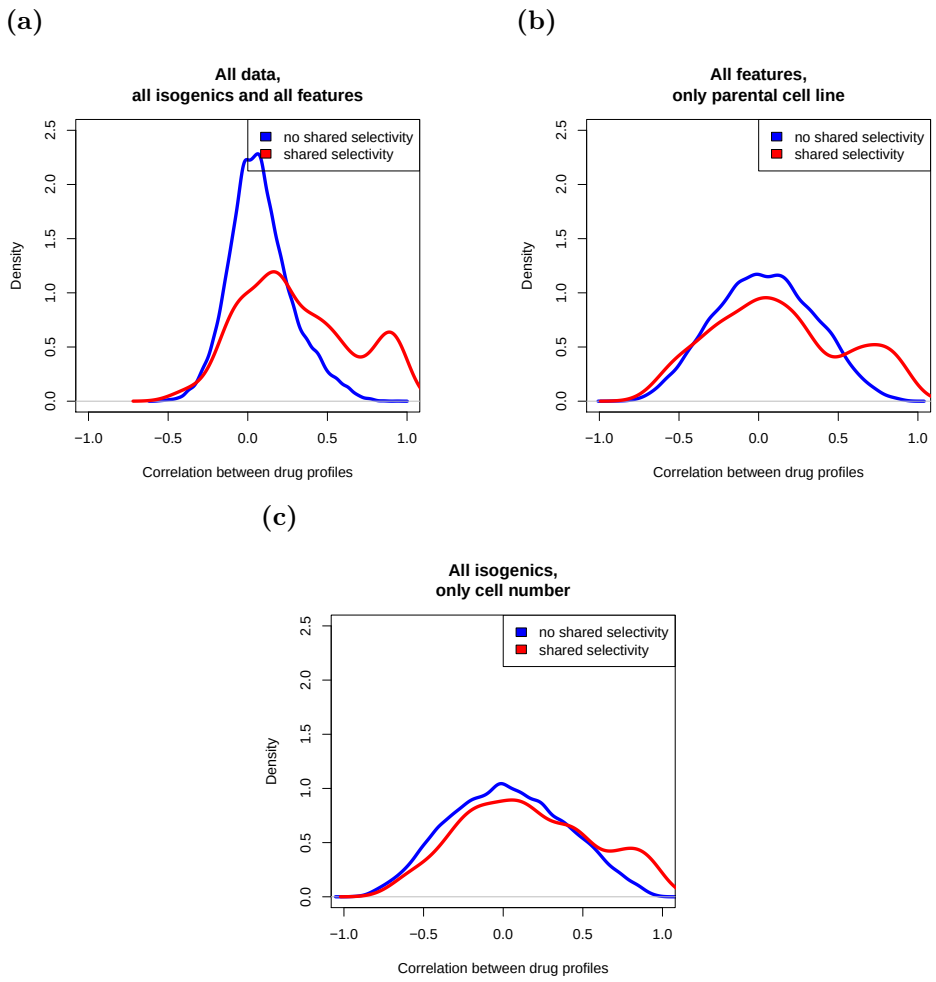


Figure 3.27: The density of correlation coefficients is plotted for the two classes of *shared* or *non-shared* target selectivity. (a) using the whole data set, (b) using only the parental cell line and all features, and (c) using only cell number as a feature for all cell lines.

or chemical similarity.

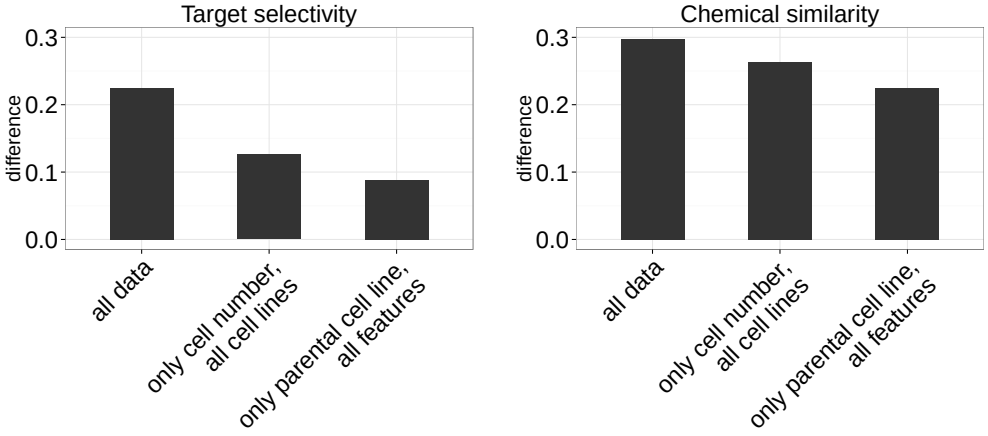


Figure 3.28: The difference in the AUC of the empirical cumulative distribution function between the classes shared target and non-shared target (left panel) and between the classes chemical similar and divergent (right panel).

The results showed that our combined approach was superior to either of the two approaches alone, revealing the added information of our combined approach of integrating high-content phenotypic profiling with pharmacogenetic interaction profiling.

3.4 Discussion

We generated the Pharmacogenetic Phenome Compendium PGPC resource by combining chemical-genetic screening with high-throughput microscopy based phenotypic profiling. Our data set is of high quality and allowed us to identify genotype specific responses to drug treatment, potential drug synergism, connections between signaling pathways, and drug mode-of-action or potential off-target effects. By using more phenotypic features, we greatly expanded the possible interaction space. Notably, some drugs only affected a single phenotypic class. Such interactions would be missed if only cellular fitness was used as readout.

Using our interaction map, we were able to uncover connections between signaling pathways. Understanding pathway structure could be used to predict possible combination therapy. For example, the link between MAPK and PI3K signaling that we observed in our data has been recently reported (Vizeacoumar *et al.*, 2013). The connections between these signaling pathways are important for understanding drug resistance and predicting possible synergistic effects of combined drug treatments (Lehár *et al.*, 2007).

With our data set we predicted drug synergism of Bendamustine and AKT inhibitors on a rational basis and we further validated this link through combinatorial drug screens. AKT inhibitors have recently entered the clinic and Bendamustine is used clinically to treat several cancer types (Keating *et al.*, 2008). Based on the fact that AKT inhibitors and Bendamustine inhibit cell growth through different pathways, a clinical study investigates the combination of Bendamustine and the AKT inhibitor MK2206 in CLL patient (NCT01369849 at ClinicalTrials.gov) (Ding *et al.*, 2014).

Additionally, we predicted and verified a synergistic drug combination of MEK inhibitors and Disulfiram, which is studied in clinical trials as an anti-cancer drug (e.g. studies NCT00742911, NCT01907165, NCT01777919 at ClinicalTrials.gov). The mode-of-action of Disulfiram is not clear and with our approach we showed that Disulfiram inhibits the activity of the proteasome. This finding is in agreement with previous findings, which showed that Disulfiram inhibits 26S proteasome activity (Lövborg *et al.*, 2006).

