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

Sascha Meiers
MSc Bioinformatcs
born in Merzig, Germany

Date of oral examination: 14.05.2018

Exploiting emerging DNA sequencing technologies to study genomic rearrangements

Referees:

Dr. Judith Zaugg
Prof. Dr. Benedikt Brors

Exploiting emerging DNA sequencing technologies to study genomic rearrangements

Sascha Meiers

14th March 2018

Supervised by Dr. Jan Korbel

Summary

Structural variants (SVs) alter the structure of chromosomes by deleting, duplicating or otherwise rearranging pieces of DNA. They contribute the majority of nucleotide differences between humans and are known to play causal roles in many diseases. Since the advance of massively parallel sequencing (MPS) technologies, SVs have been studied more comprehensively than ever before. However, in contrast to smaller types of genetic variation, SV detection is still fundamentally hampered by the limitations of short-read sequencing that cannot sufficiently cope with the complexity of large genomes. Emerging DNA sequencing technologies and protocols hold the potential to overcome some of these limitations. In this dissertation, I present three distinct studies each utilizing such emerging techniques to detect, to validate and/or to characterize SVs. These technologies, together with novel computational approaches that I developed, allow to characterize SVs that had previously been challening, or even impossible, to assess.

First, inversions—a class of SV that is notoriously difficult to ascertain—were studied in the context of the 1000 Genomes Project. These inversions had previously been predicted from low-coverage short read sequencing data, but remained inconclusive in classical PCR validation experiments. Using sequencing data from modern long-read technologies (Pacific BioSciences and Oxford Nanopore Technologies), I was able to validate hundreds of them. I then developed a computational tool to visualize long-read data, and discovered that the majority of loci harbored complex SVs rather than simple inversions. These findings suggest that the amount of complex structural variation in the human genome had so far been under-appreciated, owing to limitations in their detection using standard techniques.

In the second part, I explored the functional impact of large SVs on gene expression and chromatin organization. Previously, a series of studies described drastic effects of SVs on the regulation of genes via mechanisms that alter the three-dimensional conformation of DNA. However, these studies had focused on pathological phenotypes and on few, selected genes. We hence set out to study gene expression and chromatin conformation in highly rearranged chromosomes

of *Drosophila melanogaster* without a pathological phenotype. I first utilized Hi-C, which we applied in order to measure chromatin conformation, to characterize the rearrangements present in these chromosomes. Then, despite the presence of 15 breakpoints, we found no evidence for a conformation-related mechanism acting on gene regulation. This is particularly surprising as the majority of these breakpoints disrupted topologically associating domains. This study hence sheds a new light on the role of chromatin conformation that is complementary to the findings of previous studies. In addition, it demonstrates the capabilities of the Hi-C technology to reveal structural variation.

Third, I present the current state of a collaborative effort to enable SV detection in single cells. Studies of somatic mosaicism, i.e. on the genetic heterogeneity among cells, have so far been severly limited in the ability to discover SVs, especially copy-neutral and complex rearrangements. We hence conceived a novel method to infer—for the first time—at least seven different SV classes in single cells. This approach utlizes three independent signals that are identifiable in single-cell stranded template sequencing (Strand-seq) data. I here present a computational method called MOSAICATCHER to realize this idea and provide examples that demonstrate its feasibility. In order to explore the limitations of this method, I designed a versatile framework for the simulation of Strand-seq data and used it to assess the performance of one of the central steps of MOSAICATCHER. Once completed, this novel method will facilitate studies of SV heterogeneity and mosaicism in the context of cancer and ageing.

In summary, I utilized emerging technoogies to discover SVs—notably copy-neutral and complex rearrangements—that so far eluded detection based on MPS. This led to novel insights on the complexity and functional impact of these SVs. Moreover, I developed computational tools that advance our capabilites for SV detection and characterization, and that might aid future studies to gain a deeper understanding of the role of SVs in health and disease.

Zusammenfassung

Strukturvariationen (SV) verändern die Chromosomenstruktur, in dem sie Teile der DNA deletieren, duplizieren oder anderweitig neu anordnen. Sie sind für den Großteil der Unterschiede in der Nukleotidsequenz zwischen Menschen verantwortlich und sind kausal für verschiedene Erkrankungen. Seit dem Vormarsch von hochparallelen Sequenziermethoden (*massively parallel sequencing*, MPS) können SV umfassender denn je untersucht werden. Dennoch leidet die Detektion von SV, im Gegensatz zu jener von kleineren Formen genetischer Variation, unter den Limitierungen der Sequenzierung kurzer DNA Abschnitte, welche die Komplexität großer Genome nicht ausreichend abbilden kann. Neu aufkommende DNA-Sequenziermethoden und –protokolle bergen das Potenzial diese Limitierungen zu überwinden. In dieser Dissertation werden drei separate Studien präsentiert, die aufkommende Technologien zur Detektion, Validierung und/oder Charakterisierung von SV nutzen. Im Verbund mit neuartigen Computermethoden, die ich entwickelt habe, erlauben diese Technologien die Bestimmung von SV die zuvor schwierig, wenn nicht gar unmöglich, war.

Zunächst werden Inversionen—eine bekanntermaßen schwierig zu ermittelnde Form von SV—im Rahmen des "1000 Genomes Project" untersucht. Diese Inversionen wurden zuvor basierend auf MPS-Methoden mit niedriger Abdeckung vorhergesagt, blieben aber in klassischen PCR-basierten Validierungsexperimenten ergebnislos. Mithilfe moderner Sequenziertechnologie für besonders lange DNA-Abschnitte von Pacific BioSciences und Oxford Nanopore Technologies konnte ich hunderte davon validieren. Ich entwickelte dann eine neue Methode zur Visualisierung dieser langen DNA-Abschnitte und fand heraus, dass die Mehrzahl der Loci anstatt Inversionen letztendlich komplexe Formen von SV enthielten. Diese Entdeckung legt nahe, dass die Menge an komplexen SV im menschlichen Genom aufgrund der Limitierungen der verwendeten Technologien bisher unterschätzt wurde.

Im zweiten Teil untersuche ich den funktionellen Einfluss von großen SV auf die Genexpression und die Chromatinorganisation. Zuvor hatten eine Reihe von Studien drastische Effekte von SV auf die Genexpression gezeigt, welche über einen Mechanismus funktionieren, der die dreidimensionale Form der DNA be-

einflusst. Diese Studien fokussierten sich jedoch nur auf wenige Loci mit bekannten pathologischen Konsequenzen. Deshalb entwickelten wir ein eigene Studie, in der wir die Genexpression und Veränderungen der Chromatinstrukur in völlig neu angeordneten Chromosomen der Fruchtfliege *Drosophila melanogaster* ohne pathologischen Phänotyp untersuchen. Ich nutzte zunächst die Hi-C Technologie, die wir zur Messung der Chromatinstruktur verwendeten, um die neue Anordnung dieser Chromosomen zu bestimmen. Danach fanden wir, obwohl 15 Bruchpunkte vorhanden waren, keine Hinweise auf einen funktionellen Einfluss der Chromatinstruktur. Dies war besonders überraschend, da die Mehrzahl dieser Bruchpunkte sogenannte *topologically associating domains* zu unterbrechen scheint. Unsere Studie wirft somit ein völlig neues Licht auf die Rolle der Chromatinstruktur und ergänzt damit vorherige Studien. Zusätzlich dazu wird in dieser Studie eindrucksvoll demonstriert, wie SV mithilfe der Hi-C Technologie genauer bestimmt werden können.

Drittens wird der aktuelle Stand eines gemeinsamen Unterfanges vorgestellt, das zum Ziel hat die Detektion von SV in einzelnen Zellen zu ermöglichen. Studien zu somatischem Mosaizismus, d.h. zur genetischen Heterogenität zwischen Zellen, konnten aufgrund technologischer Limitierungen bisher nur eingeschränkt SV analysieren, vor allem nicht SV mit neutraler Kopienzahl oder komplexe Veränderungen. Wir konzipierten deshalb eine neuartige Methode die, zum ersten mal überhaupt, mindestens sieben verschiedene SV-Klassen in einzelnen Zellen ermitteln kann. Diese Methode basiert auf drei unabhänigen Signalen die in der Einzelstrang-Sequenziermethode Strand-seq identifiziert werden können. Hier stelle ich eine Software mit dem Namen MOSAICATCHER vor, die dieses Konzept umsetzt, und belege anhand von Beispielen deren Machbarkeit. Um die Grenzen der Methode bestimmen zu können entwickelte ich ein vielseitig einsetzbares Simulationsprogramm für Strand-seq-Daten und untersuchte damit die Leistungsfähigkeit von einem der zentralen Bausteine von MOSAICATCHER. Nach Fertigstellung könnte diese Arbeit Studien zu somatischem Mosaizismus, zum Beispiel im Kontext des Alterns oder von Krbeserkrankungen, vorantreiben.

In dieser Dissertation habe ich aufkommende Technologien benutzt um SV, vor allem komplexe oder jene mit neutraler Kopienzahl, zu bestimmen, die mithilfe bisheriger Technologien nicht auffindbar waren. Dies führte zu neuen Erkenntnissen zur Komplexität und zur funktionalen Rolle dieser SV. Darüberhinaus habe ich Computermethoden entwichelt, die unsere Fähigkeiten zur Bestimmung von SV erweitern und die zukünftige Studien zur Rolle dieser SV in Hinsicht auf Krankheiten ermöglichen.

Acknowledgements

First and foremost, thanks are due to my supervisor Jan Korbel, who made all of this possible. His excitement for research and his encouraging and constructive feedback have always been a great motivation to pursue my work. Thanks go also to my thesis advisors Judith Zaugg, Wolfgang Huber and Benedikt Brors, who often shared their ideas and suggestions during and outside of our regular meetings.

One of the most exciting aspects of my work was to closely collaborate with a number of amazing people. To proceed chronologically, I would like to thank Tobias Rausch, Markus Hsi-Yang Fritz, Andreas Untergasser and Adrian Stütz for our awesome team effort in the 1000 Genomes Project. Especially Adrian, the most critical and thorough old-school wet lab biologist I know, for looking through the 30 thousand plots that I created. Next, I am very happy that I had the opportunity to cooperate with Yad Ghavi-Helm and Aleksander Jankowski on our balancer project. For me, this study is the prime example of synergy resulting from our different backgrounds, and I learned a lot about research (and about flies) from both of them. I am also grateful to Eileen Furlong, who initiated, supervised, and funded this project. At last, I am quite excited about being part of the "Strand-seq nation". We have been pursuing exciting research together and generally have a lot of fun during our phone calls and hackathons! Probs to Ashley Sanders, Tobias Marschall, David Porubský and Maryam Ghareghani, but also the other "citizens" including Karen Grimes, Hyobin Jeong, and Venla Kinanen, my Master student.

Then, there are other people who influenced my professional development more than they might be aware. First of all, Tobias and Markus must be nominated in this category, with whom I spent my first months at EMBL and who I have been looking up to for their skill and their modesty to not make a big deal about it. I don't know how often I went downstairs with a question, but I always came back a little bit wiser and a little bit more relaxed. Another great source of inspiration was Sebastian Waszak, with whom I discussed many scientific and non-scientific topics, be it at EMBL or at P11. I would also like to thank David Garfield, who I often asked for counsel during the fly project and who always

gave me plenty of food for thought.

Next, I would like to acknowledge the efforts of EMBL's IT department, including GBCS, to providing and constantly improving the infrastructure for my research. Moreover, I thank those people who commented on this thesis, notably Jan Korbel, Ashley Sanders, Nina Habermann, Aleksander Jankowski and Jonas Ibn-Salem.

My time at EMBL was a thrilling, stimulating and fun experience and would not have been the same without all the friends around me. I would hence like to say thank you to all the members of the Korbel lab, including past members (Alexandros, Chris, Balca and many more), for the good times we had in the lab, downtown and on our retreats. A special thanks to Nina for lifting all the organizational burdens from our shoulders and always having an open door (figuratively, as we sit in the same room). I can also count myself lucky for having a nice batch of fellow PhD students around me, including Jessi, Jørgen, Lukas, Mariana, Marvin, and Sourabh. I am much looking forward to our next joint holiday trip!

Der größte Dank gilt jedoch meiner Familie. Auch wenn meine Eltern bis heute nicht ganz verstehen, was ich hier genau mache, haben sie immer an mich geglaubt—danke dafür! Ebenso bin ich meinen Schwiegereltern dankbar, die mich stets unterstützt haben. Die Unternehmungen mit der ganzen Familie waren nicht immer erholsam, aber ein schöner und jederzeit willkommener Ausgleich.

Zuletzt und gleichzeitig allen voran danke ich Lena und Merle. Für die tolle Zeit die war, und die, die ganz sicher kommen wird.

Contents

List of Abbreviations

ASE: allele-specific expression
BAF: B allele frequency
BIR: break-induced replication
CNV: copy number variant
CyO: Curly of Oster
dm6: the *Drosophila melanogaster* reference genome version 6
FDR: false discovery rate
FoSTeS: fork stalling and template switching
HMM: Hidden Markov Model
MEI: mobile element insertion
MMBIR: micro-homology-mediated break-induced replication
MPS: massively parallel sequencing
NAHR: non-allelic homologous recombination
NB: negative binomial
NHEJ: non-homologous end joining
PCR: polymerase chain reaction
SCE: sister chromatid exchange
SNV: single nucleotide variant
SV: structural variant
TAD: topologically associating domain
VCF: variant call format
WGS: whole-genome sequencing

List of Figures

List of Tables

General Introduction

1

In his famous "On the origin of species [...]" from 1859, Charles Darwin noted the outstanding variety of individuals within a species. Especially domesticated species, as he describes, show remarkable differences—imagine a bloodhound, terrier, spaniel, and bull-dog next to one another. But also in nature such differences occur. These difference can be passed on to offspring over generations and accumulate to such an extent that a distinct species is formed. It was this observation as well as a great amount of preceding research by him and others that led Darwin to the formulation of his famous evolutionary theory.

More than 150 years later, we have understood many of the molecular mechanisms behind this principle. We know that genetic information is encoded in DNA and that it is subject to mutations, which are partly inheritable. These changes in DNA create the variability of genetic material that is so essential to evolution. Today, we are able to study the genetic differences, which we also call genetic *variants*, between human individuals or between individuals of other species and can investigate their consequences. Many fundamental associations between the presence of genetic variants and certain traits have been found since, for example why red-green color blindness affects more males than females [Nathans et al., 1986]. Other traits, such as the expected height of a person, are less easily conceived as they appear to be affected by the combination of many genetic variants [Wood et al., 2014; Marouli et al., 2017]. We also gained a much better understanding of the causal role of genetic variation in many diseases, including Mendelian disorders and cancer [Stankiewicz et al., 2010].

In order to study the consequences of genetic variants we require methods to accurately detect them. Nowadays, this has been largely enabled with the advance of DNA sequencing technologies. Based on these methods, the average genome of many species, including humans, could be charted for the first time [International Human Genome Sequencing Consortium, 2001; Venter, 2001], and

functional units such as genes and regulatory elements were identified in great detail [ENCODE Project Consortium, 2012]. Then, population genetics studies gained further insight into the variability within the human population [1000 Genomes Project Consortium, 2015; Sudmant, Rausch, et al., 2015]. At last, the identified variants could be analyzed for a potential effect on phenotypes or disease using, for example, genome-wide association studies [Ott et al., 2015; MacArthur et al., 2017].

Variant detection has become a standard procedure in genomics research. However, not all types of genetic variants have been studied equally comprehensively. Especially larger genomic rearrangments, so-called *structural variants* (which are introduced in Section 1.2), remained difficult to ascertain using available assays, which is why their functional role is still not as well explored as for smaller types of genetic variation. However, structural variants are known to have extraordinary impact on our genetic material—after all, they constitute the majority of genetic differences within the human population [Sudmant, Rausch, et al., 2015]. The limitations of the current standard techniques had long been noticed and encouraged further method development [Onishi-Seebacher et al., 2011]. Over the last couple of years, new technologies for DNA sequencing as well as new method based on DNA sequencing have become available. These techniques hold promise to improve the abilities to study such rearrangements. In this work, I utilize such emerging technologies to characterize structural variants beyond what was possible previously.

In the rest of this chapter, I explain the relevant terminology around genetic variation (Section 1.1), introduce structural variants and the biology behind them (Section 1.2), give an overview of previous and emerging DNA sequencing technologies (Section 1.3), and outline the methodology of SV detection (Section 1.4). I continue to elaborate the current limitations of structural variant detection (Section 1.4.3). Then, I formulate the goals of my research and provide an overview of the studies covered in this dissertation (Section 1.5).

1.1. Terminology around genetic variation

Alterations in the DNA of an organism, or of a cell, can arise spontaneously via chemical or biological processes. If not repaired faithfully, they leave traces in the genetic material that we call genetic *variants*. The process of altering the DNA seqeunce is called *mutation*, but the terms mutation and variant are often used interchangeably. We already know the nucleotide sequence of the genomes of

many species, including humans. This sequence, which is supposed to represent the an average of the individuals, is stored in a so-called *reference* genome (also *reference assembly*, or simply *reference*). Thus, variants are usually defined as a difference to this reference. Some variants only affect a single nucleotide, e.g. by changing a cytosine into a thymine, and they are called *single nucleotide variants (SNVs).* Others variants delete or insert nucleotides, or fully rearrange their order, as will be introduced later.

Most metazoa are diploid, so they carry two non-identical copies of each chromosome in their cells: one of maternal and one of paternal origin. Any new variant (within a cell or organism) typically arises only on one of the two homologous chromosomes. Such a variant is said to be *heterozygous*. The genomic locus harboring this variant exists in two different versions, which we call *alleles*. Specifically it harbors a *reference allele*, which is in agreement with the reference assembly, and an *alternative allele* that describes the non-reference variant form. A site with exactly two alleles seen across a population is termed *bi-allelic*, but there are also sites that contain multiple different alleles and are thus *multi-allelic*. Chromosomes can contain many variants and, depending on the detection strategies, it is often unclear which variants reside on the same homologue. If this is known, we refer to the ensemble of variants present along a single homologue as a haploid genotype, or short *haplotype*.

metazoa *Animals. Not all animals are diploid, though. Bees and ants, for example, produce fully haploid males, yet still diploid females. Mammals, on the other hand, are believed to always be diploid [Svartman et al., 2005]*

Alleles that are present in the germ line, i.e. in cells carrying inheritable genetic material, can be propagated to offspring. This way, an individual can end up carrying the same variant of a genomic locus on both homologues, which makes it a *homozygous* variant. Variants that are seen more often in a population, specifically in at least 1% of the homologoues, are also called *polymorphisms*.

When a variant is present in an individual, but not in their parents, we call it a *de novo* variant (or *de novo* mutation). The mutation could either have occurred within the zygote during the first few cell divisions, or already beforehand in the parental germ line. The latter can sometimes be inferred if other offspring of those parents carries the same variant. Variants that occur not in the germ line but in cells of the non-inheritable part of an organism are *somatic* variants. When such an affected cell undergoes repeated divisions, a somatic variant can be present in a relevant fraction of the total cells of an individual—this is called *somatic mosaicism*. Depending on how early in development the variant occurred (and on its effect on the fitness of the cell), it can be present in all cells of the same lineage [Youssoufian et al., 2002]. Cancer is a pathological form of somatic mosaicism, which arises through the clonal expension of cells. [L. L. Campbell et al., 2007].

1.2. Structural Variation

Genomic structural variants (SVs) are commonly defined as variants affecting more than 50 consecutive base pairs of the DNA. The main purpose of this definition is to distinguish SVs from smaller indel variants or multi-nucleotide substitutions (i.e. blocks of consecutive SNVs) [Alkan, Coe, et al., 2011]. Indels are variants (of up to 50 bp) that insert or delete nucleotides, which is the same situation seen from different perspectives (hence the neologism *indel*). A more appealing definition than the arbitrary 50 bp threshold is that indels are detectable inside a contiguously mapped DNA sequencing read (introduced in Section 1.3.1) whereas SVs are detectable across alignments, yet also this definition no longer fully applies in the light of novel long-read sequencing technologies (Section 1.3.2). Fortunately, a clear distinction is not biologically relevant. SVs come in many different flavors of which the major ones are described subsequently.

SVs in the human genome are of particular relevance for health and disease. For example, they are implicated in various Mendelian diseases and in cancer [Weischenfeldt et al., 2013]. Later, in Section 3.1, I specifically discuss the phenotypic impact of SVs and present a study, in which I investigated a particular aspect of the functional consequences of SVs.

1.2.1. Different classes of structural variation

The spectrum of structural variants is broad. The major SV classes are generally be divided into copy number variants, such as deletions and duplications, or balanced rearrangements, such as inversions and translocations, yet a series of other SV forms is known. Below, I introduce the major classes of SVs that are relevant in this work.

Copy number variants (CNVs) describe the focal loss or gain of genetic material. They are termed *imbalanced*, as they do not leave the balance of the two homologues intact. A loss of DNA is called a *deletion*, and a gain either *duplication*, *triplication* or simply by its *copy number*. For example, a deletion has a copy number of one instead of the expected copy number of two in a diploid organism. A duplication that arises on one of the homologues leads to total copy number of three, and so on. Duplications are in *tandem* when the additional copy inserted in direct proximity to the original locus instead of somewhere else in the genome. The latter is referred to as *interspersed* duplication (Figure 1.1). The introduction of new sequence is called an *insertion*; however, depending on the source of the incorporated DNA, insertions can be assigned to one of several classes, only

one of which is briefly mentioned later—they are typically not counted as CNVs though.

The loss or gain of whole (or major parts of) chromosomes, historically visible under a microscope, is summarized as *aneuploidy*. Aneuploidy can range from a single chromosome (or at least the majority of the chromosome) being lost or gained, up to a complete increase or decrease of the ploidy level (of all chromosomes). The expected ploidy in diploid organisms is 2N, where N is the number of chromosomes and 2 the number of homologous copies. Ploidy can aberrantly increase to *triploidy* (3N), *tetraploidy* (4N) or even higher states covered by the general term *polyploidy* (Figure 1.1). Cells can also be in a purely *haploid* state (1N), but they are rarely viable due to problems with chromosome segregation. Mixed states, where only some chromosomes increase their copy number, are somtimes also referred to as hyperploidy.

ploidy *In humans,* N *equals 23, meaning that we carry 46 chromosomes in our cells. Interestingly, this number was falsely believed to be 48 for three decades before it was corrected by Tijo et al. [1956]*

Other types of SVs do not change the total copy number of a locus. Notably, *inversions* reverse the orientation of a locus, but generally do not include gains or losses (Figure 1.1). In fact, even an inversion can introduce (or co-locate with) CNVs depending on its mechanism of formation, which his is one of the major findings of Chapter 2. In cases where multiple SV classes occur within the same allel we term them *complex*. The most prominent examples are *inverted duplications*, which are duplications that insert in reverse orientation into the genome (Figure 1.1). Nevertheless, non-complex, i.e. *simple inversions* are the prime example of balanced SVs as they re-structure the genome without gaining or loosing genetic material.

Another class falling into the category of balanced SVs are *translocations*. In a translocation event, genetic material is exchanged between two non-homologous chromosomes. A *reciprocal translocation* is balanced because the total amount of genetic material does not change, just the assignment of certain loci (potentially of whole chromosomal arms) to chromosomes. But *imbalanced translocations* can arise, too. Here, one chromosome remains largely unchanged but a part of the homologue is duplicated and added to another chromosome, which might itself loose genetic material at the same time. This can involve whole chromosome arms, but also smaller loci, which is also covered by the definition of translocation. Typically though, translocation refers to the special case of a reciprocal translocation as shown in Figure 1.1.

Furthermore, when cells lack one of the two alleles within a larger genomic region, this is called a *loss of heterozygosity (LOH)*. LOH is an immdiate consequence of the (partial) loss of a homologue during a deletion. However, there are also copy-neutral LOH events in which the same haploid genotype is present in two

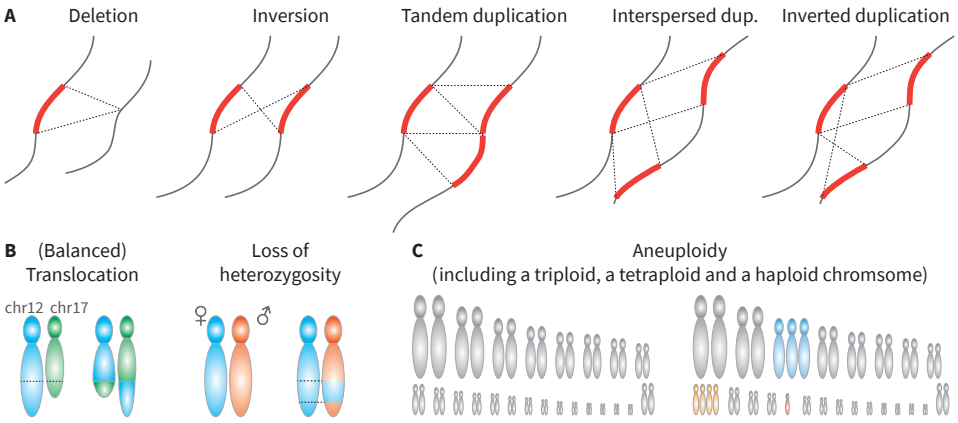


Figure 1.1: Types of structural variants. Each case is depicted by the original locus on the left and the affected locus on the right, where dashed lines are used to highlight the orientation. **A:** Different types of focal SVs of a genomic locus (red) within double-stranded DNA (represented by grey line). **B:** Chromosomes are depicted by double oval shapes. In the (balanced) translocation, chromosomes 12 and 17 are chosen exemplary to stress that exchange happens between non-homologous chromosomes. In a loss of heterozygosity, though, the maternal and paternal homologue of the same chromosome are shown. **C:** Ideograms of a normal and aneuploid cell are shown. For the sake of demonstration, the affected cells carries a haploid, a triploid and a tetraploid chromosome.

copies. This might for example occur when an individual inherits two copies of a chromosome from one parent, and none from the other parent (uniparental disomy), but it can also occur via other mechanisms (Section 1.2.2). LOH is often not observed directly, but indirectly by looking at smaller variants (notably SNVs) in a given genomic region—the absence of heterozygous variants is an indicator of LOH.

Finally, various other forms of SVs exist that are of less relevance for this work. One exception, which shall briefly be mentioned here, are *mobile element insertions (MEIs)*. Mobile elements, notably *transposons*, are DNA elements that can "jump" within a host genome. The human consists to a large fraction of the remainders of such elements [Haubold et al., 2006], which are largely prohibited from active transposition by repressive mechanisms in the host cell. A MEI may occur in a cut-and-paste or a copy-and-paste fashion and, although they principally resemble duplications or translocation, they are seen as separate class due to the fundamentally different mechanisms of formation.

1.2.2. Molecular mechanisms underlying the formation of SVs

In order to truly conceive structural variants, it is important to understand how they originate. Because we understand certain mechanisms of formation, we can today explain why SVs are not evenly distributed across the genome, for example, or why they re-occur independently in specific locations [Hastings, Lupski, et al., 2009]. More and more accurate discovery of SVs, on the other hand, has led to a better understanding of the functioning and impact of these mechanisms [Hastings, Lupski, et al., 2009; Abyzov et al., 2015]. Based on specific scars around the breakpoints of SVs, the mechanism that introduced a SV can sometimes be unraveled in retrospect. This is exactly the idea I apply in Chapter 2 to find out how the complex SVs we find were formed. Here, the major molecular mechanisms involved in formation of aneuploidy, focal copy number changes and inversions shall be introduced. They have previously been described in great detail by James Lupski and colleagues [Hastings, Lupski, et al., 2009; Carvalho et al., 2016].

Aneuploidy occurs through missegregation of single chromosomes during cell division. During meiosis, either in oocytes or spermatozoa, this can lead to inheritable aneuploidy, which was estimated to occur in 5% of human pregnancies [Templado et al., 2013]. Missegregation occurs via nondisjunction of chromosomes in meiosis I, when homologues fail to separate, or in meiosis II and mitosis, when sister chromatids are not separated properly. Alternatively, it occurs as a consequence of anaphase lag, in which a chromosome is lost in both daughter cells due to a delayed movement of chromosomes in anaphase [McMaster Pathophysiology Review (website), 2018].

Polyploidy arises differently, for example when an egg is fertilized by two sperm cells simultaneously or a fertilized ovum fuses with a sperm cell [McMaster Pathophysiology Review (website), 2018]. Missegregation of chromosomes occuring during mitosis may lead to somatic aneuploidy, which is observed in many cancer types [Gordon et al., 2012]. Also polyploidy can occur somatically, for example via repeated rounds of DNA replication without subsequent mitosis or with partial mitosis without subsequent cytokinesis. This occurs naturally, as for example in the polytene chromosomes in insect salvary glands or in hepatocytes of the human liver, but also spontaneously as frequently seen in different types of cancer [Davoli et al., 2011].

Focal SVs arise either during replication or after a break of the double-stranded DNA backbone [Hastings, Lupski, et al., 2009]. Double strand breaks and replication errors occur stochastically or result from cellular stress, but the cell actively counters such errors through its powerful repair mechanisms. DNA repair is not

always faithful, though, sometimes leading to the formation of SVs. A major mechanism of SV formation employs the homologous recombination machinery, which uses homologous sequence (from the sister chromatid) as a template to repair a break. Homologous recombination during meiosis can lead to *gene conversion*, which results in the replacement of an allele via the second allele (LOH).

However, given the repetitive nature of the human genome, homology (or near identical sequence) might not only be present at the respective locus but also in other, non-allelic loci. This *non-allelic homologous recombination (NAHR)* can create various types of SVs, including deletions, duplications, inversions and even translocation, depending on position and orientation of the ectopic homologous sequence [Carvalho et al., 2016]. The presence of homology, notably of large segmental duplications of more than 90% sequence identity and several kilobases in size, predisposes the human genome to the formation of recurrent SVs via NAHR [Carvalho et al., 2016]. Conversely, when near identical sequence is detected flanking SV breakpoints on both ends (20 bp are a usual lower threshold), such an SV is believed to be formed via NAHR [Onishi-Seebacher et al., 2011].

Other repair mechanisms do not require homology. *Non-homologous end joining (NHEJ)* is the dominant pathway during G0/G1-phase to efficiently re-ligate the ends of DNA double strands, usually leaving traces of not more than a few deleted or inserted base pairs at the junction [Lieber, 2008]. Importantly, NHEJ is available to the cell before sister chromatids are present and is a fast way to react to double strand breaks. When multiple double strand breaks arise simultaneously, NHEJ can falsely ligate genomic loci in the wrong order and thus introduce SVs. A related mechanism uses sequence identity of few (as little as 1-4) base pairs, also known as *micro-homology*, to initiate re-ligation via a mechanism called *micro-homology-mediated end joining* [Hastings, Lupski, et al., 2009].

micro-homology *It is debatable whether the term* homology *is correct here, i.e. whether the short stretches of identical DNA on both sides of an SV breakpoint in fact share a common evolutionary ancestry*

Replication of DNA, which happens prior to each cell division, is also susceptible to errors. For example, through a process called *replication slippage* smaller deletions and duplications can arise between stretches of homology within a replication fork, limited by the size of an Okazaki fragment [Hastings, Lupski, et al., 2009]. Furthermore, the DNA backbone can break within a replication fork, leaving a single-ended double strand break. Such a break can be faithfully resolved by the *break-induced replication (BIR)* mechanism: a single strand of the unfinished DNA molecule anneals to homologous sequence in the template DNA to restart replication, which can continue up to hundreds of kilobases from there [Carvalho et al., 2016]. Again, this search for homology may fail and either anneal to the homologous chromosome (instead of the sister chromatid), leading to extended stretches of LOH, or ectopically, resulting in one of several possible SV

types including CNVs and inversions.

Other replicative mechanisms of SV formation requires no, or only short stretches of micro-homology. Notably, a version of BIR that can operate independent of the homologous recombination machinerey was described, which relies only on micro-homology (4-15bp) to invade template DNA. The mechanism was consequently called *micro-homology-mediated break-induced replication (MMBIR)* [Hastings, Ira, et al., 2009]. Moreover, homology-independent rearrangements occurring during replication can include multiple complex rearrangements and are more prone to copy gains than losses, in concordance with a model of fork stalling and template switching (FoSTeS) [F. Zhang, Khajavi, et al., 2009; Hastings, Lupski, et al., 2009].

In summary, errors during cell division, replication or double-strand break repair can introduce various forms of structural variants. Especially CNVs and inversions may arise via many different mechanisms, which can sometimes, but not always, be inferred in retrospect based on the nucleotide sequence around their breakpoints.

1.3. DNA sequencing technologies

DNA sequencing refers to the process of deciphering the order of the four bases (adenine, cytosine, guanine, and thymine, abbreviated by A, C, G, and T) that constitute a DNA molecule. In the 1990s, sequencing the human DNA was a decade-long, multi-million dollar effort but it led to the successful production of a reference genome of humans and many other organisms [International Human Genome Sequencing Consortium, 2001; Venter, 2001]. Today, thanks to technological improvements, DNA sequencing has become a standard technique applied on a daily basis in genomics research. The advance of sequencing technologies has truly revolutionized genetic research and brought unforeseen capabilities also to studies of structural variation. In fact, these capabilities are not yet completely satisfactory (as described in Section 1.4) and ongoing development of new technologies and protocols is continuously pushing the boundaries of what is possible. Since DNA sequencing takes such a prominent position in my research—virtually every experiment in this thesis includes sequencing—the major techniques shall be introduced here.

1.3.1. Massively parallel sequencing

The foundation of modern DNA sequencing technologies was laid in 1977 by Frederick Sanger and his "chain termination" technique [Sanger et al., 1977]. Despite not being the first DNA sequencing method, the chain termination method brought unprecedented ease of use and accuracy [Heather et al., 2016]. It is based on DNA polymer extension via DNA-dependent DNA polymerase (i.e. replication) and the incorporation of dideoxynucleotides, which stop polymerization. With subsequent electrophoresis, partially replicated DNA fragments can be ordered by length and nucleotides identified based on radioactive or fluorescent labels. Sanger sequencing was instrumental in the Human Genome Project and, owing to the accuracy and length (around 1 kb) of sequenced fragments, it is still used today for validation purposes. It sequences single DNA fragments, though, and is thus laborious to apply in a larger scale.

The throughput could be dramatically increased with the advance of massively parallel sequencing (MPS), which is also referred to as short-read sequencing, next-generation sequencing, 2nd-generation sequencing, or high-throughput sequencing. Different commercial techniques were brought forward in the first decade of this millennium, including Pyrosequencing by 454 Life Sciences, Sequencing by Oligonucleotide Ligation and Detection (SOLiD) by Applied Biosystems, Nanoball Sequencing by Complete Genomics, Helicos Single Molecule Fluorescent Sequencing by Helicos BioSciences, and the Reversible Terminator Chemistry by Solexa [DNA sequencing (Wikipedia), 2018]. Today, the market for MPS technologies is vastly dominated by Illumina, who acquired Solexa and their technology in 2007. Here, the core principles of Illumina DNA sequencing shall be described representative for MPS in general.

Key principles of parallel DNA sequencing Like the Sanger technique, Solexa/Illumina's approach relies on the concept of sequencing by synthesis, i.e. by replication through a DNA polymerase. It in fact also utilizes the incorporation of fluorescently labeled dideoxynucleotides, which initially terminate the polymerization. A major novelty, though, is that the fluorescent label can be removed and the 3′ hydroxyl group of the dideoxynucleotide chemically restored. This technique is widely known as reversible terminator chemistry [Turcatti et al., 2008]. DNA is then replicated step by step, in each of which the incorporated nucleotides are detected using fluorescent imaging. This concept of cyclic DNA synthesis followed by fluorescent detection is shared by multiple of the aforementioned techniques, which use slighly different molecular mechanisms [Shendure

et al., 2008].

Another key concept of MPS technologies is a step of clonal amplification of DNA fragments in order to enhance the fluorescent signal detection. DNA fragments are initially ligated to adapter sequences and then, in case of Solexa/Illumina, immobilized on the flow cell and amplified via polymerase chain reaction (PCR) [Mullis, 1990], which they call bridge amplification. The aspect of parallelism comes into play when many (up to millions of) local clusters, each with a clonally amplified DNA fragment, are observed simultaneously during nucleotide incorporation. This again was driven by technological advances in high-resolution cameras, notably based on charge-coupled devices [Barbe, 1975; Shendure et al., 2008]. Due to clonal amplificaiton and high-resolution imaging, Illumina machines can sequence DNA with extremely high accuracy, with a per-base error rate is in the order of 0.1% [Fox et al., 2014].

polymerase chain reaction (PCR) *A method for amplification of DNA fragments. Did you know that PCR was invented only after Sanger sequencing, in 1983? See Mullis [1990] for a brief history*

Applications In contrast to Sanger sequencing, which targets a single locus, MPS can be applied to perform *whole-genome sequencing (WGS)*. During WGS, DNA is highly fragmented prior to the construction of a sequencing *library*, which is then sequenced via MPS to yield a large set of *sequencing reads* from all over the genome. Due to the random fragmentation this approach is also commonly known as shotgun-sequencing [Weber et al., 1997]. In a typical *re-sequencing* experiment, where a species with available reference genome is sequenced, these reads are then mapped to the reference for further analysis such as, for example, variant detection.

Apart from WGS, a large number of other sequencing protocols exist that utilize MPS to study different molecular characteristics[1]. A prominent example is RNA-seq [Morin et al., 2008; Z. Wang et al., 2009], which makes the mRNA present in cells available to sequencing by reverse-transcription into cDNA. We used this technique in Chapter 3. Sections 1.3.3 and 1.3.4 cover two other protocols based on MPS that are of particular interest in this work.

Paired-end sequencing Modern sequencing machines offer the possibility to sequence a DNA fragment from both ends. In case of Illumina/Solexa, this is achieved by a special step of bridge amplification that anneals the free end of all fragments in a clonal cluster to the surface and then frees the initially attached ends. Afterwards, sequencing continues in opposite direction to capture the other end of the DNA fragments. This approach is called *paired-end* sequencing or

[1] see *Seq, by Lior Pachter (website) [2018] for a list of such protocols

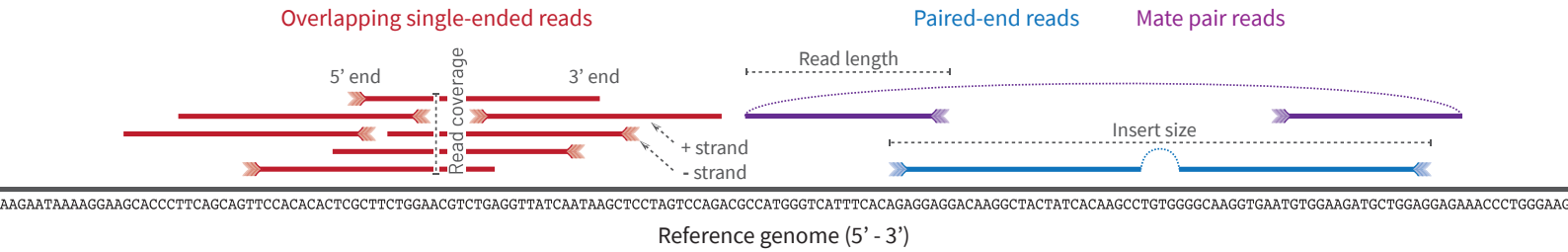


Figure 1.2: MPS sequencing reads. Sequencing reads from MPS are typically short (for example 100 bp), are sequenced from their $5'$ ends and can be single-ended or paired-end. In paired-end or mate pair libraries, the lengths of the original DNA fragment (which can be estimated after read mapping) is called insert size and usually much larger in mate pair experiments than in paired-end experiments.

paired-end tag sequencing and was used early on to study structural variants [P. J. Campbell et al., 2008] (see also Section 1.4.2). Using pairs, more bases can be sequenced at high quality than could be with a single read. Typically, the DNA fragments subject ot paired-end sequencing have a length of up to 500 bp. Larger fragments (typically around 3 kb, but up to 10 kb) can be achieved by creating *mate pair* libraries (a.k.a jumping libraries). In the approach of mate pair sequencing, a longer DNA fragment is first circularized before the connection of both ends is sequenced either single-ended or in paired-end mode [Korbel et al., 2007]. The size of the underlying DNA fragment is called *insert size*, and the number of sequenced bases *read length* (Figure 1.2). Paired-end and mate pair sequencing have played a pivotal role in SV detection, which is elaborated later in this introduction.