Along this line we identified a previously unknown off-target effect of the EGFR inhibitor Tyrphostin AG555 on the proteasome. Furthermore, we were able to render drug mode-of-action by hierarchical clustering of interaction profiles. These are important steps in early drug development (Hopkins, 2008; Al-Lazikani *et al.*, 2012). Building on our scalable approach and harvesting the information obtained from further screens will help to improve the development of effective drugs for specific genetic backgrounds as well as synergistic drug combinations at early stages of the development process.

3.4.1 Comparison with other high-throughput drug screens

Genetic interaction screens have already been performed in different eukaryotic systems. In *Saccharomyces cerevisiae* a genome-scale genetic interaction map was created by studying the effect of double mutant fitness compared to the expected fitness estimated from the single mutant effects (Costanzo *et al.*, 2010). Clustering of interaction profiles from this interaction map clustered genes into subgroups that are involved in similar biological processes and revealed connections between different cellular processes (Costanzo *et al.*, 2010). The generation of a chemical-genetic interaction map for 12 compounds and a *Saccharomyces cerevisiae* single deletion library identified genes involved in drug sensitivity (Parsons *et al.*, 2004). Integration of this data with genetic interaction profiles allowed the identification of drug targets or target pathways for the different compounds (Parsons *et al.*, 2004). Another study performed a screen for drug resistance using a yeast deletion library and additionally haploinsufficient strains of the essential yeast genes (Lee *et al.*, 2014). All approaches allowed the researchers to link drug response and sensitivity to specific genes or pathways. A drawback of these approaches is the fact that only colony formation was used as a readout.

For higher eukaryotic systems, the generation of double knockouts in a systematic way has not been possible so far. A complementary approach is the use of double gene knockdown by RNAi to generate genetic interaction maps. Investigation of ricin susceptibility by genetic interaction mapping with double-shRNA constructs recapitulated known pathways and further revealed previously unknown links and functions (Bassik *et al.*, 2013). The readout for the double-shRNA screen is the enrichment of shRNA barcodes in bulk experiments, missing out on the possibility to capture phenotypic changes in individual cells.

Microscopy readouts of combinatorial siRNA treatments are possible using combinatorial siRNA treatments plated onto screening plates by robotics. This approach has been used in mouse fibroblasts, focusing on genes involved in chromatin regulation (Roguev *et al.*, 2013). Roguev *et al.* (2013) showed that similar interaction profiles of genes could be linked to physical interactions of the corresponding proteins and that the interaction map identified connected pathways and complexes.

A similar approach was used to generate a genetic interaction map in human cancer cells by high-throughput imaging and combinatorial RNAi treatment (Laufer *et al.*, 2013). Using multi-parametric interaction mapping Laufer *et al.* (2013) identified known protein complexes.

These studies show how powerful genetic interaction approaches are in higher eukaryotes. They provide a rich data set that can be integrated with chemical-genetic interaction data, like our data set, to identify pathways involved in drug response, drug mode-of-action, and possible off-target effects.

A different approach to study drug resistance or susceptibility is the screening of large cell line panels of different backgrounds against a drug library. This strategy has been employed recently by several groups.

The Cancer Cell Line Encyclopedia (CCLE) integrated gene expression and mutation data with pharmacological profiles of 24 anti-cancer drugs for approximately 500 human cancer cell lines (Barretina *et al.*, 2012). Based on this data set, Barretina *et al.* (2012) identified genetic and gene-expression based predictors of drug sensitivity.

In a similar setup, the Cancer Genome Project (CGP) screened approximately 650 human cancer cell lines from diverse tissues against 130 drugs which are either used already in clinic or are at a preclinical stage (Garnett *et al.*, 2012). Garnett *et al.* (2012) also linked frequently mutated genes and expression profiles with drug sensitivity for a broad range of drugs based on their data.

The Cancer Therapeutics Response Portal (CTRP) is another study that screened 242 well characterized cancer cell lines against a set of 354 small molecules (Basu *et al.*, 2013). This study revealed associations between sensitivity to treatment by small molecules and specific cancer mutations (Basu *et al.*, 2013).

The aim of these studies was to provide a rich database that allows one to find associations between drug sensitivity and specific cancer mutations or cancer types. Testing and verifying these associations will help to find new biomarkers by which patients can be stratified into responsive or non-responsive groups for a given drug treatment.

All these studies use only cellular fitness as a phenotype measured across several concentrations of each drug. The biggest drawback, however, is the small overlap of findings between the different studies. A systematic comparison of the CCLE and CGP study, using common drugs and cell lines, revealed that the correlation between drug responses of cell lines was very low, while the genomic data were highly correlated (Haibe-Kains *et al.*, 2013). To improve the concordance between these large scale studies a more standardized protocol for measuring drug response seems necessary (Haibe-Kains *et al.*, 2013).

One possibility along these lines would be to increase the phenotype space using high-throughput microscopy following our approach. In this way, compound effects can be identified for several phenotypes, possibly improving the consistency between different studies.

3.4.2 Drawbacks and possible improvements of our screening system

In general, all screening approaches have certain drawbacks depending on the specific protocol used for screening. In our approach, we only used one fixed drug concentration. Therefore, interactions that would have been observed at a lower or higher drug concentration might have been missed. The same applies to the single time point used in our screen. Multiple time points would allow insights into the dynamics of drug responses, adding additional information.

Furthermore, we only used two microscopy channels for imaging the cells. With the actin and DNA signal we could already extract information on nuclear and

cellular features. Further improvements could be achieved by making use of additional microscopy channels. For example, certain stress responses or pathway specific responses could be imaged by using corresponding fluorescent markers.

Although the interaction of Ganciclovir and the KRAS WT cell line served as a positive control, it directly showed that remaining fingerprints from the genetic engineering process can influence the chemical genetic interactions observed for the different cell lines. This observed interaction is due to the expression of the Thymidin Kinase rather then the KO of the KRAS mutant allele. Therefore, conclusions drawn from screens with cell lines which contain remaining selection cassettes should be interpreted with care. For future experiments cell lines with minimal fingerprints from genomic engineering should be used to avoid these problems.

3.4.3 Outlook

Recent advances in genome editing have first been driven by Zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs). These techniques offer the possibility to generate site-specific double-strand breaks in the DNA (Gaj *et al.*, 2013). In combination with the endogenous repair mechanisms for DNA double-strand breaks, non-homologous end joining (NHEJ) orhomology-directed repair (HDR), this offers many possibilities for genome editing (Gaj *et al.*, 2013). A more recent genome editing technique makes use of clustered, regularly interspaced, short palindromic repeats (CRISPR) and Cas nucleases to introduce DNA single- or double-strand breaks (Sander and Joung, 2014; Hsu *et al.*, 2014). Using these techniques the generation of genetically modified cell lines becomes much easier and can be performed in higher throughput. This provides the tools to generate cell line panels with well defined mutations in specific genes that can be used for chemical-genetic screening approaches.

The CRISPR-Cas technology has already been employed to perform genome-wide gene knockout screens, which enable screening for positive and negative selection for a given treatment (Wang *et al.*, 2014; Shalem *et al.*, 2014). This screening methodology provides a complementary approach for chemical-genetic screening. Since it makes use of bulk experiments, it will be possible to investigate genetic drug resistance and sensitivity for different drugs using high-throughput approaches in the future.