Sequence analysis After DNA sequencing, the computational analysis of the obtained sequencing reads begins. Naturally, this analysis may be very different depending on which protocol was used. For a WGS resequencing experiment, a very common first step is to assign the short reads to their most likely origin and read orientation within the reference genome. This process is called *read mapping* or *read alignment*, and software tools to perform this task are abundant [H. Li and Durbin, 2009; Weese et al., 2009; Langmead et al., 2009; Alkan et al., 2009; H. Li, 2013]. The intricacies that hamper read mapping are sequencing errors and the repetitive nature of large genomes, which do not allow unique placement of reads in many regions. These regions are said to have low *mappability* and are difficult to deal with—often they are simply neglected. Subsequent to read mapping, downstream analyses can be carried out such as SV detection, which I

read orientation *Only one DNA strand (e.g. the $5'$ strand) is encoded in a reference genome, but fragments from both strands are sequenced. Thus, also the reverse complement sequence of each read must be mapped—we say they are mapped to the* minus *strand (Figure 1.2)*

describe in Section 1.4.

The very popular paired-end sequencing allows the mapping of a read to be informed by the placement of the second read. For example, when one read maps ambigously, it can somtimes be "anchored" by the second, uniquely mapping read. Paired-end reads, which are sequenced from their $5'$ ends towards one another, align in convergent orientation to the reference genome, whereas mate pairs align in divergent orientation (Figure 1.2).

An alternative to read mapping is *de novo* assembly, which exploits the relationship of sequencing reads to one another (i.e. a common subsequence) to restore the sequenced genome independent of a reference. These methods, however, typically fail to produce long consecutive sequences [Alkan, Sajjadian, et al., 2011].

1.3.2. Emerging long read sequencing technologies

Over the last years, new sequencing technologies have been developed which are commonly referred to as "3rd-generation sequencing" in the community. These technologies are fundamentally different to MPS technologies in that they avoid clonal amplification of DNA fragments, but sequence single molecules instead. Although single molecule sequencing had been feasible already earlier [Braslavsky et al., 2003], later commercialized by Helicos BioSciences, third-generation sequencing is commonly associated only with the techniques of Pacific BioSciences (PacBio) and Oxford Nanopore Technologies (ONT). In contrast to the Helicos platform, these techniques achieve significantly longer read lengths (up to more than 100 kb) at usually decreased accuracy. Unlike Sanger's technology, they still sequence many molecules in parallel.

Pacific BioSciences Like Illumina MPS, The technology of Pacific BioSciences relies on the concept of sequencing by synthesis using fluorescently labeled deoxynucleotides. However, the sequencing occurs in real time on single molecules and was hence termed *Single-Molecule Real-Time sequencing (SMRT)* [Eid et al., 2009]. Fluorescent image detection occurs not in cycles, but continuously (acquiring *movies* instead of images). Besides deciphering the order of nucleotides, this principle also measures kinetics of the polymerase, allowing the detection of modified bases such as methylated cytosine [Flusberg et al., 2010].

The major challenge in the development of SMRT sequencing was the detection of the fluorescence signal from a single nucleotide upon ligation. To this end, researchers had engineered microplates with so-called zero-mode wave guides

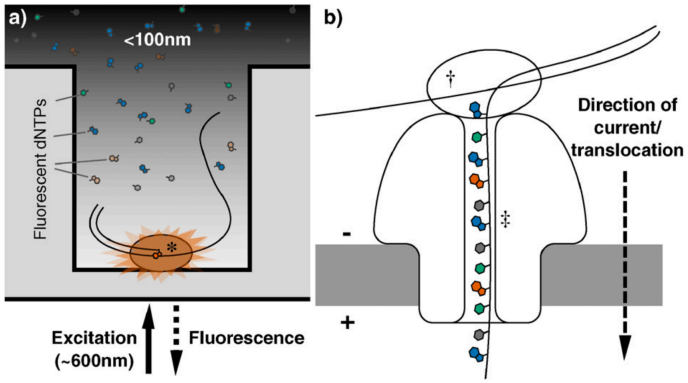


Figure 1.3: Concepts of PacBio and ONT sequencing. **A:** DNA polymerase inside a zero-mode waveguide processes a DNA molecule. The fluorescently labeled nucleotides are preferably illuminated at the bottom of the well, distinguishing their signal from the pool of nucleotides in the solution. **B:** Nanopore sequencing by passaging of a single-stranded DNA polymer through a pore in a non-conductive layer. Movement of the DNA molecule is facilitated by a voltage across the layer and decelerated by a processive enzyme (upper ellipse), e.g. a helicase that unwinds the double stranded DNA. Figure taken from "The sequence of sequencers: The history of sequencing DNA" [Heather et al., 2016] licensed under Creative Commons Attribution 4.0.

[Uemura et al., 2010]. The laser for excitation of fluorophores only illuminates the bottom of these nanowells, in which the polymerases are deposited. This way, only the nucleotides that are actively being incorporated by the polymerase can be excited and detected against a background of fluorophores outside the well [Heather et al., 2016] (see Figure 1.3 A). The base calling in single molecules is still noisier than detection in clonally amplified sequences, though, leading to reported per-base-error rates of 11-15% [Rhoads et al., 2015].

To improve accuracy, PacBio researchers promoted a technique called *circular consensus sequence (CCS)* [Travers et al., 2010]. Here, a double-stranded DNA fragment ligates to hairpin adapters on both ends to form a circular DNA molecule (the sugar phosphate backbone is one covalently bound ring), which still preserves its double-stranded structure. The polymerase can then pass this ring of DNA repeatedly, effectively sequencing the same fragment multiple times. This long read is computationally divided into sub-reads and a consensus is formed with reported accuracies of up to more than 99% [Rhoads et al., 2015].

Oxford Nanopore Technologies Oxford Nanopore Technologies utilize a fundamentally different approach to sequence single molecules. Driven by electrophoresis, a single-stranded DNA passes through a tiny pore that can detect

changes in the ionic current specific to the type of nucleotide passing through. The development of this technique spanned three decades and is based on findings from multiple labs, as Deamer et al. [2016] nicely portrayed. One major step towards the current technology was to find an appropriate pore that is just wide enough for a single-stranded DNA molecule to pass through: Initially *α-hemolysin* channels from *Staphylococcus aureus* had been used, which were later replaced by a genetically modified version of *MspA*, a porin from *Mycobacterium smegmatis* that allowed a much better signal-to-noise ratio [Butler et al., 2008]. A second crucial step was to decelerate the passage of each single nucleotide through the pore in order to allow accurate measurements of currents. While DNA polymers would naturally pass a pore at a rate of 10^{-13} sec per nucleotide even for the smallest voltages, researchers had the idea to place processive enzymes in front of the pore that slightly pause the passage of each nucleotide [Deamer et al., 2016] (Figure 1.3, B).

ONT offered the first commercially availabe nanopore sequeqncing devices, named MinION, in an early access program from 2014. Their technology utilizes an ATP-dependent helicase enzyme that unwinds double-stranded DNA before passaging the pores, with detection happening simultaneously at up to 512 pores [Jain et al., 2015]. A hairpin adapter is ligated to DNA fragments, so that after a single strand has passed the pore the complementary strand (linked covalently via the hairpin adapter) follows. This sequences the same fragment twice and is utilized to improve sequencing quality. These corrected reads are called *2D reads* and were reported to have an improved error rate of 15% [Jain et al., 2015]. Through the MinION early acces program our laboratory gained access to MinION devices, enabling us to apply the technique in the scope of the research project describe in Chapter 2.

1.3.3. Chromatin conformation capture sequencing

Studies of the three-dimensional structure of *in vivo* chromatin have for a long time been limited to imaging-based approaches. Dekker et al. [2002], however, proposed a new technique named *chromatin conformation capture (3C)* that probes the three-dimensional distance of loci using genomics methods. The basic idea is to crosslink DNA with itself in regions that are in close spatial proximity. The results of such experiments give insight into the *contact frequency* between two loci relative to other loci, which can be interpreted as an average three-dimensional distance between the loci observed across many nuclei.

The original 3C protocol relies on targeted PCR amplification and is capable

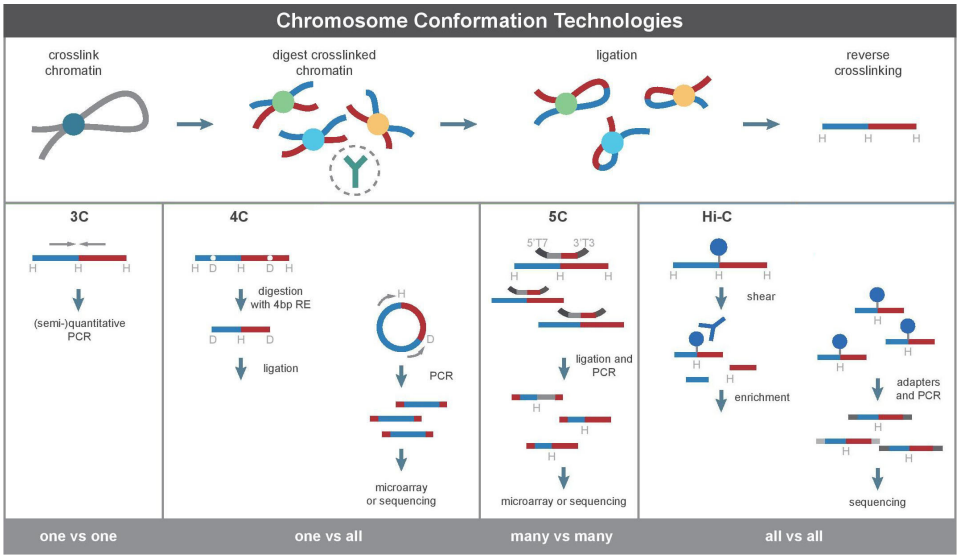


Figure 1.4: Chromatin conformation capture technologies. Chromatin conformation capture relies on cross-linking of DNA, enzymatic digestion and subsequent ligation of cross-linked loci. In the 3C-protocol, targeted interactions can be analyzed via PCR. The Hi-C protocol enriches for cross-linked fragments based on biotin labels before these fragments are paired-end sequenced. Figure modified from "Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application" [G. Li et al., 2014] licensed under Creative Commons Attribution 4.0.

of testing the interaction between exactly two loci. However, multiple protocols based on 3C have recently been shown to increase the level of parallelism, including 4C [Z. Zhao et al., 2006; Simonis et al., 2006] and 5C [Dostie et al., 2006]. Eventually, the combination of chromatin conformation capture and MPS resulted in chromatin conformation capture followed by high-throughput sequencing (Hi-C) [Lieberman-Aiden et al., 2009]. The core principle of Hi-C, as shown in Figure 1.4, involves cross-linking and digestion of DNA and then submitting cross-linked loci to paired-end MPS. After filtering of data (e.g. removing fragments ligated to themselves) both reads of a pair represent two loci that were in close three-dimensional proximity within the nucleus. Principally, the unbiased contact frequency of all loci against all other loci can be measured in this way. In practice, the resolution of these two-dimensional contact maps strongly depends on sequencing coverage. For example, to achieve 1 kb resolution—the densest maps known to date—Rao et al. [2014] required 4.9 billion pairwise contacts in a human cell line.

Hi-C was designed and used to study chromatin conformation and specific

three-dimensional features, as for example DNA loops forming between enhancers and promoters. It further revealed previously unknown structural features of the genome, which are discussed in more detail in Section 3.3.3. However, the characteristics of Hi-C contact maps predestinate them for at least two additional use cases: Intra-chromosomal interactions are more frequent than inter-chromosomal interactions, which is in concordance with the theory of chromosomal territories [Cremer et al., 2001]. Thus, Hi-C data can by utilized to cluster genomic loci by chromosome, which is especially relevant for *de novo* assembly [Burton et al., 2013]. Another observation is that the contact frequency between loci typically decays quickly with increasing genomic distance, meaning that the highest signal is detected between loci that are also close in linear genomic proximity. Because of that, larger genomic rearrangements become prominent in contact maps—this is what I utilized to characterize genomic rearrangements in Chapter 3.

1.3.4. Strand-seq

Strand-seq is a single-cell DNA sequencing protocol that preserves the identity of homologues by sequencing only the template strand of each chromosome [Falconer et al., 2012; Sanders et al., 2017]. The readout are sequencing reads, obtained via MPS (e.g. on an Illumina platform) in either paired-end or single-ended mode, which all map in the same orientation (plus strand or minus strand) to a reference genome if they originated from the same homologue[2]. In cells where homologues are inherited on opposite strands, which we call *Watson (W)* and *Crick (C)* strands, the original homologue for each read (including potential variants captured by this read) can be determined simply based on its mapping orientation. Strand-seq can thus reliably phase (i.e. distinguish haplotypes) chromosomes in their full length [Porubsky et al., 2016]. Moreover, given the consistent directionality of sequencing reads across a homologue, Strand-seq allows to detect (large) inversions in respect to the reference genome [Sanders et al., 2016]. Strand-seq was futher used to study sister chromatid exchange events (introduced in Section 4.2) [Falconer et al., 2012].

Strand-seq requires actively replicating cells, which are grown for a single round of replication in a 5-Bromo-2′-deoxyuridine (BrdU) medium. The incorporation of this thymidine analog into the newly synthesized DNA strand is the basis for obtaining stranded sequencing libraries, because after cell division each daughter cell will have only one strand labeled with BrdU (Figure 1.5). After cy-

[2]In paired-end sequencing, the first reads of all pairs map in the same orientation.

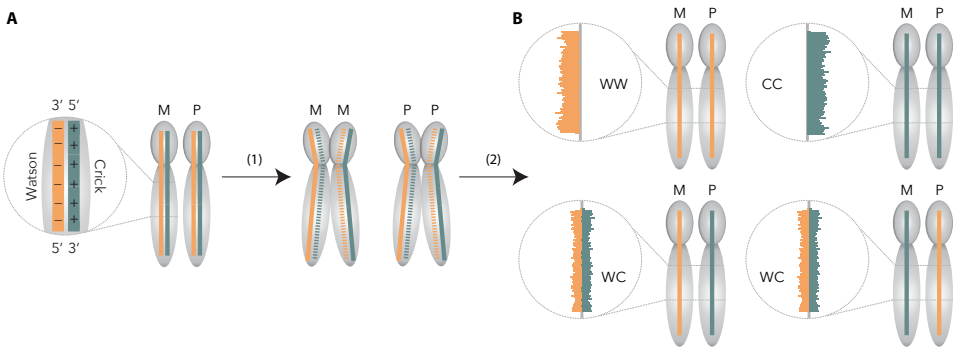


Figure 1.5: Strand-seq principle. **A:** Diploid cells (here schematically represented only for a single pair of chromosomes) contain maternal (M) and paternal (P) homologous chromosomes, each of which is a double-stranded DNA molecule. Watson (W) and Crick (C) are highlighted in orange and green. After replication in the presence of BrdU instead of Thymidine (1), cells contain two sister chromatids of each homologous chromosome, each with a different strand labeled (dotted line). After cell division, photolytic nicking of BrdU sites and library preparation (2), only sequencing reads from the non-labeled strand remain. **B:** A daughter cell inherits the two homologues in one of four different constellations, i.e. both as W (top left), both as C (top right), or both as different strands (maternal W, bottom left, or paternal W, bottom right). After read mapping cells appear in WW, CC, or WC configuration (circles). Figure modified from "Characterizing polymorphic inversions in human genomes by single-cell sequencing" [Sanders et al., 2016] licensed under Creative Commons Attribution 4.0.

tokinesis, the DNA of a daughter cell is enzymatically digested into fragments using a micrococcal nuclease enzyme (MNase). These fragments are ligated to sequencing adaptors. BrdU-containing fragments are then degraded via photolytic cleavage, so that subsequently only non-BrdU-containing fragments are amplified via PCR. Fragments from each single cell are tagged with cell-specific barcodes and are finally sequenced simultaneously, typically 96 cells together in one Illumina lane [Sanders et al., 2017].

In Chapter 4, I go into more depth on the workings of Strand-seq, notably on the computational analysis of Strand-seq data. I then present a novel approach that utilizes Strand-seq libraries across multiple single cells to detect mosaic SVs.

1.4. Structural variant detection

1.4.1. Traditional SV discovery

A traditional way to investigate chromosomal abnormalities is via *cytogenetics*. This involves the method of *karyotyping*, which functions by arresting cells in metaphase, staining chromosomes and observing them under a microscope [Spei-

cher et al., 2005]. The images of each chromosome are then ordered by chromosome to show an ideogram. Aneuploidy can then be detected by carefully inspecting the number and integrity of chromosomes, which is schematically depicted in Figure 1.1. Relevant disease-linked forms of aneuploidy could be unraveled early on this way, e.g. the trisomy of chromosome 21 that causes Down syndrome [Lejeune et al., 1959].

The classic technique has been refined to allow higher specificity and resolution, mostly via improved ways of staining, such as by quinacrine staining, Giemsa banding, or chromosome-specific labeling based on *in situ* hybridization [Speicher et al., 2005]. These techniques principally allow the detection of CNVs, inversions and translocations, yet the size range of these SVs has to be in the order of several Megabases or larger. Fluorescence *in situ* hybridization [Bauman et al., 1980], which relies on the annealing of fluorescently labeled DNA probes to their complementary DNA, is also applied in a targeted manner for validating predicted SVs of slightly smaller size.

Other means of detecting SVs include optical mapping [Schwartz et al., 1993; Teague et al., 2010] and hybridization-based microarrays. The latter ones, which are reviewed in Alkan, Coe, et al. [2011], used to be the dominating method for CNV detection before high-throughput sequencing became standard. One of the two major techniques in this category, namely array comparative genomic hybridization, utilizes the competition of the test sample's DNA and a reference DNA to a hybridization probe (e.g. short oligonucleotides) to infer the relative copy number of the tested locus [Snijders et al., 2001]. Using high-density arrays, this method can successfully detect deletions down to 500 bp in size [Conrad et al., 2010]. SNP arrays, on the other hand, utilize hybridization probes at sites of polymorphic SNVs to measure the allelic ratio within a single sample, which is called *B allele frequency (BAF)*. This way, CNVs but also LOH can be detected.

1.4.2. SV discovery in the era of massively parallel sequencing

Today, aforementioned techniques have largely been superseded by SV detection utilizing massively parallel sequencing data. SV detection methods based on MPS are generally separated into four conceptual approaches, namely read pair analysis, split-read analysis, read depth analysis, and sequence assembly [Alkan, Coe, et al., 2011]. In practice, SV prediction tools do not necessarily fit into only one of these categories. Below I will summarize the major ideas behind the different approaches as well as representative software implementing them.

Paired-end analysis SV detection based on paired-end sequencing utilizes the orientation and expected distance of two sequencing reads to another to detect rearrangements. For instance, when their mapping distance on a reference genome is larger than expected, a deletion may have occurred in the test sample anywhere in between those reads. Paired-end analysis can principally detect many different types of SVs, including CNVs, inversions, translocations, insertions and MEIs (Figure 1.6, B). This technique had first been used on bacterial artificial chromosomes [Volik et al., 2003], then on fosmid libraries [Tuzun et al., 2005], and finally in human genomes using mate pair sequencing [Korbel et al., 2007].

Paird-end read analysis is nowadays one of the dominating principles of SV discovery and is implemented in well-established software tools such as BreakDancer [Ken Chen et al., 2009], Delly [Rausch et al., 2012], CLEVER [Marschall et al., 2012], or LUMPY [Layer et al., 2014]. Besides the richness in detectable SV classes, an advantage of the paired-end read signature is that it can identify the breakpoints of an SV. The breakpoint accuracy depends on sequencing coverage and insert size distribution, but it is typically in the range of few to several hundred base pairs.

Split-read analysis Split-read approaches utilize the fact that sequencing reads, if long enough, can be divided and separately assigned to different locations of the reference assembly (Figure 1.6, C). This is different from intra-read gaps or mismatches, which are still tolerated within each alignment and which are typically used to detect SNVs and small indels. Based on the position and orientation of the partial alignments, split-read analysis detects SVs in the same way as paired-end analysis. The major advantage, though, is that breakpoints can be determined much more accurately, often down to the exact nucleotide. Split-read approaches had been explored, for instance, early on during the 1000 Genomes Project (Section 2.1.1) on 400 bp single-ended reads [Z. D. Zhang et al., 2011]. Tools such as Pindel [Ye et al., 2009] or BreakSeq [Lam et al., 2010] were among the first ones to specifically implemented the split-read approach.

Nowadays, strongly encouraged by increasing read lengths form standard MPS machines (e.g. up to 2 x 300 bp on an Illumina MiSeq platform nowadays), read mapping software has been refined towards the ability to directly perform split-read mapping, as exemplified by tools such as the widely used bwa mem [H. Li, 2013] or specialized tools like YAHA [Faust et al., 2012] and SplazerS [Emde et al., 2012]. This allowed other popular paired-end analysis detection tools to incorporate the split-read approach, for example in Delly, MATE-CLEVER [Marschall et al., 2013], and LUMPY.

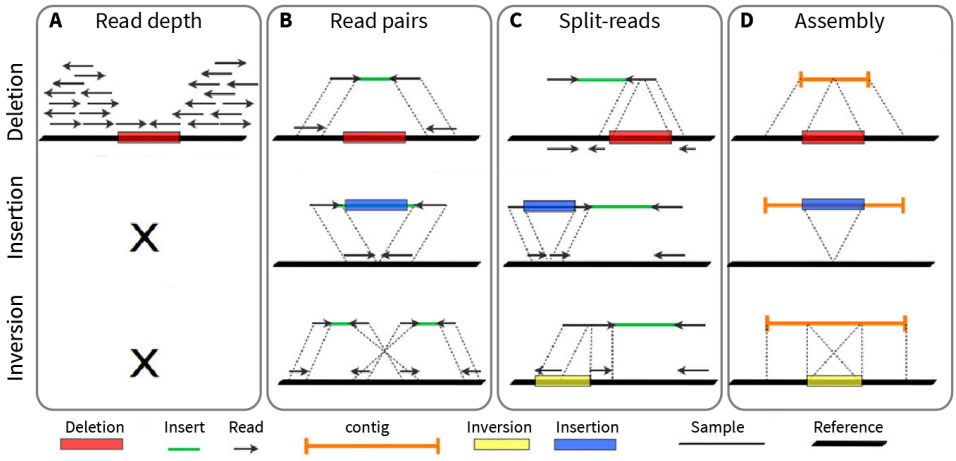


Figure 1.6: Principles of SV detection in MPS data. Schematic representation of the four distinct mechanisms for SV detection on three examples: A deletion, an insertion (of duplicated sequence, for example), and an inversion. Figure modified from "Detection of Genomic Structural Variants from Next-Generation Sequencing Data" [Tattini et al., 2015] licensed under Creative Commons Attribution 4.0.

Read depth analysis A complementary method for detecting CNVs utilizes the total read depth signal inside an SV (Figure 1.6, A). This resembles the methodology of microarrays, yet with improved resolution since all of the (mappable) genome can be covered instead of selected loci. SegSeq [D. Y. Chiang et al., 2009] and CNV-seq [Xie et al., 2009] were among the first tools that utilized read depth in a sample-vs.-control scenario to detect CNVs. mrFast [Alkan et al., 2009] and CopySeq [Waszak et al., 2010] extended this approach to single-sample CNV calling, and later CNVnator [Abyzov et al., 2011] and genome STRiP [R. E. Handsaker et al., 2015] gained more popularity. Normalization of read depth is a major challenge in these approaches owing to an uneven sequencing coverage, which is why these tools typically perform best when an internal reference (e.g. a control sample) is available. The popular population-scale CNV caller genome STRiP even suggests a minimum of 20 to 30 sample genomes in order to perform well. In contrast to paired-end or split-read detection, the read depth method cannot accurately predict breakpoints, which becomes a major disadvantage especially for smaller variants. For very large variants, on the other hand, read-depth methods can still robustly predict CNVs even when their breakpoints reside in repetitive regions.

Sequence assembly At last, sequence assembly-based methods do not rely on the information provided by read mapping software but perform *de novo* assembly instead, as demonstrated by Yingrui Li et al. [2011], for example. A comparison of the sample sequence to the reference genome, e.g. via sequence alignment, then reveals the presence of SVs (Figure 1.6, D). While whole-genome assembly still remains limited and computationally expensive ["Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species" 2013], there are methods that perform local re-assembly of reads, for example TIGRA [K. Chen et al., 2014] or the recent NOVOBREAK [Chong et al., 2017]. Especially for SV types that are difficult to detect by paired-end or split-read approaches, notably for insertion of novel DNA sequence, assembly can yield a great benefit. Tools that address this are NOVELSEQ [Hajirasouliha et al., 2010], MINDTHEGAP [Rizk et al., 2014], and BASIL/ANISE [Holtgrewe et al., 2015].

In practice, SV detection using any of the four different approaches typically requires an additional filtering step after initial SV prediction. Such filters can rely on the quality metrics provided by the prediction tool, but optimally they involve an independent signal, such as read coverage for paired-end predicted CNVs. One useful signal that shall be highlighted here is B allele frequency. This idea stemming from microarrays is applicable to MPS data, too: At sites of heterozygous SNVs that reside within a putative SV, the sequencing reads supporting both alleles can be contrasted to infer the copy number of the locus. Notably, in Section 3.3.2 this principle was instrumental to validate predicted duplications.

1.4.3. State of the art and limitations of SV studies

Structural variants in the human genome have been studied many times, driven by the availability of new technology. Initial population studies mapped large CNVs in several individuals using microarrays [Jonathan Sebat et al., 2004; Iafrate et al., 2004; Sharp et al., 2005; Redon et al., 2006]. With refinements of these techniques, CNV discovery could later be expanded to hundreds of individuals and down to a detection size of 1 kb (or even 500 bp), which scaled up the number of detected variants tremendously [McCarroll et al., 2008; Conrad et al., 2010].

Further improvements in the discernible size range, in accuracy of breakpoint, and in the types of SVs detectable were reached with the application of MPS technologies. Korbel et al. [2007] and Kidd et al. [2008] were among the first studies to utilize paired-end-like approaches to study SVs, including balanced ones, in few individuals. A series of studies followed that explored all the different technical approaches mentioned in Section 1.4.2.

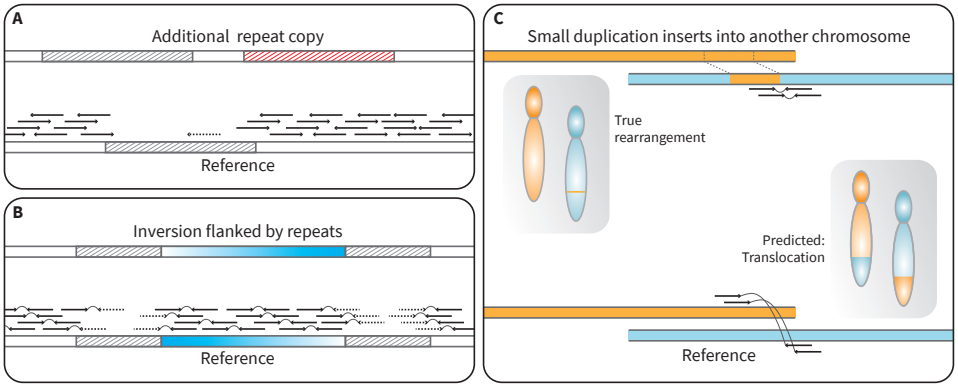


Figure 1.7: Examples for limitations of MPS-based SV detection. Three examples for cases in which current MPS-based SV detection methods fail. The upper half in each panel shows sample DNA carrying an SV; below is shown where reads from that sample map to the reference assembly. **A:** A duplication of a repetitive element (shaded boxes) occurred, which is not detected because read mapping is masked within the repetitive region. **B:** An inversion flanked by repetitive elements. Because paired-end reads cannot be mapped uniquely inside the repeats, this inversion remains undetected. With an read length or insert size larger than the size of these repeats, the inversion could be revealed. **C:** Insertion of a small piece of DNA into another chromosome leads to the prediction of a reciprocal translocation. The fundamental principal behind this limitation is that standard MPS-based calling only detects the breakpoints of a copy-neutral rearrangement, but cannot reason about their inner state. This is different in a CNV, for which the inner state (i.e. its read coverage) can be utilized for calling. The issue depicted here also arises for other SV classes, notably inversions.

The first phase of the 1000 Genomes Project presented the then most comprehensive SV call set, which was based on on low-coverage sequencing data of 179 individuals [Mills et al., 2011]. However, especially inversion detection faced major limitations within the project, as I describe in more detail in Chapter 2. Further studies continued SV characterization in the human population (e.g. [Sudmant, Mallick, et al., 2015; Hehir-Kwa et al., 2016]) and in disease, revealing thousands of copy number variable loci that are linked to pathological phenotypes [Swaminathan et al., 2012; Forbes et al., 2011]. SVs were also mapped extensively in cancer genomes [Weischenfeldt et al., 2016; P. J. Campbell et al., 2017] and in other organisms such as *Drosophila melanogaster* [Massouras et al., 2012; Zichner et al., 2013] or *C. elegans* [Maydan et al., 2010].

An increased sequencing depth led to improvements of the sensitivity of SV calling and of the accuracy of breakpoint detection. Nevertheless, it did not overcome specific limitations owed to the repetitive nature of the human genome. Notably, SVs attributed to NAHR are known to be flanked by repeat sequence, in

which read mapping (and consequently paired-end SV detection) often fails. This has been termed the "short-read dilemma" [Onishi-Seebacher et al., 2011]. Unfortunately, the human genome consists to a large portion of repeats. Sequence analyses found that up to two third of the human genome are derived from repetitive elements (mostly transposable elements) [de Koning et al., 2011] with around 5% of the genome containing large (>10 kb) and highly identical segmental duplications [International Human Genome Sequencing Consortium, 2001]. Especially repeat-embedded inversions cannot be detected based on traditional techniques (including MPS) "at high throughput and high resolution" [Sanders et al., 2016]. Figure 1.7 depicts three exemplatory scenarios in which repeats confuse MPS-based SV detection.

Challenges related to the repetitive nature of our genomes have been discussed extensively in other areas, notably for *de novo* assembly [Alkan, Sajjadian, et al., 2011] and haplotype phasing [Browning et al., 2011]. In both cases, blocks of information (e.g. phasing blocks or contigs) fail to span through repetitive genomic regions. These challenges are even greater in other species with more repetitive or polyploidy genomes, which additionally tend to have reference assemblies of lower quality. Livestock and crop are two examples with outstanding environmental relevance that suffer from these limitations [Bickhart et al., 2014; Saxena et al., 2014].

At last, MPS experiments have been noted to be only limitedly suited for the detection of subclonal structural variation [Forsberg et al., 2017]. Such SVs, which may be present in a tissue carrying somatic mosaicism, as for example in the case of cancer, only affect a fraction of cells. Thus, in a standard bulk MPS experiment, which averages the signal across thousands to millions of cells, variants present at low frequency often remain undetected or are rejected as biological noise. As Section 4.1 further elaborates, this shortcoming has hampered studies on mosaic SVs in the past. Recent technological advances have allowed the analysis of single cells, which greatly increased the detction of mosaic CNVs. However, also these techniques cannot overcome the challenges in the detection of copy-neutral SVs in single cells (Section 4.1).

1.5. Research goals and thesis overview

Studies of structural variation are still fundamentally hampered by the shortcomings of current SV detection methods. This especially affects studies of balanced or complex rearrangements, which had often remained cryptic in previous stud-

ies. In this dissertation, I aim at uncovering and further examining SVs that had been difficult to ascertain beforehand.

In order to do so, I utilize emerging sequencing technologies and protocols—namely the techniques introduced in Sections 1.3.2 to 1.3.4. My work is structured into three separate research projects, in which I explore one of the techniques, each. Below, the specific aims of each project are outlined. A common theme throughout these projects is the design of new bioinformatics methods for data analysis and visualization. I believe that the full potential of novel technology can only be unlocked by the simultaneous development of new computational approaches. I hence thrive for advancing the state of the art of current methodology by building—and making available—software tools that future research benefits from.

In Chapter 2, I present my work for the final phase of the 1000 Genomes Project, in which my role was to validate inversion predictions in the human population. These inversion predictions had initially remained inconclusive in PCR-based validation experiments. We then utilized targeted long-read sequencing on both PacBio and ONT MinION platforms to examine the respective loci more closely. My immediate goal in this project was to verify inversion calls based on long-read information. Moreover, the subsequent goal was to further characterize these loci and investigate why previous validations had failed. Afterwards, an optional goal was to investigate the validated inversions deeper to understand the biological mechanisms that created them.

In Chapter 3, I focused on the functional aspects of structural variation. Together with my collaborators, we set out to study the consequences of chromosomal rearrangements on gene expression and on the three-dimensional organization of chromosomes—the latter had gained much attention in recent years and is introduce in depth in Section 3.1. We utilized Hi-C to study chromatin conformation of highly rearranged chromosomes of *Drosophila melanogaster*. A crucial initial step of the project was to map the rearrangements and other variation present in these chromosomes. My first goal thus was to characterize SVs, including large pericentric inversions, which had been known only at cytogenetic resolution beforehand. As a part of this goal, I aimed at exploring the usability of Hi-C data for validation and characterization purposes. The second and more general goal of this study was to find out whether, and how these rearrangements affect gene expression. My specific milestone towards this goal was to robustly determine differential gene expression between the rearranged and a non-rearranged (wild type) chromosome. At last, this information should be integrated with findings on chromatin structure (provided by collaborators) to be able to conclude on the

impact of rearrangements and on the mechanisms that are involved.

In Chapter 4, my collaborators and I explored the potential of Strand-seq to discover SVs in single cells. While this had been previously done for inversions, we aimed at extending this principle—for the very first time—to at as many SV classes as possible. Unlike previous efforts using Strand-seq, the particular goal was to enable this detection also for variants that are present in only a low fraction of cells. Here, I present the current state of this ongoing endeavor. The first goal in this study was to develop a general concept of how different SV classes can be revealed based on the signals available in single-cell Strand-seq data. Then, the next goal was to prove feasibility of this concept by implementing the approach and applying it to available as well as newly sequenced cell line data harboring structural variation. A third important aspect was to test the limitations of our approach in a controlled environment.

At last, Chapter 5 concludes on the achievements resulting from my work. This includes the specific questions of the three projects as well as a more general view on how emerging technologies improved SV characterization in the three different scenarios. I then briefly review recent developments by others in the community that occurred around the same time as my research. This should give a broader impression on how emerging technologies are about to change studies on structural variation and places my work within the broader context of recent advances in the field.

Complex Inversions in the Human Genome

2

In 2014 and 2015 I had the opportunity to collaborate with a large consortium of scientists on the 1000 Genomes Project. My supervisor Jan Korbel was the co-leader of the structural variation subgroup and, together with my colleagues Tobias Rausch, Adrian Stütz, Benjamin Raeder, Markus Hsi-Yang Fritz, and Andreas Untergasser, I approached the validation and characterization of inversions. This chapter covers my work for the 1000 Genomes Project, which not only turned out to solve an interesting mystery but also resulted in a co-authorship in Sudmant, Rausch, et al. [2015]. I continue by describing subsequent work, including a side project on sequence match visualization that came into being from collaboration with Markus Hsi-Yang Fritz (Section 2.4), as well as an analysis of inversion breakpoints (Section 2.3). The latter results were presented in form of a poster at the German Conference for Bioinformatics 2016 in Berlin. There is supplementary information to this chapter enclosed in the appendix (Appendix B).

2.1. Introduction

2.1.1. The 1000 Genomes Project

The 1000 Genomes Project was the effort of a large international panel of scientists to capture the genetic diversity of the human population as comprehensively as possible. The project can be regarded as the hitherto final step in a series of projects that unraveled and characterize the human genome: The Human Genome Project, which, founded in 1990, was likely the first truely large-scale collaborative work in biology, as well as their private competitor Celera Genomics (from 1998) aimed at completing the first human genome sequence. These tremendous efforts, published simultaneously in early 2001 [International Human Genome Sequencing Consortium, 2001; Venter, 2001], had both used DNA of a few individuals (among them Celera's then-CEO Craig Venter) to determine

an average human genome. The focus was shifted to the variation present in the human population with the initiation of the International HapMap Project in 2002. By performing whole-genome SNV genotyping on hundreds of individuals from multiple continents, including family-offspring trios, the project generated a first haplotype map of the human population [International HapMap Consortium, 2005; International HapMap Consortium, 2007; International HapMap 3 Consortium, 2010]. Over the years the HapMap Project was scaled up to more than 1,000 individuals, to discover rare SNVs (minor allele frequency < 5%) and to call copy number variants. The samples originally collected for the HapMap Project, which are still available today as immortalized lymphoblastoid cell lines, were also used by the 1000 Genomes Project. Starting long before my own participation, with the pilot phase publication dating back to 2010 [1000 Genomes Project Consortium, 2010], the project even extended this set of samples by multiple other populations [1000 Genomes Project Consortium, 2012]. However, the fundamental novelty was to apply low-coverage whole-genome sequencing which allows to assess a much broader spectrum of variants in a much larger portion of the genome.

In the third and final phase the project had performed mean 7.4x-coverage sequencing on 2,504 samples and additionally expanded the palette of computational methods that were applied as well as added new types of variation to be studied [1000 Genomes Project Consortium, 2015]. The final data set consists of approximately 88 million phased variants, which was estimated to include more than 99% of non-rare SNVs in the studied populations. Besides novel insights on, for example, seemingly dispensable genes, the primary achievement of the project was to make an unprecedented resource of genetic variation available to the public. This came along with technical advances in methods and standardizations such as the nowadays commonly used file format variant call format (VCF)[1]. The 1000 Genomes data has since been utilized in many research studies, for example to refine human population history [Veeramah et al., 2014], to impute [Howie et al., 2012] missing SVs in many genome-wide association studies [Wood et al., 2014, for example], or to discriminate somatic from germline mutation in cancer studies [Hiltemann et al., 2015]. At last, since the start of the 1000 Genomes Project a considerable number of similarly large sequencing studies has popped up that focus on specific populations [UK10K Consortium, 2015; Sulem et al., 2015; Telenti et al., 2016] or disease [P. J. Campbell et al., 2017] and which eventually scale up the catalogue of known variation in the human population.

[1]Established for the 1000 Genomes Project, the format is nowadays maintained by the Global Alliance for Genomics & Health (`https://www.ga4gh.org`

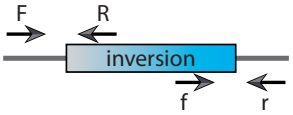
2.1.2. Predicted inversions and the validation problem

The Structural Variation subgroup of the 1000 Genomes Project focused on the identification of various classes of genomic rearrangements. With a total of 83 scientists involved and led by eight principle investigators with different expertise, the group assembled the to date most comprehensive resource of structural variation in the human population. Copy number variants, inversions, mobile element insertions and other types of SV were predicted from low-coverage whole-genome sequencing data, statistically phased and validated using a number of different approaches.

Among his many contributions to this project, Tobias Rausch had predicted inversions using the paired-end signature-based SV detection tool Delly. This set of inversions, which was strictly filtered according to validation results and population genetics aspects, contains a total of 786 inversion calls in the range of 500 bp - 50 kb (median size 1.70 kb) and ended up being the only inversion call set used in the project. In comparison to phase one of the 1000 Genomes Project, this is a more than 15-fold increase in the number of inversions.