To better reflect the disease model of interest, the generation of patient-derived induced pluripotent stem cells (iPS cells) seems a promising approach to generate disease models that can be used for high-throughput screening. First steps in this direction have been taken, e.g., for studying cardiovascular disease (Matsa *et al.*, 2014). Patient-derived iPS cells are already comercially available and used as a model system for drug screening and testing for cell toxicity of candidate drugs (Oksana *et al.*, 2014). Furthermore, if the disease can be causally linked to a mutation, the genomic loci containing this mutation could be corrected or introduced by the previously mentioned genetic engineering approaches in patient-

derived iPS cells. This provides the possibility of screens with paired study designs of disease and WT alleles which might be particularly informative.

The generation of cell lines carrying variants or mutations that were identified by GWAS approaches would also allow for better functional characterization and identification of the responsible variant.

A combination of the mentioned approaches and building up a resource of high quality microscopy images for various drug screens in different disease or toxicity models will be a valuable resource for further research and drug discovery. Integrating the data from such a resource in early drug development will improve the development process. Based on the data, toxicities could be identified and possibly circumvented at an early stage. Genotype-specific responses should be tested additionally and could be used for better patient stratification in clinical trials, leading to more successful outcomes.

Combining all the mentioned approaches and possibilities will help to improve patient-stratified medicine, where patients will receive specific drug treatments based on their individual genotype.

Chapter 4

Conclusions

Science has always been driven by the development of new technologies. Nowadays, in the genomic era, high-throughput studies using automated microscopy or sequencing technology are powerful tools to further enhance our understanding of fundamental biological processes and how the are connected. Today, more and more of these studies are performed, generating a huge amount of data. Handling and analyzing these data has become a challenge. Ever improving experimental protocols require constant adaptation and development of new methods to analyze the generated data. This has led to many collaborative projects in which experimentalists and data analysts work together to tackle these challenges.

After working on such collaborative projects which used high-throughput microscopy and sequencing technologies, I provided a comprehensive description of the biological and technological background and described my contributions to the analysis and results of these projects. I discussed the challenges and methods required for the analysis of the different data types and which conclusions can be drawn from them. In summary, this dissertation provides a good overview of the potential of the respective projects and which hypotheses can be addressed by similar projects.

A ongoing problem with large scale high-throughput studies is the lack of reproducibility for the results. This is due to missing data or analysis tools, which are not publicly provided. To address this problem, open source software and publishing of the data and analysis pipelines is becoming more and more important. For example, the Bioconductor platform provides a great open source tool set in the **R** programming environment for the analysis of high-throughput experiments. To address these needs, I developed the two **R** packages `FourCSeq` and `PGPC` which both include vignettes that can be used to rerun the analysis pipelines and reproduce the results.

A current challenge is the integration of multivariate data from different technology platforms, which will become more and more important to further understand biological processes and networks. Good examples of these studies are the large Pan-Cancer studies, such as the TCGA (http://cancergenome.nih.gov), where different cancer samples are profiled using different high-throughput tech-

nologies and integrated with clinical data to improve the understanding of the molecular basis of different cancer types.

Having worked a lot on chromosome conformation capture data, I think that especially its integration with live cell super-resolution microscopy data will be important to obtain a more complete picture of the dynamics involved in gene regulation.

Other currently fast developing fields with a strong need for analysis tools are metagenomics, single-cell technologies and super-resolution microscopy. The latter two provide the possibility to access the cell to cell variability and hence will need a kind of probabilistic description in the future to describe the biology of the observed states.

Bibliography

Adams, D., Karolak, M., Robertson, E., and Oxburgh, L. (2007). Control of kidney, eye and limb expression of Bmp7 by an enhancer element highly conserved between species. *Developmental Biology*, **311**(2), 679–90.

Al-Lazikani, B., Banerji, U., and Workman, P. (2012). Combinatorial drug therapy for cancer in the post-genomic era. *Nature Biotechnology*, **30**(7), 679–92.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular Biology of the Cell*. Garland Science; 5 edition (November 28, 2007).

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Developmental Cell*, **16**(1), 47–57.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, **21**(3), 381–95.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A., and Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell*, **144**(2), 214–26.

Barbie, D. a., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. a., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T., and Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**(7269), 108–12.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. a., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa,

A., Jané-Valbuena, J., Mapa, F. a., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., and Garraway, L. a. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**(7391), 603–7.

Bassik, M. C., Kampmann, M., Lebbink, R. J., Wang, S., Hein, M. Y., Poser, I., Weibezahn, J., Horlbeck, M. a., Chen, S., Mann, M., Hyman, A. a., Leproust, E. M., McManus, M. T., and Weissman, J. S. (2013). A systematic Mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell*, **152**(4), 909–22.

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., Ebright, R. Y., Stewart, M. L., Ito, D., Wang, S., Bracha, A. L., Liefeld, T., Wawer, M., Gilbert, J. C., Wilson, A. J., Stransky, N., Kryukov, G. V., Dancik, V., Barretina, J., Garraway, L. a., Hon, C. S.-Y., Munoz, B., Bittker, J. a., Stockwell, B. R., Khabele, D., Stern, A. M., Clemons, P. a., Shamji, A. F., and Schreiber, S. L. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, **154**(5), 1151–61.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**(1), 289–300.

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**(5705), 2242–6.

Bliesath, J., Huser, N., Omori, M., Bunag, D., Proffitt, C., Streiner, N., Ho, C., Siddiqui-Jain, A., O'Brien, S. E., Lim, J. K., Ryckman, D. M., Anderes, K., Rice, W. G., and Drygin, D. (2012). Combined inhibition of egfr and ck2 augments the attenuation of pi3k-akt-mtor signaling and the killing of cancer cells. *Cancer Letters*, **322**(1), 113 – 118.

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczynski, B., Riddell, A., and Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, **44**(2), 148–56.

Borisy, A. a., Elliott, P. J., Hurst, N. W., Lee, M. S., Lehar, J., Price, E. R., Serbedzija, G., Zimmermann, G. R., Foley, M. a., Stockwell, B. R., and Keith, C. T. (2003). Systematic discovery of multicomponent therapeutics. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(13), 7977–82.

Branco, M. R. and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biology*, **4**(5), e138.

Breinig, M., Klein, F. A., Huber, W., and Boutros, M. (in preperation). High-content phenotyping and epistasis analysis to predict mode-of-action and synergisms of drugs in cancer cells.

Bridges, C. B. (1922). The origin of variation. *American Naturalist*, **56**, 51–63.

Bryant, H. E., Schultz, N., Thomas, H. D., Parker, K. M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N. J., and Helleday, T. (2005). Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature*, **434**(7035), 913–7.

Campos-Ortega, J. and Hartenstein, V. (1985). *The embryonic development of Drosophila melanogaster*. Springer-Verlag, Berlin, Heidelberg.

Cao, Y., Charisi, A., Cheng, L.-C., Jiang, T., and Girke, T. (2008). ChemmineR: a compound mining framework for R. *Bioinformatics (Oxford, England)*, **24**(15), 1733–4.

Catalanotti, F., Reyes, G., Jesenberger, V., Galabova-Kovacs, G., de Matos Simoes, R., Carugo, O., and Baccarini, M. (2009). A Mek1-Mek2 heterodimer determines the strength and duration of the Erk signal. *Nature Structural & Molecular Biology*, **16**(3), 294–303.

Chambeyron, S. and Bickmore, W. a. (2004). Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & Development*, **18**(10), 1119–30.

Chan, D. a. and Giaccia, A. J. (2011). Harnessing synthetic lethal interactions in anticancer drug discovery. *Nature Reviews Drug Discovery*, **10**(5), 351–64.

Chazaud, C., Oulad-Abdelghani, M., Bouillet, P., Décimo, D., Chambon, P., and Dollé, P. (1996). AP-2.2, a novel gene related to AP-2, is expressed in the forebrain, limbs and face during mouse embryogenesis. *Mechanisms of Development*, **54**(1), 83–94.

Conrad, C. and Gerlich, D. W. (2010). Automated microscopy for high-content RNAi screening. *The Journal of Cell Biology*, **188**(4), 453–61.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. a., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. a., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A.-C., Morris, Q. D., Kim, P. M., Kaiser, C. a., Myers, C. L., Andrews, B. J., and Boone, C. (2010). The genetic landscape of a cell. *Science*, **327**(5964), 425–31.

Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, **2**(4), 292–301.

Crumpacker, C. S. (1996). Ganciclovir. *New England Journal of Medicine*, **335**(10), 721–729. PMID: 8786764.

Danesh, S. M., Villasenor, A., Chong, D., Soukup, C., and Cleaver, O. (2009). BMP and BMP receptor expression during murine organogenesis. *Gene Expression Patterns*, **9**(5), 255–65.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W., and Steinmetz, L. M. (2006). A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(14), 5320–5.

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., Gibbons, R. J., Vernimmen, D., Yoshinaga, Y., de Jong, P., Cheng, J.-F., Rubin, E. M., Wood, W. G., Bowden, D., and Higgs, D. R. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science*, **312**(5777), 1215–7.

de Laat, W. and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, **502**(7472), 499–506.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, **295**(5558), 1306–11.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P., Dean, A., and Blobel, G. (2012). Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. *Cell*, **149**(6), 1233–1244.

Di Maira, G., Salvi, M., Arrigoni, G., Marin, O., Sarno, S., Brustolon, F., Pinna, L. a., and Ruzzene, M. (2005). Protein kinase CK2 phosphorylates and upregulates Akt/PKB. *Cell Death and Differentiation*, **12**(6), 668–77.

Ding, W., Shanafelt, T. D., Lesnick, C. E., Erlichman, C., Leis, J. F., Secreto, C., Sassoon, T. R., Call, T. G., Bowen, D. a., Conte, M., Kumar, S., and Kay, N. E. (2014). Akt inhibitor MK2206 selectively targets CLL B-cell receptor induced cytokines, mobilizes lymphocytes and synergizes with bendamustine to induce CLL apoptosis. *British Journal of Haematology*, **164**(1), 146–50.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**(7398), 376–80.

Dostie, J., Richmond, T. a., Arnaout, R. a., Selzer, R. R., Lee, W. L., Honan, T. a., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, **16**(10), 1299–309.

Drissen, R., Palstra, R.-J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S., and de Laat, W. (2004). The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes & Development*, **18**(20), 2485–90.

Fatica, A. and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, **15**(1), 7–21.

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**(1).

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E., and Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**(7269), 58–64.

Gaj, T., Gersbach, C. a., and Barbas, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, **31**(7), 397–405.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R. J., Bignell, G. R., Tam, A. T., Davies, H.,

Stevenson, J. a., Barthorpe, S., Lutz, S. R., Kogera, F., Lawrence, K., McLaren-Douglas, A., Mitropoulos, X., Mironenko, T., Thi, H., Richardson, L., Zhou, W., Jewitt, F., Zhang, T., O'Brien, P., Boisvert, J. L., Price, S., Hur, W., Yang, W., Deng, X., Butler, A., Choi, H. G., Chang, J. W., Baselga, J., Stamenkovic, I., Engelman, J. a., Sharma, S. V., Delattre, O., Saez-Rodriguez, J., Gray, N. S., Settleman, J., Futreal, P. A., Haber, D. a., Stratton, M. R., Ramaswamy, S., McDermott, U., and Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**(7391), 570–5.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, **512**(7512), 96–100.

Gostissa, M., Yan, C. T., Bianco, J. M., Cogné, M., Pinaud, E., and Alt, F. W. (2009). Long-range oncogenic activation of Igh-c-myc translocations by the Igh 3' regulatory region. *Nature*, **462**(7274), 803–7.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. a., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., Sturgill, D., Tuch, B. B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R. E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J. E., Wan, K. H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P. J., Brenner, S. E., Brent, M. R., Cherbas, P., Gingeras, T. R., Hoskins, R. a., Kaufman, T. C., Oliver, B., and Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, **471**(7339), 473–9.

Gregor, T., Garcia, H. G., and Little, S. C. (2014). The embryo as a laboratory: quantifying transcription in Drosophila. *Trends in Genetics*, **30**(8), 364–375.

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J. W. L., and Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. *Nature*, **504**(7480), 389–93.

Hartman, J., Garvik, B., and Hartwell, L. (2001). Principles for the buffering of genetic variation. *Science*, **291**, 1001–4.

Hausser, J. and Zavolan, M. (2014). Identification and consequences of miRNA-target interactions - beyond repression of gene expression. *Nature Reviews Genetics*, **15**(9), 599–612.

Helder, M. N., Ozkaynak, E., Sampath, K. T., Luyten, F. P., Latin, V., Oppermann, H., and Vukicevic, S. (1995). Expression pattern of osteogenic protein-1 (bone morphogenetic protein-7) in human and mouse development. *Journal of Histochemistry & Cytochemistry*, **43**(10), 1035–1044.

Hérault, Y., Fraudeau, N., Zákány, J., and Duboule, D. (1997). Ulnaless (Ul), a regulatory mutation inducing both loss-of-function and gain-of-function of posterior Hoxd genes. *Development (Cambridge, England)*, **124**(18), 3493–500.

Hérault, Y., Rassoulzadegan, M., Cuzin, F., and Duboule, D. (1998). Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nature Genetics*, **20**(4), 381–4.

Hindorff, L. a., Sethupathy, P., Junkins, H. a., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. a. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(23), 9362–7.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, **4**(11), 682–90.

Horn, T., Sandmann, T., Fischer, B., Axelsson, E., Huber, W., and Boutros, M. (2011). Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nature Methods*, **8**(4), 341–6.

Hsu, P. D., Lander, E. S., and Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell*, **157**(6), 1262–1278.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18 (Suppl.**(1997).

Huber, W., Toedling, J., and Steinmetz, L. M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics (Oxford, England)*, **22**(16), 1963–70.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. a., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, pages 1–5.

Jones, T. R., Carpenter, A., and Golland, P. (2005). Voronoi-based segmentation of cells on image manifolds. In *Proceedings of the First International Conference on Computer Vision for Biomedical Image Applications*, CVBIA'05, pages 535–543, Berlin, Heidelberg. Springer-Verlag.

Kaelin, W. G. (2005). The concept of synthetic lethality in the context of anticancer therapy. *Nature Rreviews Cancer*, **5**(9), 689–98.