Before its final release, this call set (just like any other call set in the project) had to undergo cycles of validation and filtering. Adrian Stütz and Benjamin Raeder were in charge of running PCR-based validation experiments of these inversions. In a nutshell they applied a Four primer strategy on inversion loci, which is capable of distinguishing the genotype of inversions based on the combination of bands resulting from the PCRs. However, this requires that the breakpoints of the inversion were accurately predicted and also lie in regions accessible to targeted PCR amplification. The latter was initially thought to be a problem, as inversions are known to often be flanked by inverted repetitive sequence due to their origination through NAHR (see Section 1.2.2).

Four primer strategy



Four separate PCRs are run using primer pairs FR, fr, Ff and Rr. In samples not carrying the inversion only FR and fr will yield bands, in homozygous carriers only Ff and Rr, and in heterozygous carriers all four reactions will yield bands

When they tested PCR verification experiments on a subset of 96 predicted inversion loci they were puzzled: for the vast majority of loci the received band patterns did not match the expectations, neither clearly validating nor invalidating the locus. When counted as invalidations, these results would have given rise to an false discovery rate (FDR) of up to 80%. Despite multiple trials to improve PCR conditions and primer locations, the experimental verification of the inversion call set remained unsuccessful for a long time and posed a scientific mystery.

Method	#Total	#Val.	#Inval.
PCR	96	27	2
PacBio	308	186	37
ONT MinION	96	54	6
Fosmid	34	28	4
Total	308	229	47

Table 2.1: Inversion validations. Number of predicted inversions analysed (#Total), validated (#Val.) and invalidated (#Inval.) using one of several techniques performed at EMBL or by collaborators.

Class	Count
Simple inversion	42
Inverted duplication	110
Inv. and del.	29
Inv. and multi-del.	14
Highly complex	11

Table 2.2: Complex inversion classes. Summary of classes of inversions found at 206 long read-resolved loci (*Inv. and del.*: Inversion flanked by deletion, *Inv. and multi-del.*: Inversion flanked by multiple deletions)

2.2. Results I: Long-read sequencing unravels unexpected levels of complexity

By the time I became involved in the inversion validation project my colleagues had already decided for a new strategy. They selected a subset of small inversions in a range of up to 10 kb including the surrounding genomic regions via amplification through long-range PCR. The resulting amplicons were then submitted to long-read sequencing in order to uncover the true genomic structure underlying the predicted inversions. At the same time collaborators within the 1000 Genomes Project had performed independent inversion validation experiments that will not be addressed in detail here. Just to mention them briefly: 35 loci were inserted into fosmid libraries, clonally expanded and then submitted to PacBio sequencing; another set of loci was genotyped in the cell line CHM1, which had recently been sequenced with PacBio [Chaisson et al., 2014] and then validated using the publicly available long read data; another 29 loci present in the individual NA12878 were evaluated in additional available whole-genome PacBio sequencing data for this individual [Pendleton et al., 2015]; at last, small set of inversion breakpoints was submitted to Sanger sequencing [Sudmant, Rausch, et al., 2015, supplementary methods].

2.2.1. Validation and characterization of inversion loci

The validation experiments carried out at EMBL were three-fold: First, a small subset of inversion calls, mostly loci with two-sided support from Delly calls, could be validated using PCR and aforementioned four primer strategy. For the majority of predictions however, PCR validations had remained inconclusive so

that long-read sequencing was applied. So secondly, amplicons were submitted to PacBio single-molecule sequencing[2] at Baylor College of Medicine and at MPI for Plant Breeding Research, Cologne[3]. Thirdly, a smaller number of loci were sequenced on a ONT MinION device[2] at EMBL. These numbers are summarized in Table 2.1.

I received both the raw sequencing data of the amplicons in multiple stages. At first I explored the differences between raw PacBio reads and the circular consensus sequences provided by PacBio 's internal software. In virtually all cases the circular consensus sequence showed the same information as its raw subreads, yet with less sequencing errors, which is why I preferred the consensus sequences from that point on. After mapping to the human reference genome using Blasr [Chaisson et al., 2012], I selected reads mapping their respective locus for visualization. I then generated dotplots of each long read against its reference locus using a custom script, which was developed by Markus Hsi-Yang Fritz and me and which is further described in Section 2.4. I used a very similar approach to analyze ONT MinION data: I obtained both 1D reads as well as 2D reads from multiple different sequencing runs. I then utilized Last [Kielbasa et al., 2011] to map reads to an artificial reference only containing the loci of interest. Again I selected mapping reads for visualization through dotplots.

Finally, I scanned the approximately 30,000 resulting plots together with Adrian Stütz for signs of inversions at the predicted loci. We categorized loci based on our visual inspection into the three classes: "validated", "invalidated", and "inconclusive". We required at least five independent reads supporting the alternative allele to count an inversion-type variant as validated. On the other hand, we deemed a locus as invalidated only if there were at least 30 reads supporting the reference allele and not a single one with an inversion signal. This asymmetric measure is motivated by an allelic imbalance yielding from the PCR, as I briefly describe in Section 2.2.2. The majority of validations were successful through support from PacBio reads, which outperformed ONT MinION data in both quality and quantity in our setup.[4]. In 20 cases ONT MinION was instrumental in validating loci that could not be resolved by PacBio due to a lack of coverage. Reassuringly, among the set of 37 loci that were both informative in PacBio and ONT

[2]Details on library preparation and sequencing can be found in the supplementary methods of Sudmant, Rausch, et al. [2015]

[3]Data is publicly available through the European Nucleotide Archive via accession numbers SAMEA3299513, SAMEA3257441, and SAMEA3257407

[4]Note that I do not carry out a detailed comparison of both technologies. Instead, we primarily used PacBio and additionally explored a very early version of ONT MinION, which turned out to be helpful for a subset of loci

MinION data, 35 loci agreed in showing an inversion-type variant and the other two agreed in showing the absence thereof. The final FDR for inversions in the 1000 Genomes project was estimated to be 17%, although this estimate might be conservative because of potential allelic dropouts (as explained in Section 2.2.2).

Most interestingly, we observed that only a minority of cases turned out to be simple inversions in the sense that a contiguous stretch of DNA is re-oriented within its original locus. Instead, most positive loci harbored additional complexity, i.e. other types of variation mixed with the inversion. We identified five major classes among the inversion-positive loci, which are presented in Figure 2.1: simple inversions, inverted duplications, inversion flanked by deletions, inversion flanked by multiple deletions, and loci with even higher complexity. In summary, a mere 20% of inversions validated by long reads are simple; the majority, on the other hand, are inverted duplications (see Table 2.2). Precisely, this is the case for the set of inversions with a single breakpoint predicted by Delly. There is a smaller subset of 21 validated inversion predictions with double-sided support from read pairs that are all simple inversions. The estimated FDR for inversions with double-sided support is 9%.

After the completion of the 1000 Genomes Project, Tobias Rausch adapted Delly to be able to distinguish double-sided (hence likely simple) inversion calls from potentially complex ones. To do so, overlapping paired-end signatures are combined and tested for the possibility of the locus showing an inverted duplication or an inversion flanked by a deletion. Nevertheless, this approach would most likely not have been successful in the scope of the 1000 Genomes Project due to the low average coverage of 7.5x, since without sufficient coverage it becomes less likely to find enough breakpoint-spanning read pairs for both ends of the complex rearrangement.

In the aftermath, when the high level of complexity is seen, it is clear why PCR validations were bound to fail. In fact, for simple inversions that were predicted as such (due to double-sided support from read pairs) PCR verification experiments worked about as good as expected. Nevertheless, complex loci were only visible to long read sequencing. Now that we know how these types of variants look like, paired-end mapping-based tools can be trained to capture these events and Delly is one of the first tools capable of identifying such classes of complex SVs. At last, the abundance of complexity was higher than anticipated; yet the total number of identified complex inversions is still small compared to other, e.g. copy-number variants, and their functional role remains to be shown.

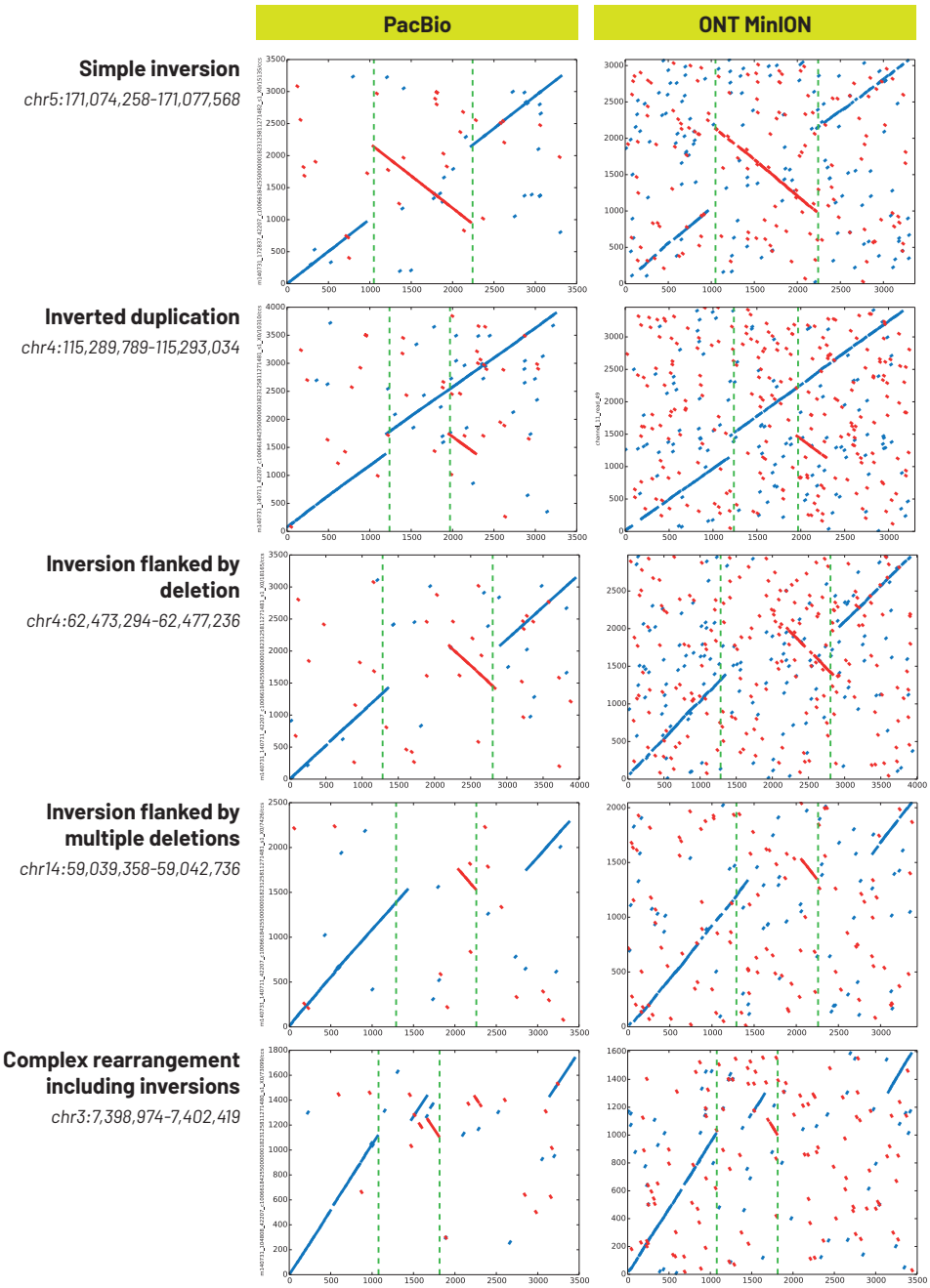


Figure 2.1: Complex inversion types revealed by PacBio and ONT MinION reads. Exemplatory loci were selected to represent the most commonly seen classes of complex inverted variants (listed in Table 2.2). Green lines denote inversions predicted by DELLY. PacBio circular consesus sequences and raw ONT MinION reads (vertical) are compared to their reference locus (horizontal) via dotplots using maximal unique matches of lengths at least 10 (PacBio) or 8 (MinION). Axes denote sequence lengths in base pairs. Original read names are included for later reference. See Section 2.4.1 for an explanation of the concept of dotplots.

2.2.2. Artifacts in amplicon sequencing

During the analysis of amplicons I made a number of interesting observations that can most likely be attributed to errors occurring in PCR or long-read sequencing. Since these observations led to confusion (or at least to discussions) during data analysis I describe them below and propose possible explanations for them.

A first type of artifacts is linked to PacBio's circular consensus sequences. As mentioned in Section 1.3.2, the quality of a PacBio read can be increased by sequencing the same DNA fragment multiple times in a ring-like structure. PacBio 's software resolves this repetitive signal by splitting it at the motif of the adapter sequence used to close the ring and by forming a consensus of the resulting subreads. In a few cases this software reported reads that consisted of an amplicon locus concatenated to a reverse copy of itself. A possible explanation of these cases would be that the software misses an adapter sequence, e.g. due to mutations occurring in early PCR cycles or due to insufficient sequencing quality. Yet also the possibility of a PCR artifact cannot be excluded. Eventually this signature was readily identified and did not pose a problem in the analysis.

A more severe artifact was allelic imbalance among the PCR amplicons. I observed up to 100-fold ratios in the number of reads supporting alternative and reference allele. Intriguingly there was a clear trend towards the shorter allele being more abundant. For example in an inverted duplication, the reference allele was usually more abundant than the inversion allele, whereas in the case of an inversion flanked by a deletion, the alternative allele was seen more frequently. This is, however, not suprising as it is a known effect of PCR to prefer shorter fragments and one of the reasons why many DNA protocols (such as standard Illumina sequencing for example) contain a size selection step. This is also why we required at least 30 reads of the reference allele to be seen before declaring a locus invalidated. Yet it could mean that some of the invalidated loci are in fact true inversion-type alleles but the alternative allele dropped out. This was not futher explored so that FDR estimates for the amplicon-based inversion validation can be considered conservative.

A last, initially very puzzling observation is also linked to long-range PCR and demonstrated in Figure 2.2. Up on inspection of dotplots of long reads I identified at least six loci that seemed to carry three different alleles. In all cases the same two alternative alleles were present in addition to a reference allele: an inverted duplication and a deletion. At first I verified that these observations are also present in ONT MinION data to exclude a PacBio sequencing artifact (also Figure 2.2). Next, I began investigating the possibility of the whole genomic locus

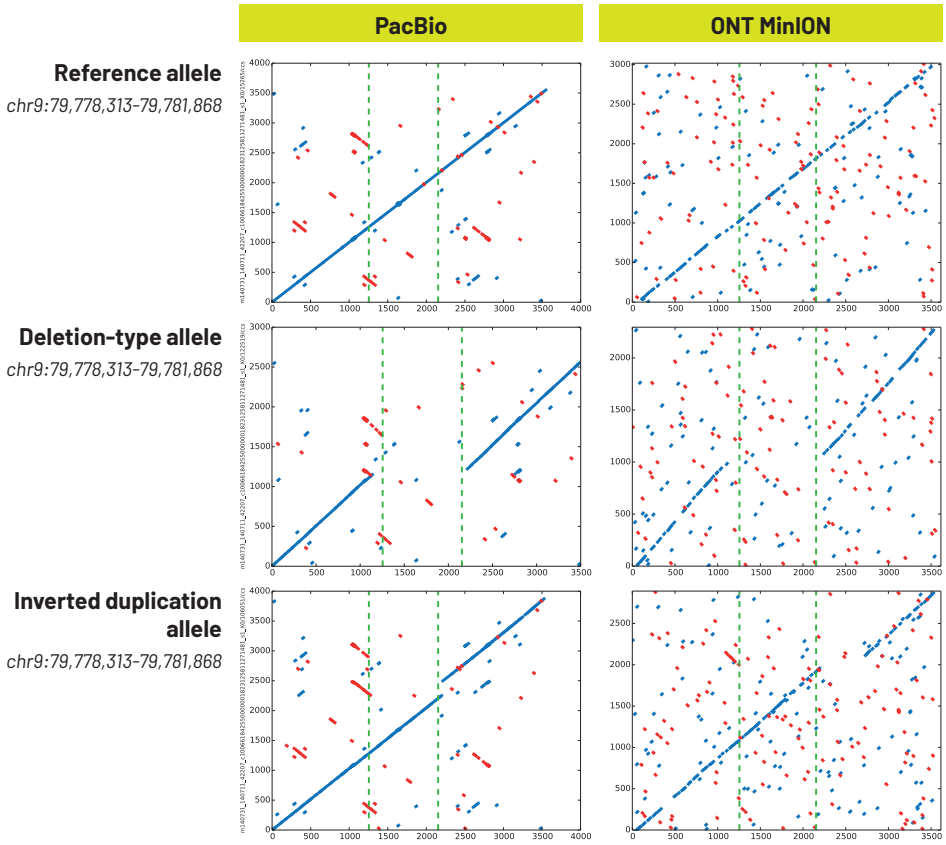


Figure 2.2: Loci showing three alleles, likely resulting from PCR artifacts. Example locus showing more than two alleles in a single diploid sample, plotted the same way as in Figure 2.1.

being duplicated within the genome, which could explain the presence of more than two alleles. I did not find such evidence based on read mapping and neither did the duplication call set created for the 1000 Genomes Project contain any information supporting this theory. The possibility remains that a rather recent event duplicated the same locus in the analyzed cell lines, yet a much simpler explanation is on hand. We have reason to believe that the deletion was caused during PCR by the presence of the inverted duplication. Such PCR artifacts have been previously described [W. Ji et al., 1994; Hommelsheim et al., 2015] and this theory is further supported by the fact that the deletion breakpoints in all cases roughly align with the insert position of the inverted copy. Even when this replication error, caused by the polymerase jumping to an annealing homologous region, occurred only once during the amplification process the then-shorter al-

lele could have been favored during subsequent PCR cycles. We decided to count these loci as inverted duplications, hence also as validated.

2.3. Results II: Analysis of inversion breakpoints

After completion of phase 3 of the 1000 Genomes Project, which includes the results described in Section 2.2, I was further interested in gaining a deeper understanding of this unexpectedly abundance of complexity in the inversion call set. These inversions are present in the germline of at least one of 2,504 individuals, representing an estimated allele frequency of approximately 0.05%, yet some of them are present much more frequently. The origin of such inherited variation, which might have occurred thousands of generations ago, is not accessible to functional assays. However, as various different studies revealed in the past (reviewed in Onishi-Seebacher et al. [2011] and Hastings, Lupski, et al. [2009] and partly explained in Section 1.2.2), the mechanisms generating SVs leave stochastic but specific traces in the genome that can later be revealed and potentially tracked. In more detail, these traces are found around the breakpoint junctions of an SV. As Hastings, Lupski, et al. [2009] explain, NAHR is known to induce inversions between two inverted homologous regions of at least several hundred base pairs. Hence inversions that are found between such flanking inverted repeats are likely to be generated by NAHR. On the other hand, homology-independent mechanisms such as NHEJ do not require annealing stretches of DNA. NHEJ is associated with very small (typically not more than a few base pairs) deletions or, more rarely, the insertion of free DNA. Moreover, multiple mechanisms are further associated to micro-homology, a feature that is demarcated by short stretches of identical DNA on both ends of the breakpoint. The sequence around breakpoints of SVs can thus be informative about the operating mechanisms.

In order to obtain nucleotide resolution sequence data I developed a computational pipeline to perform accurate assembly of PacBio reads of the amplicon loci described above. This pipeline consists of five computational steps and a final semi-manual validation stage and is described subsequently. With high-quality assemblies of the complex inversions at hand I could then search for characteristic traces of, for example, micro-homology around their breakpoints. I further expanded the list of analyzed loci by another type of SV and by additional PacBio-independent data sets. In the following section I describe they way I obtained high-resolution sequences and the findings resulting from those. In Section 2.4 I

further describe the visualization method that resulted as a side products of this work is.

2.3.1. Assembly of PacBio reads achieves nucleotide resolution sequence information

The five computational steps towards PacBio assembly cover both steps performed by external software as well as by own tools specifically designed for this task. At first, I extracted PacBio reads mapping to a specific locus using SAMtools [H. Li, B. Handsaker, et al., 2009] and depleted them for reads of the reference allele. This depletion relies on the orientation of sequence alignments computed by Last. Next, I fed the remaining reads into wgs-Assembler [Myers, 2000] version 8.2 to perform the actual assembly step per locus. The outcome is a list of contigs, of which I hope at least one to represent the true genomic structure of the inversion-type allele. Then I mapped PacBio reads back onto these contigs using Blasr and implemented several quality filtering steps, including trimming of low-coverage regions and a rudimentary detection of artifacts. Finally I "polish" the assembly using PacBio 's tool Quiver (PacBio software) []. This last step is the only point in the pipeline that utilizes the full potential of the quality metrics reported by the PacBio sequencers alongside the sequence reads, and turned out to be crucial in achieving near-perfect quality.

contigs Contigous *stretches of DNA derived from, in this case, stitching together sequencing reads based on their overlap. The order and orientation of these contigs to one another can often not be determined by the assembly software [Alkan, Sajjadian, et al., 2011]*

Afterwards I prepared the contigs for visualization-guided manual validation. I inserted all contigs into a database and used an early version of Maze (see Section 2.4) to prepare dotplots. The goal was to decide whether they (a) indeed show a correct inversion signal and not the reference allele or a spurious contig, and (b) which class of inversion they represent (Figure 2.1). This type of quality control is rather difficult to automatize completely, yet it is an easy task for the human eye. This is why I designed a simple, web browser-based user interface featuring keyboard shortcuts to be able to quickly annotate a large number of assemblies. Out of 17,000 initial contigs generated using this approach I managed to identify high quality assemblies for 153 complex loci, which were then used for breakpoint analysis.

Once I achieved satisfying quality with PacBio assemblies I complemented the dataset by other types of data that had become available: Sanger sequencing of 65 breakpoint loci that were generated by Adrian Stütz, and Illumina short read assemblies of multiple loci generated by Tobias Rausch. Furthermore I included another class of SV, namely proximal duplications, which were predicted by Tobias Rausch and which closely resemble inverted duplications yet without the

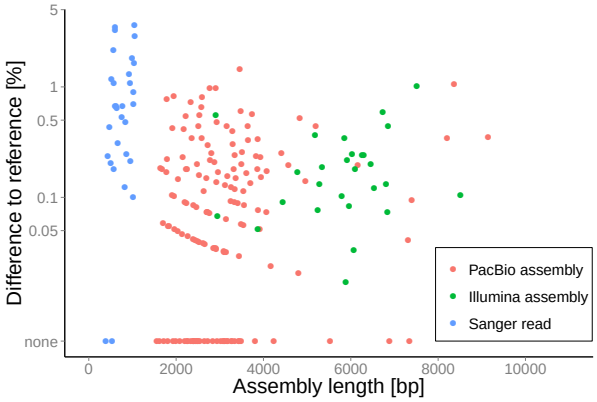


Figure 2.3: Quality of assemblies. Overview of length and quality of the 210 loci (listed in Table B.1) that were assessed using any of three technologies. Quality is approximated by number of mismatches to the reference genome, taking into account known variantion of the sample. Note that unlike Illumina-assembled reads, PacBio assemblies are restricted to the size of the amplicons.

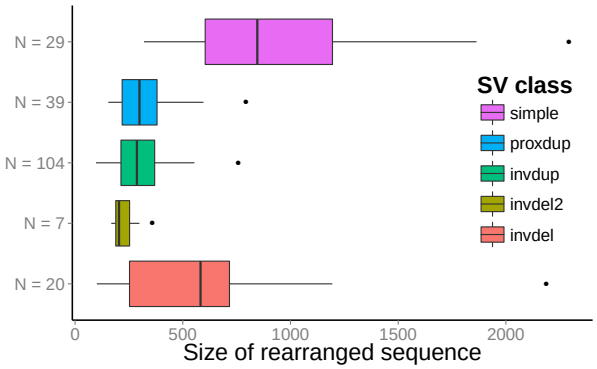


Figure 2.4: Sizes of rearranged sequences. Sizes were determined from assembled loci using LAST-split as described in Section 2.3.2. Note that the overall size of the amplicons is limited, which potentially explains shorter sizes for the rearranged part in loci with higher complexity.

copy being inverted.

As a measure of quality I determined the number of mismatches of the assembly to the reference locus. This comparison was performed within each mapping segment separately (e.g. within the inverted part and the flanking parts separately) and known SNVs were not counted. Eventually I gathered high quality, basepair-resolution sequence information for 210 loci (and 365 sample) of different classes of variants. The sizes and qualities of the best sequence per locus are shown in Figure 2.3 and are the technology yielding the best data per locus is explicitely listed in Table B.1.

2.3.2. Sizes of rearranged sequences suggest origin through a common mechanism

I used sequence alignments calculated by the split-mode of LAST, as implemented in MAZE, to determine each mapping segment and to resolve breakpoints (see Section 2.4.2 for more detail). This way I resolved the sizes of the rearranged sequence of all different classes of SV, as shown in Figure 2.4.

Interestingly, an immediate observation was that inverted duplications and proximal duplications show the same very tight size distribution of approximately 200 - 400 bp (Figure 2.4). Furthermore, not only the size of the additional

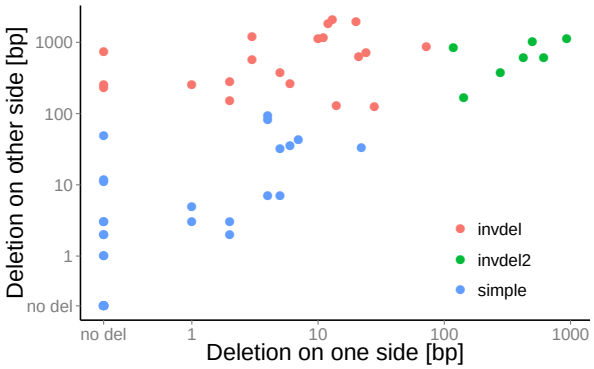


Figure 2.5: Size of deletions that flank inversions. Size of deletions (detected via LAST-split, see method in Section 2.4.2) next to inverted sequences in all assembled loci that were classified as "simple inversion", "inversion flanked by deletion", or "inversion flanked by multiple deletions".

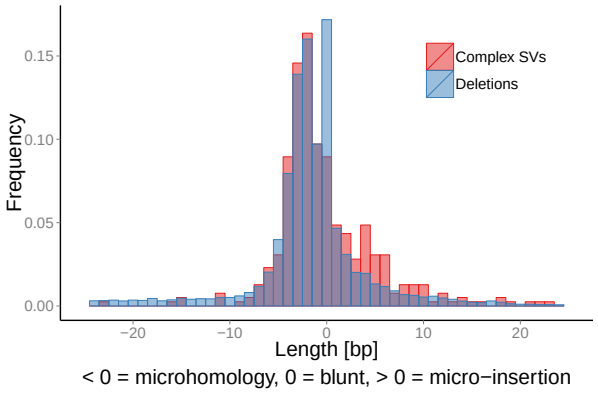


Figure 2.6: Breakpoint characteristics of complex loci vs. deletions. Micro-deletions or -duplications (a.k.a. microhomology) as identified by LAST-split are reported for assembled loci with exactly two breakpoints (199/210) as well as for a set of 12,206 single-breakpoint deletions from Sudmant, Rausch, et al. [2015].

copy but also the distance to the insertion point is highly similar, between 0.5 and 3 kb. While these sizes could be limited due to the amplicon approach, which is dependent on the capabilities of long-range PCR, the predictions made by DELLY suggest not. The paired-end calling by DELLY does not penalize the detection of larger versions of the same types of variants but reported only few, which suggests that these classes of variation indeed follow such a tight size distribution. Of course germline variants underlie selective forces; it is thus unclear whether the mechanism of origin exclusively creates this short size spectrum or whether larger inverted or promximal duplications are strongly selected against. Nevertheless, the similarities between inverted and proximal duplication suggest both to be generate by a similar, possibly replication-based mechanism (Section 1.2.2).

Moreover, I also observed an interesting feature of the deletions flanking some inversions. The classification into "simple inversion", "inversion flanked by deletion", and "inversion flanked by multiple deletions" relies on a rule-of-thumb estimate during visual inspection. As Figure 2.5 suggests, deletions flanking inversions display a continuous size spectrum instead of falling into clear categories (deletion vs. no deletion). This again hints at a common mechanism generating these three classes of inversions, which sloppily handles the ends of the to-be-inverted sequence and hence creates more or less large deletions at one or both ends.

2.3.3. Breakpoints analysis suggests mechanisms other than NAHR operating

In order to further explore the mechanism creating complex inversions I searched for homologous sequence around the breakpoints that would be a clear sign for NAHR. Except for a few low-complexity repeats no true homology could be found around the breakpoints. This is, however, not surprising, as the paired-end read-based detection strategy by Delly is known to have lower power in repetitive (hence not uniquely mappable) regions.

I continued the investigation of mechanism by inspecting breakpoints for the presence of micro-homology. I selected 199/210 loci with exactly two breakpoints (the remaining ones being "highly complex") and analyzed the breakpoints of the sequenced alleles in terms of their matches to the reference sequence. This analysis can unravel small insertions of non-templated sequence, exact breakpoints or the presence of micro-homology (see Figure 2.7 for example). The definition of such base-level features is highly susceptible to alignment methods and –parameters. To make statements about the breakpoint characteristics of the set of complex SVs[5] I compared them to another set of SVs using the exact same methodology. I employed a data set of short read assemblies of deletion breakpoints available from the 1000 Genomes Project [Sudmant, Rausch, et al., 2015], parts of which had been analyzed beforehands [Abyzov et al., 2015]. I filtered to this data set down to 12,206 single-breakpoint events without additional complexity and compared the breakpoint characteristics of complex SVs and deletions as shown in Figure 2.6. Interstingly, deletions and complex SVs did not differ strongly in the amount and length of micro-homology. However, the distributions in general differed significantly (p-value < 0.001, Kolmogorov–Smirnov test) with complex SVs harboring less exact breakpoints than deletions and slightly longer stretches of non-templated insertion of DNA (median 1bp for deletions vs. median 4bp for complex SVs, p-value < 0.001, Wilcoxon-Mann-Whitney test).

In summary I identified multiple features of the set of complex SVs that provide subtle hints towards the mechanisms they originate from. Clearly they arise independent of homology, a fact that is tightly coupled to the detection method applied. Among the non-homology dependent mechanisms there is no evidence of a micro-homology-based mechanism such as FoSTeS/MMBIR being favored. Nevertheless, there are significant differences compared to the class of deletions, which acts as a proxy for non-complex SVs, namely in the number and length

[5] I refer to the set of 199 loci as "complex SVs" from now on although technically it contains also simple (non-complex) inversions

of non-templated sequence insertions. Notably, I described multiple different classes of SVs, of which, as the data suggests, many appear to be generated by the same mechanisms: inverted and non-inverted, proximal duplications follow exactly the same tight size distribution, and the deletions that sometimes flank inversions span a continuous size range rather than showing a binary pattern (present or absent). At last, the insertion of a few base pairs of DNA at the breakpoints of complex SVs could be traces of NHEJ or BIR mechanisms operating [Hastings, Lupski, et al., 2009; Carvalho et al., 2016], yet I deemded the singal not strong enough to make these distinctions.

2.4. Results III: Maze—a tool for match visualization and breakpoint inspection

A common theme of this thesis is how the development of computational methods for analysis and, notably, visualization of biological data are instrumental to accelerate scientific research. This chapter shall not fall behind in that respect. Together with Markus Hsi-Yang Fritz I developed a tool to easily visualize the alignment of two DNA sequences and to inspect the breakpoints of potential matches between them. Termed Maze, this piece of software resulted from the necessities I faced during the analyses described in Sections 2.2 to 2.3. It is available online at `https://github.com/dellytools/maze` or, in a version with reduced functionality, at `https://gear.embl.de/maze/`. In this section I describe the method, design decisions, and its use cases.

2.4.1. Interactive match visualization in the web browser

Maze is designed to quickly compare two nucleotide sequences against another. As an easy-to-grasp, visual comparison method we chose dotplots, which have a long-standing tradition in sequence analysis [Fitch, 1969; Gibbs et al., 1970]. The principle of dotplots is to graphically highlight regional identity of two different sequences or of the same sequence to itself in a 2-dimensional matrix. The top panel of Figure 2.2 shows such a graphical comparison of two similar sequences, and Figure 2.1 shows examples of sequences carrying different types of inversion compared to a non-inverted sequence. In contrast to a sequence alignment, which generally aims at computing the constellation of two sequences having the closest distance, a dotplot simultaneously compares all substrings of both sequences to one another.

This idea has recurred many times in the history of bioinformatics with different goals in mind. For example, some dotplots originally designed for peptide comparisons color matches based on exact, single-character matches. Dotplots for DNA comparison, on the other hand, tend to require longer matches or calculate an average of the identity of multiple characters within a sliding window. This is to reduce the background noise that naturally occurs when comparing sequences with a small alphabet, such as DNA[6]. Furthermore, dotplots have been optimized for differently scaled inputs. A genome-scale comparison[7] likely requires a different design and different data structures to find matches than the comparison of many small sequences.

maximal unique match
Continuous stretch of exact matches that is not included in any other such stretch, hence cannot be expanded to either side. A maximal unique *match additionally require the match to occur exactly once in both sequences*

The dotplots generated by Maze display maximal unique matches or, alternatively, just maximal matches). Those can be fast computed by MUMmer [Kurtz et al., 2004], which internally uses an efficient suffix tree data structure. They are then filtered to a minimum length and displayed in a web browser using Javascript. This is basically the same as offered by MUMmer's dotplot function[8]. Although maximal unique matches have many of the aforementioned advantages for DNA comparisons, they are bad at handling small insertions or deletions as they show only exact matches. This is why the minimal length offers a trade-off between sensitivity (shorter matches) and reduction of background noise (longer matches).

The selling point of Maze is its ease of use tailored towards quick inspection of long-read sequencing data of, for example, amplicons. The sequences can be entered as standard FASTA or FASTQ files through a user interface in the web browser. Dotplots are calculated on the fly and displayed as in Figure 2.7 (left panel). Maze also computes matches to the reverse complement sequence and displays those in red. Besides a one-vs.-one mode, Maze can compare many sequences to a single one, just as I needed in Section 2.2 to inspect long reads of amplicon sequencing. Moreover, Maze can also compare multiple pairs of sequences, a scenario that became relevant once optimal assemblies for diferent loci were available in Section 2.3. In those sections an early version of Maze was used and plots were written as PDF files instead of displayed in the web browser.

The current version of Maze uses the popular d3 library[9] to display dotplots. A

[6]For instance, a single-character comparison of two random DNA sequences would fill as much as one quarter of the matrix

[7]See for example `http://last.cbrc.jp/hg18-mm9.png` for a human vs. mouse comparison

[8] More precisely, by the *mummerplot* command. Examples can be found online at `http://mummer.sourceforge.net/examples/`

[9]Found online at `https://d3js.org/`

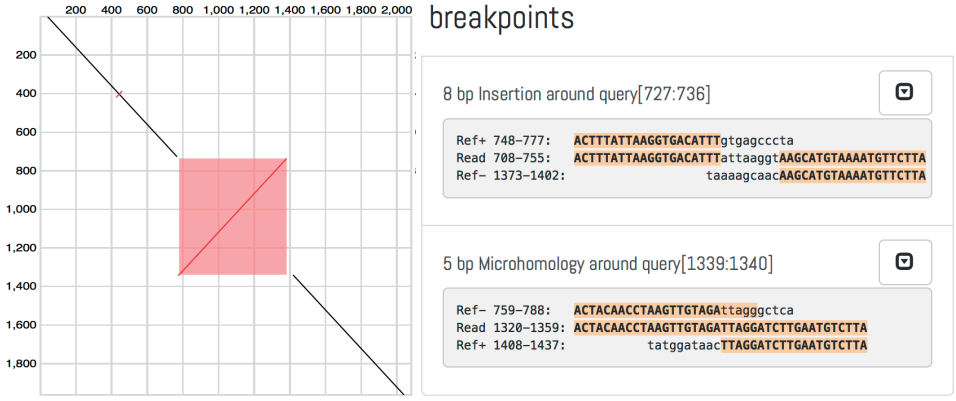


Figure 2.7: Screenshot from MAZE. On the left, MAZE displays a dotplot of two sequences (here inversion assembly, vertically, vs. its reference locus, horizontally) and highlights sequence alignments (red). On the right, the sequence alignment around the breakpoints is highlighted and classified into cases of micro-insertions, blunt ends or micro-homology.

user can readily adapt the minimum length and type of matches to be calculated via in-app settings. The dotplot can be panned and zoomed-in using the computer mouse, saved as PDF, and by using keyboard shortcuts a user can quickly browse through multiple dotplots. At last, after I successfully utilized MAZE 's dotplots to solve the inversion validation mystery of Section 2.2, I extended its functionality by an analysis module dedicated to breakpoints of sequence matches, as I describe in the upcoming section.

2.4.2. Breakpoint identification and characterization

Any analysis of SV breakpoints depends a reasonable definition of what breakpoints actually are and how to determine them. I found a convincing solution in an algorithm called LAST-split [Frith et al., 2015]: given a sequence and a reference genome, it finds the optimal alignments that maximally cover the query sequence without allowing overlaps. This idea treats the contiguous sequence (from assembled reads, in this case) as the true allele and tries to split it into consecutive parts that can be assigned to their most likely origins in the reference genome. This also means that while the sequence is split into consecutive stretches, matches on the reference are unconstrained and can potentially overlap, e.g. in the case of a duplication. This method was specifically designed with inversions and complex rearrangements in mind and therefor determines breakpoints accurately in the given scenario.

The breakpoint module of Maze runs Last-split on the two given sequences to determine breakpoints. First of all, the aligning parts are highlighted inside the dotplot as rectangles that become visible when a user crosses the respective entry with the computer mouse (Figure 2.7, on the left). In addition to that, a separate field on the right displays the exact sequence alignment in form of text (not shown in Figure 2.7). Given the breakpoints from Last-split, Maze can compute the characteristics of that breakpoint in respect to the reference sequence. For example, if the two consecutive parts of the sequence match to regions of the reference genome that slightly overlap, the exact position of the breakpoint is ambiguous. This is what is commonly understood as micro-homology (see Section 1.2.2 and compare to figure 1 in Hastings, Lupski, et al. [2009]) and can be evidence for a micro-homology-dependent mechanism forming the SV. If the two regions are not exactly contiguous in the sequence because no according match to the reference is found, an insertion of other, possibly non-templated DNA occurred. Maze graphically arranges the alignments around the breakpoints, detects such features and reports them in a separate panel on the right, as can be seen in Figure 2.7 (right panel). The respective matches are highlighted within the alignment to direct the user's attention to the right position belonging to the breakpoint. A current shortcoming is that the extension of this alignment does not allow gaps. In order to do so, another alignment would have to be computed, yet of a region which was deemed irrelevant by the Last aligner. Additionally to the breakpoints on the query sequence, a third panel focuses on the breakpoints of the reference sequence (not shown in Figure 2.7). Neighboring reference matches also can overlap, meaning the affected bases are duplicated in the sequence, or can be distinct with a small gap in between. These micro-duplications and micro-deletions are similarly displayed up to a size of 100 bp by Maze.

I used an early version of this breakpoint analysis tool to obtain the results in Section 2.3. Together with custom scripts the analysis was automated to process a large number of loci sequentially, allowing me to generate Figures 2.5 to 2.6. In the current available version[10] of Maze this functionality was abandoned for the sake of simplicity. I utilized expandable fields to support a user in showing or hiding certain information and adopted the modern look and feel of the main window of Maze that was designed by Markus Hsi-Yang Fritz.