Keating, M. J., Bach, C., Yasothan, U., and Kirkpatrick, P. (2008). Bendamustine. *Nature Reviews Drug Discovery*, **7**(6), 473–4.

Kisselev, A. F., van der Linden, W. a., and Overkleeft, H. S. (2012). Proteasome inhibitors: an expanding army attacking a unique target. *Chemistry & Biology*, **19**(1), 99–115.

Klein, F. A., Anders, S., Pakozdi, T., Ghavi-Helm, Y., Furlong, E. E. M., and Huber, W. (submitted for publication). FourCSeq: Analysis of 4C sequencing data.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**(3), R25.

Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013). Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nature Methods*, **10**(5), 427–31.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., and Carey, V. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, **9**.

Lee, A. Y., St Onge, R. P., Proctor, M. J., Wallace, I. M., Nile, A. H., Spagnuolo, P. a., Jitkova, Y., Gronda, M., Wu, Y., Kim, M. K., Cheung-Ong, K., Torres, N. P., Spear, E. D., Han, M. K. L., Schlecht, U., Suresh, S., Duby, G., Heisler, L. E., Surendra, A., Fung, E., Urbanus, M. L., Gebbia, M., Lissina, E., Miranda, M., Chiang, J. H., Aparicio, A. M., Zeghouf, M., Davis, R. W., Cherfils, J., Boutry, M., Kaiser, C. a., Cummins, C. L., Trimble, W. S., Brown, G. W., Schimmer, A. D., Bankaitis, V. a., Nislow, C., Bader, G. D., and Giaever, G. (2014). Mapping the cellular response to small molecules using chemogenomic fitness signatures. *Science*, **344**(6180), 208–11.

Lehár, J., Zimmermann, G. R., Krueger, A. S., Molnar, R. a., Ledell, J. T., Heilbut, A. M., Short, G. F., Giusti, L. C., Nolan, G. P., Magid, O. a., Lee, M. S., Borisy, A. a., Stockwell, B. R., and Keith, C. T. (2007). Chemical combination effects predict connectivity in biological systems. *Molecular Systems Biology*, **3**(80), 80.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, **12**(14), 1725–1735.

Lettice, L. a., Williamson, I., Wiltshire, J. H., Peluso, S., Devenney, P. S., Hill, A. E., Essafi, A., Hagman, J., Mort, R., Grimes, G., DeAngelis, C. L., and Hill, R. E. (2012). Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. *Developmental Cell*, **22**(2), 459–67.

Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(14), 4936–42.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. a., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. a., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**(5950), 289–93.

Lövborg, H., Oberg, F., Rickardson, L., Gullbo, J., Nygren, P., and Larsson, R. (2006). Inhibition of proteasome activity, nuclear factor-KappaB translocation and cell survival by the antialcoholism drug disulfiram. *International Journal of Cancer*, **118**(6), 1577–80.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*.

Lower, K. M., Hughes, J. R., De Gobbi, M., Henderson, S., Viprakasit, V., Fisher, C., Goriely, A., Ayyub, H., Sloane-Stanley, J., Vernimmen, D., Langford, C., Garrick, D., Gibbons, R. J., and Higgs, D. R. (2009). Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(51), 21771–6.

Lucas, J. S., Zhang, Y., Dudko, O. K., and Murre, C. (2014). 3D Trajectories Adopted by Coding and Regulatory DNA Elements: First-Passage Times for Genomic Interactions. *Cell*, **158**(2), 339–352.

Luo, J., Emanuele, M. J., Li, D., Creighton, C. J., Schlabach, M. R., Westbrook, T. F., Wong, K.-K., and Elledge, S. J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell*, **137**(5), 835–48.

MacDonald, M. L., Lamerdin, J., Owens, S., Keon, B. H., Bilter, G. K., Shang, Z., Huang, Z., Yu, H., Dias, J., Minami, T., Michnick, S. W., and Westwick, J. K. (2006). Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nature Chemical Biology*, **2**(6), 329–37.

Mani, R., St Onge, R. P., Hartman, J. L., Giaever, G., and Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(9), 3461–6.

Maskey, D., Yousefi, S., Schmid, I., Zlobec, I., Perren, A., Friis, R., and Simon, H.-U. (2013). ATG5 is induced by DNA-damaging agents and promotes mitotic catastrophe independent of autophagy. *Nature Communications*, **4**, 2130.

Matsa, E., Burridge, P. W., and Wu, J. C. (2014). Human Stem Cells for Modeling Heart Disease and for Drug Discovery. *Science Translational Medicine*, **6**(239), 239ps6.

Mirny, L. a. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, **19**(1), 37–51.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., De Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls hox genes transcription in digits. *Cell*, **147**(5), 1132–1145.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**(5881), 1344–9.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**(7469), 59–64.

Neumann, B., Walter, T., Hériché, J.-K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., Cetin, C., Sieckmann, F., Pau, G., Kabbe, R., Wünsche, A., Satagopam, V., Schmitz, M. H. a., Chapuis, C., Gerlich, D. W., Schneider, R., Eils, R., Huber, W., Peters, J.-M., Hyman, A. a., Durbin, R., Pepperkok, R., and Ellenberg, J. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, **464**(7289), 721–7.

Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., Eussen, B., de Klein, A., Singer, R. H., and de Laat, W. (2011). Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature Cell Biology*, **13**(8), 944–951.

Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., and Duboule, D. (2014). Temporal dynamics and developmental memory of 3d chromatin architecture at hox gene loci. *eLife*.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**(7398), 381–5.

Northcott, P. a., Lee, C., Zichner, T., Stütz, A. M., Erkek, S., Kawauchi, D., Shih, D. J. H., Hovestadt, V., Zapatka, M., Sturm, D., Jones, D. T. W., Kool, M., Remke, M., Cavalli, F. M. G., Zuyderduyn, S., Bader, G. D., VandenBerg, S., Esparza, L. A., Ryzhova, M., Wang, W., Wittmann, A., Stark, S., Sieber, L., Seker-Cin, H., Linke, L., Kratochwil, F., Jäger, N., Buchhalter, I., Imbusch, C. D., Zipprich, G., Raeder, B., Schmidt, S., Diessl, N., Wolf, S., Wiemann,

S., Brors, B., Lawerenz, C., Eils, J., Warnatz, H.-J., Risch, T., Yaspo, M.-L., Weber, U. D., Bartholomae, C. C., von Kalle, C., Turányi, E., Hauser, P., Sanden, E., Darabi, A., Siesjö, P., Sterba, J., Zitterbart, K., Sumerauer, D., van Sluis, P., Versteeg, R., Volckmann, R., Koster, J., Schuhmann, M. U., Ebinger, M., Grimes, H. L., Robinson, G. W., Gajjar, A., Mynarek, M., von Hoff, K., Rutkowski, S., Pietsch, T., Scheurlen, W., Felsberg, J., Reifenberger, G., Kulozik, A. E., von Deimling, A., Witt, O., Eils, R., Gilbertson, R. J., Korshunov, A., Taylor, M. D., Lichter, P., Korbel, J. O., Wechsler-Reya, R. J., and Pfister, S. M. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*.

Oksana, S., Jayne, H., Ivan, R., and F., C. E. (2014). High-content assays for hepatotoxicity using induced pluripotent stem cell-derived cells. *ASSAY and Drug Development Technologies*, **12**(1), 43 – 54.