In summary, Maze is a user-friendly and easy-to-install software specifically designed for inspecting long reads or assemblies in the range up to 50 kb. It allows a user to visually browse through their sequence data to (1) judge the quality

[10] https://github.com/dellytools/maze/commit/106803d350719188181ce39cbb5c832822b724da

(compare, for example, PacBio to ONT MinION data in Figure 2.1), (2) distinguish successful sequencing reads from artifacts and (3) discover interesting alleles such as complex rearrangements. The additional breakpoint module of Maze comes into play whenever a user is interested in characteristics of the breakpoint junctions of SVs. Maze uses current web browser-based technology and a modern design to offer a responsive user experience and was instrumental in gaining the scientific insights described in Sections 2.2 and 2.3.

2.5. Conclusions

The third and final phase of the 1000 Genomes Project created an unprecedented resource of variation in the human population [1000 Genomes Project Consortium, 2015]. One of the major achievements was the integrated map of structural variation [Sudmant, Rausch, et al., 2015], for which various classes of SVs were detected and genotyped in 2,504 individuals. My colleagues and I were responsible for the set of inversions, which are notoriously difficult to ascertain as also other population-scale studies note [Kidd et al., 2008; Hehir-Kwa et al., 2016]. With 768 loci it is to date the most comprehensive high-accuracy inversion-type call set of the human population.

These inversions had been predicted from low-coverage paired-end sequencing data, but PCR-based validation experiments proved difficult. For a while it was a major mystery why inversions that seemed to be accurately predicted could not be experimentally validated. This mystery was solved with the help of long-read sequencing, notably PacBio and ONT MinION sequencing. A subset of 208 predicted loci (all below 10 kb in size) were amplified together with surrounding genomic DNA, sequenced and then visualized via dotplots. This way, more than 80% of analyzed loci could be verified, exemplifying the potential of third-generation sequencing.

Strikingly, we found that the vast majority of loci turned out to be complex instead of simple. This realization explained the initial dilemma of PCR-based verification experiments and it revealed a previously unexpected amount of complexity. Besides simple inversions, we discovered inverted duplications as the most abundant class of complex SVs with inversion signature, but also inversions flanked by deletions on one or both sides and variants with even higher complexity (shown in Figure 2.1). The existence and amount of such complex variants has since been confirmed and even extended to more complex classes by independent studies (notably Collins et al. [2017], but also Sanders et al. [2016] and English et

al. [2015, fig. 7]).

A presumably abundant source of inversions is NAHR, which creates inversions embedded in inverted repeats [Carvalho et al., 2016]. For example, up to 12% of the human genome were estimated to be susceptible to NAHR-mediated inversions [Dittwald et al., 2013]. However, paired-end mapping strategies are known to have limited power when SV breakpoints align with repetitive sequence (Section 1.4.3). Korbel et al. [2007] attributed a mere 14% of paired-end-detected SVs to the NAHR mechanism and the SV calls (notably inversions) of the 1000 Genomes Project underlie a similar detection bias. In contrast, studies using alternative technologies reveal higher levels of NAHR-mediated inversions: Kidd et al. report 67% of inversions identified to be flanked by inverted repeats using fosmid libraries (Kidd2008), Chaisson find et al 68% of 34 inversions detected via PacBio assemblies [Chaisson et al., 2014], and also optical mapping reportedly has a greater ability to detect SVs embedded in repetitive DNA [Teague et al., 2010, fig. 6]. Furthermore, studies using breakpoint-independent approaches report inversions in a much larger size range (i.e. a minimum of 200kb in Bansal et al. [2007]; a median size of 176kb in Sanders et al. [2016] using Strand-seq).

It can thus be noted that the inversions predicted by the 1000 Genomes Project are complementary to previously revealed inversions, both in terms of size and in the high level of complexity. While generally the functional relevance of this class of complex SVs remains to be shown, a few variants overlap genes as for example *Ras* homologue family member H *(RHOH)* [Sudmant, Rausch, et al., 2015]. NAHR appears not to be relevant in creating these SVs, but the effective mechanisms still remain obsure. A higher number of micro-insertions, as I found through detailed breakpoint analysis based on nucleotide-resolution assembly sequences, could hint at NHEJ or a replication-based mechanism (e.g. BIR) operating. Collins et al. [2017], who did a more extensive analysis of complex rearrangements, suggest the synchronous repair of simultaneous double-strand breaks as seen in cases of chromothripsis might play a role here.

Finally we learned that complex classes of inversions are present in the human genome and that this complexity should be taken into consideration in future efforts of SV discovery. Delly is one of the first tool that is capable of doing so, yet manual inspection of complex SVs is still highly recommended (also reported for other tools [Collins et al., 2017]). Maze, which resulted as a side product from this work, is a handy tool to inspect complex SVs sequenced with long reads and can be used in the future to further investigate complex loci.

3 Effects of SVs on Gene Expression and Chromatin Organization in *D. Melanogaster*

In the project described here I teamed up with Yad Ghavi-Helm, Aleksander Jankowski and Eileen Furlong to study the functional impact of SVs, specifically of large chromosomal rearrangements, in the model organism *Drosophila melanogaster*. Recent sequencing technologies, here Hi-C, were of fundamental importance for the characterization of SVs, which was an essential step of the project. All wet lab experiments described below were performed by Yad Ghavi-Helm with the help of Rebecca Rodríguez Viales. Aleksander Jankowski implemented all Hi-C-related analyses, which are hence only depicted briefly. All other computational analyses including all figures are my own work if not explicitly stated differently. Supplementary information can be found in Appendix C. At the time of writing a manuscript of this study was in preparation [Ghavi-Helm, Meiers, Jankowski, et al., 2018].

3.1. Introduction

3.1.1. Impact of SVs in health and disease

Population-scale studies revealed that SVs are a major contributor to genetic variation in the human population [Conrad et al., 2010] – in fact they contribute more base pair differences than SNVs [Sudmant, Rausch, et al., 2015]. Besides their abundance in healthy individuals, SVs had earlier already been recognized for their role in a number of diseases, as F. Zhang, Gu, et al. [2009], Weischenfeldt et al. [2013], and Carvalho et al. [2016] review in more detail.

The consequences certain SVs imply can range from purely molecular phenotypes, such as altered gene expression, to little severe (e.g. the efficiency of starch

digestion depending on the copy number of *AMY1* [Perry et al., 2007]) or more severe phenotypes (e.g. red-green blindness, caused by CNVs on chromosome X [Nathans et al., 1986]), to disease or to an increased susceptibility towards a disease. The latter can include complex diseases such as autoimmunity [Fanciulli et al., 2007] or autism [J. Sebat et al., 2007]. Especially CNVs, the most-frequently studied type of SV, have been linked to Mendelian diseases [F. Zhang, Gu, et al., 2009, see table 2] and are frequently involved in cancer [Beroukhim et al., 2010]. Yet SV classes that contribute to disease are by far not restricted to only deletions or duplications. Aneuploidy, to start with, can play a major role in diseases such as Down syndrome, which was among the earliest identified genetic diseases [Lejeune et al., 1959]. Also inversions were identified as the cause of several Mendelian diseases [Feuk, 2010]. Recurrent translocations, often creating gene fusions, are known to be driving many cancer types [Mertens et al., 2015] and, at last, transposable elements have been noted be to play a role in cancer [Burns, 2017].

The mechanisms by which SVs induce these phenotypic consequences are manifold. Breakpoints of SVs that fall into a gene may cause its loss of function. Aforementioned gene fusions can create novel chimeras that potentially exert fundamentally different functions from the original genes [Mertens et al., 2015]. Copy number gains or losses lead to dosage imbalance, with critical consequences for many cancer-associated genes [Fehrmann et al., 2015], and a heterozygous deletion or a LOH event can reveal recessive mutations that had been rescued by a functional allele beforehand. Furthermore, SVs may impact on the molecular level, e.g. on gene expression, which can be studied even if not associated with an observable phenotype. In the context of expression quantitiative trail loci analyses, more and more such effects have lately been found [Sudmant, Rausch, et al., 2015; C. Chiang et al., 2017].

expression quantitiative trail loci *Genomic loci with different alleles that influence the expression of genes (typically a single gene) that are not neccessarily in close proximity*

However, causative SVs need not affect the actual gene itself, but could target regulatory sequences in proximity or elsewhere in the genome. This is typically the case for aforementioned expression quantitative trait loci, yet the (then very surprising) discovery was already made in 1979 in a row of heritable blood diseases named thalassemia [Fritsch et al., 1979]. These *position effects* have long been noted to have an impact in health and disease, yet the underlying mechanism are in many cases not yet understood [Kleinjan et al., 2005].

Over the past years the community has gained more insight into a novel mechanism which SVs neither affects genes nor regulatory sequences themselves. Instead, this mechanism re-models the three-dimensional organization of chromatin, as I elaborate further below.

3.1.2. Three-dimensional chromatin conformation

The advance of chromatin conformation capture techniques, notably Hi-C (introduced in Section 1.3.3), revealed a feature of the spatial organization of chromatin named topologically associating domains (TADs) [Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012; Rao et al., 2014] (see Appendix C.1 for comments on the cited literature). TADs are physical domains of DNA that are characterized by an increase in chromatin interactions inside them and a relative insulation of contacts across distinct domains. These structures show up as characteristic triangles in contact frequency maps, as illustrated in Figure 3.1. Despite molecular differences in, for example, the involved architectural proteins, the phenomenon of TADs was observed in a range of species across the tree of life ranging from humans and mice to *Drosophila melanogaster* and *Caenorhabditis elegans*. This suggests that these structures are a universal feature of metazoan genomes [Dekker et al., 2015].

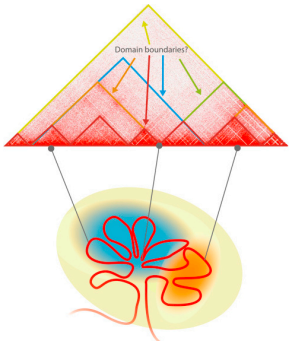


Figure 3.1: Schematic of topologically associating domains. Characteristic triangles in a mammalian Hi-C map (top) belong to spatial domains of DNA inside the nucleus (bottom). Figure taken from "Structure meets function: How chromatin organisation conveys functionality" [Ruiz-Velasco and Zaugg, 2017] licensed under Creative Commons Attribution 4.0.

TADs have gained extraordinary attention in the field and their potential function has been discussed and reviewed intensely in recent years [Gibcus et al., 2013; Gorkin et al., 2014; Sexton et al., 2015; Denes Hnisz et al., 2016; Ruiz-Velasco, Kumar, et al., 2017, among others]. Key characteristics of TADs are that the expression of genes within TADs tends to be orchestrated [Le Dily et al., 2014; Nora et al., 2012, see figure 4b] and that they align with epigenetic features such as histone marks [Nora et al., 2012] and DNA replication timing [Pope et al., 2014; Le Dily et al., 2014]. Furthermore, TAD boundaries are associated with the insulator element-binding protein CTCF, enriched in house-keeping genes and conserved across cell types [Dixon et al., 2012; Rao et al., 2014; Schmitt et al., 2016].

Well-characterized long-range interactions between promoters and enhancers, which can be revealed through 3C-based techniques, appear to be confined within TADs (reviewed by Smallwood et al. [2013]). Additionally, units of enhancers and promoters with correlated activity were found to align well with them [Shen et al., 2012]. It is unclear whether the DNA loops connecting enhancers and promoters are cause or consequence of active expression, but it was observed that many such contacts establish long before gene activation [Ghavi-Helm et al., 2014]. Further research using functional assays supported that genes are co-regulated within TADs. Notably, Symmons et al. [2014] inserted reporter genes at several hundred sites of the mouse genome and observed tissue-specific activity in spatial blocks correspond to TADs.

Together, these results suggest that TADs exert a crucial regulatory function. A current hypothesis is that they confine promoter-enhancer interactions inside,

as X. Ji et al. [2016] call them, insulated neighbourhoods. As I described in the next section, a series of recent perturbation studies gave additional substance to this hypothesis.

3.1.3. Consequences of disrupting chromatin conformation

Several studies followed up on the question what would happen when the boundaries between TADs were disrupted. In line with the hypothesis of insulated neighborhoods, they found that the merging of two TADs can alter the "search space" of an enhancer element, which can then suddenly drive expression of another gene to non-physiological amounts. This mechanism was termed *enhancer adoption* [Lettice et al., 2011] or *enhancer hijacking* [Northcott et al., 2014].

The cause for a TAD boundary disruption can be a SV that deletes or inverts the genomic region harboring this boundary. This was quite remarkably shown in various cancer types including medulloblastoma [Northcott et al., 2014], lung cancer and colorectal cancer [Weischenfeldt et al., 2016] and acute lymphoblastic leukemia [D. Hnisz et al., 2016]. Furthermore, computational studies linked this mechanism to genetic diseases via mining of public databases and showed, among other things, that around 10% of known disease-associated deletions potentially function via an enhancer adoption mechanism [Ibn-Salem et al., 2014; R. Li et al., 2016; Zepeda-Mendoza et al., 2017].

Moreover, as Lupiáñez et al. [2016] and Krijger et al. [2016] review, a number of studies specifically tested this hypothesis on particular genomic loci by genome editing using CRISPR/Cas [Doudna et al., 2014]. For example, Guo et al. [2015] altered the binding sites of an architectural protein in proximity of an enhancer element and observed looping as well as ectopic gene expression across the boundary. Similarly, Narendra et al. [2015] observed heterochromatin spreading when they deleted such a boundary at the *Hox* locus in *Drosophila*. In fact, already Nora et al. [2012] observed this mechanisms when they studied an additional mouse line with a TAD boundary deletion and observed a merged TAD and ectopic gene expression instead of the two clearly separated TADs and regulatory insulation seen in wild type mice.

Recently, Lupiáñez et al. [2015] impressively replicated limb abnormalities in mice that were known from human genetic diseases. They first showed that these abnormalities originated from misregulated gene expression in a developmental stage, which was caused by TAD boundary deletion. When they engineered similar deletions in mice using the CRISPR/Cas system, they observed a change in chromatin conformation, ectopic gene expression across the boudary and indeed

the same phenotype of limb malformations that had been observed in human patients.

In conclusion, all these studies provide strong evidence for TADs to play a central role for gene regulation. When TADs break—and fuse with another TAD—genes encounter a new regulatory environment including enhancers, which may lead to severe misregulation via en enhancer adopiton mechanisms.

3.2. Design of the study

In this study we set out to understand how genomic aberrations caused by large chromosomal rearrangements and other SVs can affect the regulation of genes. Specifically, we were interested in rearrangements that lead to the disruption of TADs or the formation of new TADs. Unlike previous studies (outlined in Section 3.1.3), we wanted to explicitly focus on a phenotypically healthy system where no dominating effects, but rather modest changes in molecular phenotypes were to be expected. We think that this aspect has been vastly neglected by the previous studies, which had investigated rather pathological situations. Hence our work will be complementary and open a new perspective on the functional impact of TADs. Moreover, instead of selecting a single locus we aimed at testing multiple rearrangements in a genome-wide fashion to gain a broader understanding on the generality of such effects.

3.2.1. Balancer chromosomes carry large rearrangements

We found a suitable model system for this task in so-called balancer chromosomes of *Drosophila melanogaster*. These naturally derived chromosomes have a long tradition in fly genetics and are frequently used as a tool to keep recessive lethal mutations from being lost from the population. Three relevant features allow them to perform this task: (1) they carry recessive lethal mutations so that homozygous offspring will not live, effectively balancing the alleles in a population. This further requires that haplotypes remain intact, so consequently (2) balancer chromosomes suppress recombination by disrupting homologous pairing. This is achieved by the introduction of multiple large inversions via, for example, X-ray mutagenesis. (3) Balancer chromosomes include dominant marker alleles so that carriers can be readily detected based on their phenotype in adult stage. Such balancer chromosomes are nowadays available for all the major chromosomes 2, 3, and X. Interestingly, despite their common usage, balancer chromosomes had only recently been characterized via whole-genome sequencing [Miller et

X-ray mutagenesis
*Balancer chromosomes date back to the work of Hermann Joseph Muller, who in the 1920s studied mutagenesis through X-ray radiation [Muller, 1928] and received a Nobel prize in 1948. See also this blog post from Laurence Moran (*goo.gl/zGi76W*)*

al., 2016; Miller et al., 2018].

We chose balancer chromosomes for chromosome 2 and 3, namely Curly of Oster (CyO) and TM3 [Tinderholt, 1960] to create a double balancer line, named F_1. See Figure 3.2 for an overview of the *D. melanogaster* genome of both the wild type line and the double balancer line utilized in this study. For both balancer chromosomes the major rearrangements had been previously characterized by karyotyping (see note[1] and Figure 3.8). With two balancer chromosomes, we effectively increased the number of genomic rearrangements to be studied, which was an important consideration early on in the project. Together, both balancer chromosomes carry 16 breakpoints of large, partly centromere-spanning inversions and we expected them to contain additional sub-microscopic SVs related to their mutagenic origin. Accordingly, one of the first steps I carried out in this study was a deep characterization of the mutational landscape of the balancer chromosomes, which is presented in Section 3.3.

CyO *Chromosome 2 balancer, derived from Oster [1956] and based on work of Ward [1923] (see Miller et al. [2018]), which carries the Cy allele, hence showing a characteristic phenotype of curly wings in adult flies*

3.2.2. Studying *cis*-regulation through allele-specific gene expression and haplotype-resolved chromatin conformation

To study the effects of SVs on gene regulation we wanted to compare global gene expression in the balancer chromosomes in comparison to their wild type homologues. This could principally be done by comparing two different fly strains, one carrying genomic rearrangements and one without. However, this approach would make it difficult to distinguish *cis*-regulation, i.e. the effect of alterations of the chromosome itself, from *trans*-regulation such as an altered expression of some transcription factor. This is why a central foundation of the study was to compare wild type chromosomes and balancer chromosomes within the same fly line via allele-specific expression (ASE) analysis. This is achieved by measuring gene expression separately for both alleles, which differ by naturally occurring variation, notably SNVs. Since we were able to resolve chromosome-long haplotypes in this study, alleles could even be aggregated within each gene and across all genes. We could thus distinguish expression changes into up- or down-regulation of a balancer allele (in respect to the wild type allele). Sections 3.4.1 and 3.4.2 provide more detail on the procedure. We assume in the ASE analysis that the vast majority of SNVs simply tag alleles but do not have an effect on their regulation, or that this effect is negligible in comparison to larger rearrangements.

Another key point of the study was to explore the three-dimensional chromatin conformation of the highly rearranged balancer chromosomes and to link it to observed ASE. We hence performed a Hi-C experiment and again used SNVs to

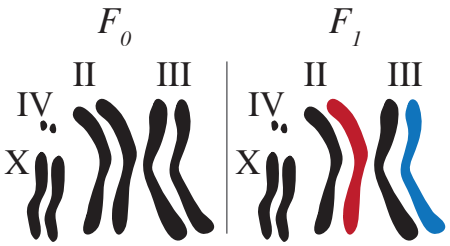


Figure 3.2: Genome overview of balanced fly line. Schematic of the four chromosomes of (female) *D. melanogaster* flies of the wild type line F_0 and of the double balancer cross F_1. CyO red, TM3 blue.

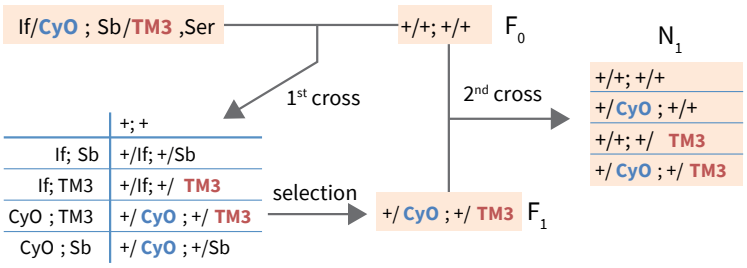


Figure 3.3: Crossing scheme. The fly lines used in this study are derived from a homozygous wild type line, denoted by +/+;+/+ or F_0 and a double balancer line (*If/CyO;Sb/TM3,Ser*). After a first cross adult flies are selected for markers of both balancer chromosomes, yielding an F_1 generation (a.k.a. double balancer cross). A backcross with the initial wild type line generates a pool of four different genotypes (N_1) of which on average 25% of both chromosome 2 and 3 are balancer chromosomes.

distinguish fragments belonging to balancer- or wild type chromosomes. As a matter of fact the resulting haplotype-resolved maps of contact frequency can additionally be utilized to characterize genomic rearrangements, as I describe in Section 3.3.3.