Parada, L. a., McQueen, P. G., and Misteli, T. (2004). Tissue-specific spatial organization of genomes. *Genome biology*, **5**(7), R44.

Parsons, A. B., Brost, R. L., Ding, H., Li, Z., Zhang, C., Sheikh, B., Brown, G. W., Kane, P. M., Hughes, T. R., and Boone, C. (2004). Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature Biotechnology*, **22**(1), 62–9.

Parsons, A. B., Lopez, A., Givoni, I. E., Williams, D. E., Gray, C. a., Porter, J., Chua, G., Sopko, R., Brost, R. L., Ho, C.-H., Wang, J., Ketela, T., Brenner, C., Brill, J. a., Fernandez, G. E., Lorenz, T. C., Payne, G. S., Ishihara, S., Ohya, Y., Andrews, B., Hughes, T. R., Frey, B. J., Graham, T. R., Andersen, R. J., and Boone, C. (2006). Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell*, **126**(3), 611–25.

Pau, G., Fuchs, F., Sklyar, O., Boutros, M., and Huber, W. (2010a). EBImage - an R package for image processing with applications to cellular phenotypes. *Bioinformatics*, **26**(7), 979–981.

Pau, G., Zhang, X., Boutros, M., and Huber, W. (2010b). *imageHTS: Analysis of high-throughput microscopy-based screens*. Bioconductor Package.

Pelechano, V. and Steinmetz, L. M. (2013). Gene regulation by antisense transcription. *Nature Reviews Genetics*, **14**(12), 880–93.

Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., Wu, L. F., and Altschuler, S. J. (2004). Multidimensional drug profiling by automated microscopy. *Science*, **306**(5699), 1194–8.

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. a., Guo, C., Sun, Y., Bland, M. J., Wagstaff, W., Dalton, S., McDevitt, T. C., Sen, R., Dekker, J., Taylor, J., and

Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**(6), 1281–95.

Pyronnet, S., Imataka, H., Gingras, a. C., Fukunaga, R., Hunter, T., and Sonenberg, N. (1999). Human eukaryotic translation initiation factor 4G (eIF4G) recruits mnk1 to phosphorylate eIF4E. *The EMBO Journal*, **18**(1), 270–9.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2014). *fda: Functional Data Analysis*. R package version 2.4.3.

Riddle, R. D., Johnson, R. L., Laufer, E., and Tabin, C. (1993). Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell*, **75**(7), 1401–16.

Roguev, A., Talbot, D., Negri, G. L., Shales, M., Cagney, G., Bandyopadhyay, S., Panning, B., and Krogan, N. J. (2013). Quantitative genetic-interaction mapping in mammalian cells. *Nature Methods*, **10**(5), 432–7.

Ruf, S., Symmons, O., Uslu, V., and Dolle, D. (2011). Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nature Genetics*.

Sander, J. D. and Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, **32**(4), 347–55.

Sanyal, A., Baù, D., Martí-Renom, M. a., and Dekker, J. (2011). Chromatin globules: a common motif of higher order chromosome structure? *Current Opinion in Cell Biology*, **23**(3), 325–31.

Saunders, A., Core, L. J., Sutcliffe, C., Lis, J. T., and Ashe, H. L. (2013). Extensive polymerase pausing during Drosophila axis patterning enables high-level and pliable transcription. *Genes & Development*, **27**(10), 1146–58.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**(3), 458–72.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. a., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**(6166), 84–7.

Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer, New York.

Soltoff, S. P. (2001). Rottlerin is a mitochondrial uncoupler that decreases cellular ATP levels and indirectly blocks protein kinase Cdelta tyrosine phosphorylation. *The Journal of Biological Chemistry*, **276**(41), 37986–92.

Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, **13**(9), 613–26.

Spitz, F., Herkenne, C., Morris, M. a., and Duboule, D. (2005). Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes. *Nature Genetics*, **37**(8), 889–93.

Splinter, E., de Wit, E., van de Werken, H. J. G., Klous, P., and de Laat, W. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods (San Diego, Calif.)*, **58**(3), 221–30.

Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R.-J., Wendt, K. S., Grosveld, F., van Ijcken, W., and Soler, E. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature Protocols*, **8**(3), 509–24.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, **24**(3), 390–400.

Symmons, O., Pan, L., Remeseiro, S., Klein, F. A., Aktas, T., Huber, W., and Spitz, F. (in preperation).

Tan, X., Hu, L., Luquette, L. J., Gao, G., Liu, Y., Qu, H., Xi, R., Lu, Z. J., Park, P. J., and Elledge, S. J. (2012). Systematic identification of synergistic drug pairs targeting HIV. *Nature Biotechnology*, **30**(11), 1125–30.

Tanimoto, T. (1957). An Elementary Mathematical Theory of Classification and Prediction. *Internal IBM Technical Report*.

Thongjuea, S., Stadhouders, R., Grosveld, F. G., Soler, E., and Lenhard, B. (2013). r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Research*, **41**(13), e132–e132.

Torres, J. and Pulido, R. (2001). The tumor suppressor PTEN is phosphorylated by the protein kinase CK2 at its C terminus. Implications for PTEN stability to proteasome-mediated degradation. *The Journal of Biological Chemistry*, **276**(2), 993–8.

Tsujimura, T., Klein, F. A., Langenfeld, K., Glaser, J., Huber, W., and Spitz, F. (submitted for publication). Topological and regulatory autonomy of the adjacent Tfap2c and Bmp7 genes in vivo.

van de Werken, H. J. G., Landan, G., Holwerda, S. J. B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Oz, Y., Bouwman, B. a. M., Verstegen, M. J. a. M., de Wit, E., Tanay, A., and de Laat, W. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods*, **9**(10), 969–72.

Vernimmen, D., De Gobbi, M., Sloane-Stanley, J. a., Wood, W. G., and Higgs, D. R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *The EMBO Journal*, **26**(8), 2041–51.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. a. (2007). VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Research*, **35**(Database issue), D88–92.

Vizeacoumar, F. J., Arnold, R., Vizeacoumar, F. S., Chandrashekhar, M., Buzina, A., Young, J. T. F., Kwan, J. H. M., Sayad, A., Mero, P., Lawo, S., Tanaka, H., Brown, K. R., Baryshnikova, A., Mak, A. B., Fedyshyn, Y., Wang, Y., Brito, G. C., Kasimer, D., Makhnevych, T., Ketela, T., Datti, A., Babu, M., Emili, A., Pelletier, L., Wrana, J., Wainberg, Z., Kim, P. M., Rottapel, R., O'Brien, C. a., Andrews, B., Boone, C., and Moffat, J. (2013). A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. *Molecular Systems Biology*, **9**(696).

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**(6166), 80–4.

Wang, X., Yue, P., Chan, C.-B., Ye, K., Ueda, T., Watanabe-Fukunaga, R., Fukunaga, R., Fu, H., Khuri, F. R., and Sun, S.-Y. (2007). Inhibition of mammalian target of rapamycin induces phosphatidylinositol 3-kinase-dependent and Mnk-mediated eukaryotic translation initiation factor 4E phosphorylation. *Molecular and Cellular Biology*, **27**(21), 7405–13.

Waskiewicz, a. J., Flynn, a., Proud, C. G., and Cooper, J. a. (1997). Mitogen-activated protein kinases activate the serine/threonine kinases Mnk1 and Mnk2. *The EMBO Journal*, **16**(8), 1909–20.

Weinbach, E. C. and Garbus, J. (1969). Mechanism of Action of Reagents that uncouple Oxidative Phosphorylation. *Nature*, **221**(5185), 1016–18.

Young, D. W., Bender, A., Hoyt, J., McWhinnie, E., Chirn, G.-W., Tao, C. Y., Tallarico, J. A., Labow, M., Jenkins, J. L., Mitchison, T. J., and Feng, Y. (2007).

Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nature Chemical Biology*, **4**(1), 59–68.

Young, D. W., Bender, A., Hoyt, J., McWhinnie, E., Chirn, G.-W., Tao, C. Y., Tallarico, J. a., Labow, M., Jenkins, J. L., Mitchison, T. J., and Feng, Y. (2008). Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nature Chemical Biology*, **4**(1), 59–68.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, **38**(11), 1341–7.

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**(7269), 65–70.

Zuin, J., Dixon, J. R., van der Reijden, M. I. J. a., Ye, Z., Kolovos, P., Brouwer, R. W. W., van de Corput, M. P. C., van de Werken, H. J. G., Knoch, T. a., van IJcken, W. F. J., Grosveld, F. G., Ren, B., and Wendt, K. S. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(3), 996–1001.

Appendix

Table of clusters from the drug-cell line interaction profiling.

The following table was taken from (Breinig *et al.*, in preperation).

Table 1: Table of clusters from the drug-cell line interaction profiling.

Cluster	**Compounds**	**Target selectivity / Biological process**	**Associated references**
C1	Podophyllotoxin Colchicine Vinblastine Vincristine Taxol CHM-1 hydrate Nocodazol	microtubule	-
C2	U0126 PD98,059	MEK1/2	-
C3	PD169316 SB202190	p38 MAPK	-
C4	Betamethasone Beclomethasone	steroidal anti-inflammatory glucocorticoids	-
C5	DMAT TBBz	CK2	-
C6	5'dFUrd 5-FU	DNA/RNA metabolism	(Lum et al., 2004) this study used yeast chemo-genomics and identified that 5-FU preferentially interferes with RNA metabolism

C7	Etoposide	Topoisomerase	(Iorio et al., 2010) this study used transcriptional profiling and identified a link between CDK2 inhibitors and topoisomerase inhibitors;
	Amsacrine	Topoisomerase	(Pourquier et al., 2000) this study identified a link between ara-c and topoisomerase
	NU2058	CDK2	
	Ara-c	DNA metabolism	
	Cyclo-c	DNA metabolism	
C8	Mitoxantrone	Topoisomerase	(Bertrand et al., 1991) this study identified links between calcium levels and topoisomerase inhibitor activity
	Camptothecin	Topoisomerase	
	Thapsigargin	Increased intracellular calcium levels (SERCA inhibitor)	
	Calcimycin	Increased intracellular calcium levels (Ca ionophore)	
C9	Carboplatin	DNA alkylating	-
	CB1954		
C10	Bendamustine	DNA alkylating	(Leoni et al., 2008) this study showed that bendamustine has a different mode-of-action as compared to standard DNA alkylating agents;
	Iodoacetamine	Alkylating	(Strom et al., 2006) this study identified pifithrin-μ as a compound that interferes with p53 signaling;
	Pifithrin-μ	P53	(Kwok et al., 2001) this study showed that parthenolide interferes with IKKbeta due to modifying cysteine residues;
	Parthenolide	IKKbeta, Cysteine modifying	(Macpherson et al., 2007) this study identified compounds, including supercinnamaldehyde and iodoacetamide, that interfere with TRPA1 due to modifying cysteine residues
	Supercinnamaldehyde	TRPA1, Cysteine modifying	

C11	Ouabain	Na/K ATPase	(Farr et al., 2009) this study showed Golgi mediated transport of Na/K pump subunits to the plasma membrane
	Dihydro-Ouabain	Na/K ATPase	
	Brefeldin	Golgi/ER	
C12	CGP-74514A	CDK1	(Boutros et al., 2007) this review summarizes the interplay between CDK1, cdc25, and translational control during the cell cycle
	Emetine	Translation	
	NSC95297	Cdc25	
C13	BAY11-7082	IKB-α	(Grivennikov and Karin, 2010) this review summarizes the links between NFκB signaling and JAK/STAT signaling
	BAY11-7085	IKB-α	
	STATTIC	STAT3	
C14	YC-1	guanylyl cyclase activator	(Lubbe et al., 2009) this study showed a link between guanylyl cyclase activity and MMPs
	ARP101	MMP2	
C15	Cantharidic acic	PP2A	(Hernandez et al., 2010) this study showed a link between PP2 and GSK3
	Cantharidin	PP2A	
	BIO	GSK3 > CDKs	
C16	Phenanthroline	Iron chelator	(Oppenheim et al., 2000)(Crider et al., 2012) this study and review summarize links between iron metabolism, folate metabolism and DNA methylation
	5-azacytidine	DNA methyltransferase	
	Aminopterin	Folate metabolism/DHFR	
	Methotrexate	Folate metabolism/DHFR	

C17	Rottlerin Niclosamide	PKCδ Oxidative phosphorylation	(MacDonald et al., 2006)(Maioli et al., 2012) this study and review highlight that rotterlin does not inhibit PKC and interferes with oxidative phosphorylation; (Weinbach and Garbus, 1969) this study showed that niclosamide uncouples oxidative phosphorylation
C18	Disulfiram ZPCK Tyrphostatin AG555 CAPE	ALDH Chymotrypsin-A EGFR NFκB	(Kisselev et al., 2012)(Grivennikov and Karin, 2010) these reviews address proteasome functions and NFκB signaling
Arrow 1	Ara-c 5-Azacytidine	DNA metabolism DNA methyltransferase	(Christman, 2002) this review summarizes the DNA methyltransferase activity of 5-azacytidine
Arrow 2	BIO Indirubin-3-oxim	GSK3 > CDKs CDKs > GSK3	(Meijer et al., 2003) this study revealed that indirubin derivatives can inhibit GSK3 as well as CDKs and showed that BIO preferentially inhibits GSK3 whereas indirubin-3-oxime preferentially inhibits CDKs

References in table:

Bertrand, R., Kerrigan, D., Sarang, M., and Pommier, Y. (1991). Cell death induced by topoisomerase inhibitors. Role of calcium in mammalian cells. Biochem. Pharmacol. 42, 77–85.

Boutros, R., Lobjois, V., and Ducommun, B. (2007). CDC25 phosphatases in cancer cells: key players? Good targets? Nat. Rev. Cancer 7, 495–507.

Christman, J.K. (2002). 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. Oncogene 21, 5483–5495.

Crider, K.S., Yang, T.P., Berry, R.J., and Bailey, L.B. (2012). Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. Adv. Nutr. Bethesda Md 3, 21–38.

Farr, G.A., Hull, M., Mellman, I., and Caplan, M.J. (2009). Membrane proteins follow multiple pathways to the basolateral cell surface in polarized epithelial cells. J. Cell Biol. 186, 269–282.

Grivennikov, S.I., and Karin, M. (2010). Dangerous liaisons: STAT3 and NF-kappaB collaboration and crosstalk in cancer. Cytokine Growth Factor Rev. 21, 11–19.

Hernandez, F., Langa, E., Cuadros, R., Avila, J., and Villanueva, N. (2010). Regulation of GSK3 isoforms by phosphatases PP1 and PP2A. Mol. Cell. Biochem. 344, 211–215.

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A., et al. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc. Natl. Acad. Sci. U. S. A. 107, 14621–14626.

Kisselev, A.F., van der Linden, W.A., and Overkleeft, H.S. (2012). Proteasome inhibitors: an expanding army attacking a unique target. Chem. Biol. 19, 99–115.

Kwok, B.H., Koh, B., Ndubuisi, M.I., Elofsson, M., and Crews, C.M. (2001). The anti-inflammatory natural product parthenolide from the medicinal herb Feverfew directly binds to and inhibits IkappaB kinase. Chem. Biol. 8, 759–766.

Leoni, L.M., Bailey, B., Reifert, J., Bendall, H.H., Zeller, R.W., Corbeil, J., Elliott, G., and Niemeyer, C.C. (2008). Bendamustine (Treanda) displays a distinct pattern of cytotoxicity and unique mechanistic features compared with other alkylating agents. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 14, 309–317.

Lubbe, W.J., Zuzga, D.S., Zhou, Z., Fu, W., Pelta-Heller, J., Muschel, R.J., Waldman, S.A., and Pitari, G.M. (2009). Guanylyl cyclase C prevents colon cancer metastasis by regulating tumor epithelial cell matrix metalloproteinase-9. Cancer Res. 69, 3529–3536.

Lum, P.Y., Armour, C.D., Stepaniants, S.B., Cavet, G., Wolf, M.K., Butler, J.S., Hinshaw, J.C., Garnier, P., Prestwich, G.D., Leonardson, A., et al. (2004). Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. Cell 116, 121–137.

MacDonald, M.L., Lamerdin, J., Owens, S., Keon, B.H., Bilter, G.K., Shang, Z., Huang, Z., Yu, H., Dias, J., Minami, T., et al. (2006). Identifying off-target effects and hidden phenotypes of drugs in human cells. Nat. Chem. Biol. 2, 329–337.

Macpherson, L.J., Dubin, A.E., Evans, M.J., Marr, F., Schultz, P.G., Cravatt, B.F., and Patapoutian, A. (2007). Noxious compounds activate TRPA1 ion channels through covalent modification of cysteines. Nature 445, 541–545.

Maioli, E., Torricelli, C., and Valacchi, G. (2012). Rottlerin and cancer: novel evidence and mechanisms. ScientificWorldJournal 2012, 350826.

Meijer, L., Skaltsounis, A.-L., Magiatis, P., Polychronopoulos, P., Knockaert, M., Leost, M., Ryan, X.P., Vonica, C.A., Brivanlou, A., Dajani, R., et al. (2003). GSK-3-selective inhibitors derived from Tyrian purple indirubins. Chem. Biol. 10, 1255–1266.

Oppenheim, E.W., Nasrallah, I.M., Mastri, M.G., and Stover, P.J. (2000). Mimosine is a cell-specific antagonist of folate metabolism. J. Biol. Chem. 275, 19268–19274.

Pourquier, P., Takebayashi, Y., Urasaki, Y., Gioffre, C., Kohlhagen, G., and Pommier, Y. (2000). Induction of topoisomerase I cleavage complexes by 1-β-d-arabinofuranosylcytosine (ara-C) in vitro and in ara-C-treated cells. Proc. Natl. Acad. Sci. 97, 1885–1890.

Strom, E., Sathe, S., Komarov, P.G., Chernova, O.B., Pavlovska, I., Shyshynova, I., Bosykh, D.A., Burdelya, L.G., Macklis, R.M., Skaliter, R., et al. (2006). Small-molecule inhibitor of p53 binding to mitochondria protects mice from gamma radiation. Nat. Chem. Biol. 2, 474–479.

Weinbach, E.C., and Garbus, J. (1969). Mechanism of action of reagents that uncouple oxidative phosphorylation. Nature 221, 1016–1018.
default default

List of Figures

List of Tables

Acknowledgments

At the end of my PhD thesis I would like to express my gratitude to all the people who contributed to make my PhD at EMBL a great time for me.

First of all, I would like to thank Wolfgang Huber for giving me the opportunity to do my PhD in his research group, for his support and feedback on my projects, and the good scientific input and guidance.

Special thanks go to Paul Bertone and Caroline Friedel for agreeing to be the first and second examiner of my PhD, and for their feedback during my TAC meetings. I also thank Lars Steinmetz for being on my TAC committee.

I would like to acknowledge Michael Boutros and Nassos Typas for joining my defense committee.

Thanks to all my fellow group members, I always had a great time with you guys, especially at our retreats. Greg, thanks a lot for helping me to get started at the beginning of my PhD. My special thanks goes to Bernd Fischer for helping me with the analysis of the microscopy interaction screens and his constant feedback. I also thank Aleks and Bernd Klaus for the discussions on the 4C and HiC projects. Thanks a lot Simon for the critical feedback on my method paper and your support to push it out. I would like to acknowledge Julian for introducing me to emacs and org-mode. A special thanks to Joe for proofreading my thesis.

I would like to acknowledge the support and feedback from all my collaborators. Thanks to Yad, Tibor, and Eileen for constant feedback on my analysis and the possibility to work on this project. I am thankful for the nice mouse projects with Taro, Silvia, and Orsolya from the Spitz group and the great feedback and comments by François Spitz. Thanks to Marco and Michael for the high-throughput microscopy projects and insights into the new developments of screening technologies.

Furthermore, I thank all my fellow PhD students from my year for the great time during the Predoc course, PhD retreats, and other occasions.

I would like to acknowledge the EC FP7 projects Systems Microscopy (NoE) and Radiant for providing financial support of my Ph.D.

Finally, special thanks go to my brother Moritz, and my parents Beate and Karl-Friedrich for their never-ending encouragement and ongoing support.

Last and foremost, I thank my wife Carolina for giving me key feedback on writing my thesis, proofreading my thesis, and supporting and motivating me in the tough phases of my PhD.

Publications

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, **512**(7512), 96–100.

Klein, F. A., Anders, S., Pakozdi, T., Ghavi-Helm, Y., Furlong, E. E. M., and Huber, W. (submitted for publication). FourCSeq: Analysis of 4C sequencing data.

Tsujimura, T., Klein, F. A., Langenfeld, K., Glaser, J., Huber, W., and Spitz, F. (submitted for publication). Topological and regulatory autonomy of the adjacent Tfap2c and Bmp7 genes in vivo.

Symmons, O., Pan, L., Remeseiro, S., Klein, F. A., Aktas, T., Huber, W., and Spitz, F. (in preperation).

Breinig, M., Klein, F. A., Huber, W., and Boutros, M. (in preperation). High-content phenotyping and epistasis analysis to predict mode-of-action and synergisms of drugs in cancer cells.