Early on we made a decision to study fly embryos (instead of adult flies) for mainly two reasons: To begin with, *D. melanogaster* embryos are well described, there is an outstanding amount of external data available [Gramates et al., 2017; Celniker et al., 2009], and the Furlong lab has years of experience on embryogenesis Furlong et al., 2001; Ghavi-Helm et al., 2014, among others. Secondly, it is experimentally very difficult (if not infeasible) to extract intact nuclei from adult flies, which is a crucial requirement for Hi-C experiments. However, collecting double balancer embryos is not directly possible based on their phenotypic markers, which are only expressed in an adult stage. We hence decided to collect fly embryos from a backcross of the double balancer line (F_1) with the original wild type line (F_0). The resulting generation, termed N_1, is a mix of genotypes as shown in Figure 3.3, in which effectively 25% of chromosomes 2 and 3 are balancer chromosomes. This cross was used both to measure ASE and chromatin conformation.

In the subsequent sections I describe the methodology and our findings on the effect genomic rearrangements have on chromatin organization and gene regulation.

3.3. Results I: Mutational landscape of balanced chromosomes

We performed deep paired-end whole-genome sequencing (WGS) (approximately 100 x and 200 x, respectively) of both the wild type line (F_0) and the double balancer line (F_1) with read lengths of 300 and 200 bp on an Illumina MiSeq platform. To be able to better resolve SVs we additionally sequenced mate pair libraries of both samples with a read length of 2·100 bp and a median insert size of circa 4 kb. I utilized this data as well as Hi-C data (experiments described in Section 3.6.1) to characterize the mutations present on balancer and wild type chromosomes in respect to a common reference genome (DM6). Below, the mutational landscape of the balanced chromosomes is described.

3.3.1. Single nucleotide variants

I utilized WGS data of both F_0 and F_1 cross simulataneously to call SNVs and small indels with FreeBayes [Garrison et al., 2012]. The comparison of genotypes of the homozygous wild type line and the heterozygous cross enabled me to assign mutations to individual haplotypes: if, for example, a SNV is heterozygous in the cross and homozygous alternative in the wild type sample, the variant is located on the wild type chromosome and not on the homologous balancer chromosome.

Of the 761,348 detected SNVs on chromosomes 2 and 3, 38.9% could be assigned to the balancer chromosomes, 29.5% to the wild type chromosomes and 29.8% were shared between both. Only a fraction of 1.8% of SNVs did not match any of the expected genotypes, which can mostly be attributed to regions in the wild type chromosomes not being fully homozygous (Table 3.1). These numbers imply that, on average, there is one SNV every 210 bp that distinguishes balancer and wild type haplotypes. This density was an important parameter as SNVs were utilized later to separate sequencing data such as RNA-seq into haplotypes (Sections 3.4.2 and 3.6.1).

The number of balancer-specific SNVs was around 1.3 times higher than wild type-specific ones. Moreover, when I compared the observed variation to the *Drosophila* reference panel (DGRP) [Mackay et al., 2012; Huang et al., 2014], a panel of fully sequenced inbred *D. melanogaster* lines derived from a natural population, I found that a striking majority of SNVs is represented in the panel, yet balancer-specific SNVs to a lower fraction (86.1%) than wild type-specific ones (92.1%). These observations are not surprising since balancer chromosomes, due to their inability to undergo recombination, accumulate and likely tolerate more

Genotype	Chr2	Chr3	ChrX
balancer-specific	2717.1	2717.6	956.3
wild type-specific	2247.0	1905.6	844.0
common	2338.7	1873.7	2212.3
wt heterozygous	98.6	120.4	176.2
errorneous	17.6	16.7	5.7

Table 3.1: Number of SNVs per Megabase. The majority of SNV calls are in concordance with the study design (balancer-specific, wild type-specific, or common), except in short regions of remaining heterozygosity of the wild type chromosomes and a negligible number of false calls. Note that ChrX is not balanced but also paired with a homologue from a divergent line.

mutations over time than normal chromosomes [Araye et al., 2013].

Moreover, I had a closer look at the distribution of base substitutions, i.e. the mutation spectrum of the balancer chromosomes. The idea behind this is that balancer chromosomes, which had been derived by X-ray mutagenesis, could exhibit distinct pattern in the distribution of base substitutions. However, after removing SNVs shared with the DGRP I did not observe any striking differences in the mutation spectrum of SNVs between wild type and balancers (see Figure C.1; detailed methods in Appendix C.3).

mutation spectrum *As a summary of the total set of mutations I count the relative frequencies of different base substitutions (e.g.* C>T*) and their tri-nucleotide contexts (one base up- and downstream). The spectrum of somatic SNVs is often analysed in cancer genomics in this way, in order to extract* mutational signatures *that are linked to mutagenic processes*

3.3.2. Copy number variants

Deletions I called deletions from WGS data of the F_0 and F_1 samples using Delly, which employs read pair and split-read signatures, and developed an extensive set of filters to reduce false positive predictions. For both reasons of filtering and with validation in mind I categorized deletion calls into three size ranges: (1) Below 50 bp: these calls, predicted from split reads, historically belong into the indel category and an additional set in this size range was predicted with FreeBayes. (2) 50–159 bp: a size range where predictions based on a split read signature typically yields reliable results, yet these calls are large enough for validation with PCR. (3) 160 bp and larger: in this size range additional filters based on a read depth signal are applied prior to PCR-based experimental validations. The latter is generally recommended for CNV predictions based on a paired-end read signature and was highlighted also for other tools than Delly [Layer et al., 2014, for example]. I developed an ad hoc filtering strategy for larger (160+ bp) deletion calls, which compares locally normalized read depths between of between heterozygous and homozygous samples (see Appendix C.4 for details). As a con-

sequence, shared deletions are excluded in this size range. This is not a limitation to our study, though, as we are primarily interested in genetic variation distinguishing the homologues.

Deletion subset	#Calls	V	I	Empirical FDR
<50 bp	3,072	0	0	*unknown*
50 - 159 bp	737	24	1	4%
160+ bp, low mappability	75	25	0	0%
160+ bp, high mappability	395	24	1	4%
Weighted average	4,279			*0.375%

Table 3.2: Deletion validation via PCR. Number of calls as well as number number of validated (**V**) and invalidated (**I**) loci as determined by PCR. The proportion of validated calls determines the empirical FDR. * Calls below 160 bp were not PCR-validated and did not enter the weighted emprical FDR)

After that, Yad Ghavi-Helm performed PCR verification experiments on 75 randomly picked loci for three subsets of deletion calls (Table 3.2). A locus was considered validated if the resulting PCR bands of both F_0 and F_1 sample matched the expected sizes; otherwise it was classified as invalidated. Common variants were not submitted to PCR validations to allow comparison between homozygous and heterozygous samples. The final deletion call set based on Delly predictions contained 4,279 calls. In order to get a single set of deletions, I chose a lower size cutoff of 15 bp and merged the Delly call set with small deletions predicted by FreeBayes. This led to a total set of 8,340 calls on chromosomes 2, 3, and X with the size distribution shown in Figure 3.4.

Duplications Next, I predicted tandem duplications using Delly on WGS and mate pair data simultaneously (see Appendix C.5). Typically this class of SV is harder to ascertain than deletions both utilizing paired-end and read depth signals. Yet also for duplications, orthogonal information can be utilized to gain confidence in predictions. Specifically these are read depth, which should be considered in the context of local mappability, and BAF. After the application of initial filters, I generated overview plots containing this information for 352 loci on chromosomes 2 and 3. An example is shown in Figure 3.6, and further examples, including a negative case, can be found in Figures C.2 and C.3. I looked for a change in read depth between samples as well as for a characteristic change in BAF to validate calls. In a balancer-specific duplication, for instance, the read-

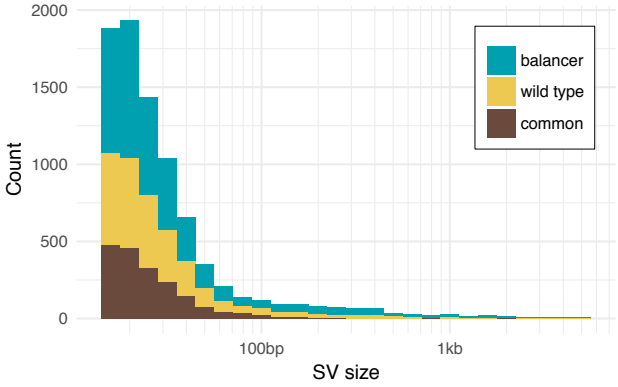


Figure 3.4: Deletion size distribution. The size distribution of final deletion calls (based on DELLY and FREEBAYES) is shown. Common deletions in the range 160+ bp were excluded. I observed a consistently higher number of balancer-specific deletions (compared to wild-type ones) and noted that this ratio increases with larger sizes.

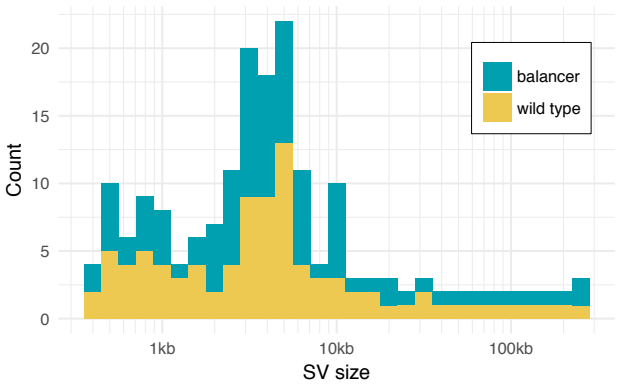


Figure 3.5: Duplication size distribution. The final duplication set consisting of manually validated tandem duplication and a set of three non-tandem duplications is shown.

depth should increase only in the F_1 sample and the BAF of SNVs should switch from 50% to around 33% and 66%.

In addition to tandem duplications predicted by DELLY I found another three non-tandem duplications by manual inspection of the BAF signal across the genome, which were then included in the set of duplications prior to manual validation. In total I verified 122 duplications in the size range from several hundred base pairs up to 258 kb using this strategy (Figure 3.5).

Summary The CNVs I found on the balancer and wild type chromosomes contain an abundance of short deletions from 15 bp up to 5.2 kb. The short size range of these deletions is in line with a study that performed SV calling in a population of flies, where a median deletion size of 178 bp at a lower cutoff of 50 bp is reported [Zichner et al., 2013]. Larger deletions likely underlie strong selection owed to the gene-dense character of the *D. melanogaster* genome. With an experimentally estimated FDR of 3.75% the deletion call set can be considered of high accuracy. Although very large deletions were detected neither in wild type nor balancer chromosomes, I did recognize a slightly higher rate of deletions on the balancer chromosomes. The ratios were 1.08 for deletions in the size range below 50 bp, 1.3 for 50 - 159 bp, and 1.9 for larger ones. Interstingly, a study in human cancer reported an increase in deletion rate as a consequence of ionizing radiation [Behjati et al., 2016], suggesting that the subtle increase in deletion number

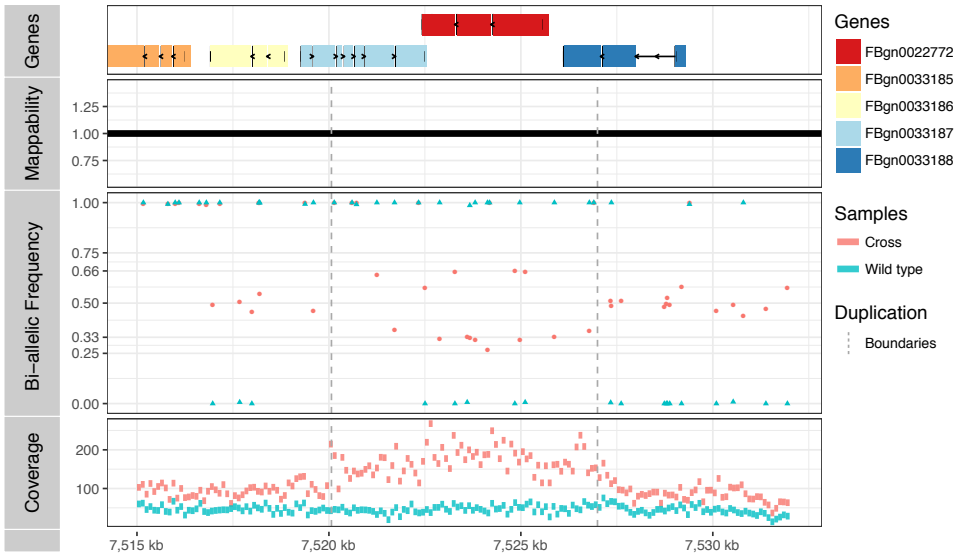


Figure 3.6: Signals used for duplication validation shown in an example. A tandem duplication at locus *chr2R:7,520,066–7,526,996* is shown, which was predicted based on paired-end read signature using DELLY (dashed lines). The additional signals included for a visual validation are (1) a mappability track (fraction of uniquely mappable reads), (2) the BAF, i.e. the fraction of reads supporting the alternative allele of an SNV, and (3) the total read coverage in 100 bp windows. BAF is typically the clearest signal to discriminate true from false copy number variants, but requires the presence of SNVs.

(notably in the larger size spectrum) could be a trace of the irradiation applied during generation of the balancer chromosomes.

Duplications, on the other hand, which underlay thorough manual validation, where distributed evenly on balancer and wild type chromosomes. An exceptionally large duplication of 258 kb was found on CyO and this duplication effectively increases the copy number of 33 genes.

3.3.3. Validation of large rearrangements using Hi-C data

I further aimed at refining the breakpoint junctions of the major rearrangements on the balancer chromosomes. These junctions had previously been known only at microscopic resolution[1]. Aleksander Jankowski generated Hi-C maps of contact frequencies for both balancer and wild type haplotypes separately in respect

[1] Described in terms of cytological bands, i.e. CyO = *2Lt-22D1 33F5-30F 50D1-58A4 42A2-34A1 22D2-30E 50C10-42A3 58B1-2Rt* and TM3 = *3Lt-65E 85E-79E 100C-100F2 92D1-85E 65E-71C 94D-93A 76C-71C 94F-100C 79E-76C 93A-92E1 100F3-3Rt.* Source: `https://bdsc.indiana.edu/stocks/balancers/balancer_bps.html`.

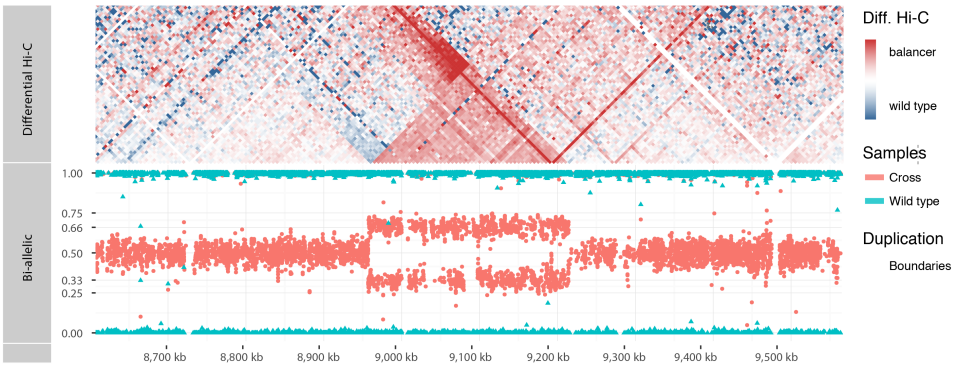


Figure 3.7: Large duplication on CyO characterized by differential Hi-C. Genomic regions around the duplication *chr2L:8,957,000-9,215,000*. The top panel displays differential unnormalized Hi-C fragment counts ($\log_2 \frac{\mathrm{bal}}{\mathrm{wt}}$) colored from red (positive) to blue (negative). The panel below shows the biallelic frequency. This duplication was initially discovered based on the characteristic biallelic frequency signal and is validated by a total increase of locus-internal Hi-C contacts on the balancer chromosome, as seen as the red triangle. Moreover, a decrease in balancer-specific contacts close to the diagonal left of the locus (blue) as well as an increase left of the upper tip of the triangular region suggest the duplication was inserted to the left in inverted orientation. Hi-C data processing and plots were contributed by Aleksander Jankowski.

to the same reference genome (see Section 3.6.1). In these maps, which are shown in Figure 3.8, we identified strong contact signals off the diagonal with a characteristic "bowtie" shape that coincided with gaps close to the diagonal. Undoubtedly these signals stemmed from genomic rearrangements and by comparing them to the available cytogenetic information, Yad Ghavi-Helm could assign 15/16 of these breakpoints to the respective inversions. The last breakpoint appears to lie outside the mappable reference genome and is thus inaccessible to read-mapping based methods. I then searched through inversion and translocation[2] predictions called with Delly (on WGS and mate pair sequencing data) to track down breakpoints at base pair resolution. This yielded base pair resolution breakpoints in 14/15 cases, which are reported in Table C.1.

Since our (yet unpublished) Hi-C- and MPS-based breakpoint refinement, two consecutive studies had mapped—for the first time—the genomic coordinates of several balancer rearrangements including CyO and TM3 [Miller et al., 2016; Miller et al., 2018]. Reassuringly, when I compared our results to the positions stated in these studies I found precise agreement for 12/12 breakpoints provided by the studies. However, relying on an approach using read pair and split-read

[2] These are not actual translocation, they just exhibit translocation signatures because the left and right arms of chromosomes 2 and 3 are reported separately in dm6

signatures in MPS data followed by local assembly, these studies failed to identify three breakpoints: for two of them (namely *2L:2.14 Mb* and *2L:12.71 Mb* on CyO) estimates at 10 kb resolution were provided, but they are off from our reported breakpoints by several kilobases. Another breakpoint (*3R:20.3 Mb* on TM3) remained fully obscure, whereas we at least provide a "best guess" based on Hi-C information. This strongly highlights the technological advantage in our study: Based on Hi-C data, we were able to track down the genomic positions of breakpoints down to 5 kb resolution. Then, both short- and long-insert paired sequencing data was searched for read pairs spanning the breakpoint junction.

We found, again in agreement with these previous studies, that only one of the breakpoint junctions in the balancer chromosomes showed a precise re-ligation, whereas most of them were scarred by a loss of genetic material (9 cases, max. 1.8 kb, median 114 bp) or small sequence duplications (4 cases, median 9 bp, max 281 bp; see Appendix C.6). Additionally, we predict that the breakpoint on TM3 at *3L:20.3 Mb* contains a deletion of around 17 kb, which could explain the difficulties in ascertainment by Miller et al. [2016].

Characterization of large SVs using Hi-C I also utilized the haplotype-resolved contact maps further to validate and characterize other large SVs. Notably, an inversion of approximately 40 kb on TM3 (*chr3L:14.61-14.64 Mb*). was validated by a characteristic Hi-C signal (data not shown). I also inspected aforementioned 258 kb duplication on CyO closer, which I had predicted by a characteristic BAF signal (described in Section 3.3.2). Interestingly, the differential Hi-C contact map (Figure 3.7) confirms the duplication by total increase in contacts on the balancer haplotype (seen as a red triangle). However, beyond the pure copy number the Hi-C signal further suggests that the additional copy of the locus was inserted in inverted orientation to its left end. There are two lines of evidence for this: A reduction in contacts of the locus to its left neighboring region (blue, i.e. decrease on the balancer chromosome), suuggesting that this end of the locus is no longer proximal to its original left neighboring region—as if an insertion had happened there. Secondly, there is a striking increase in contact frequency of the right end of the duplicated locus to the left neighboring region. Taken together this is best explained by an inverted duplication occurring in tandem.

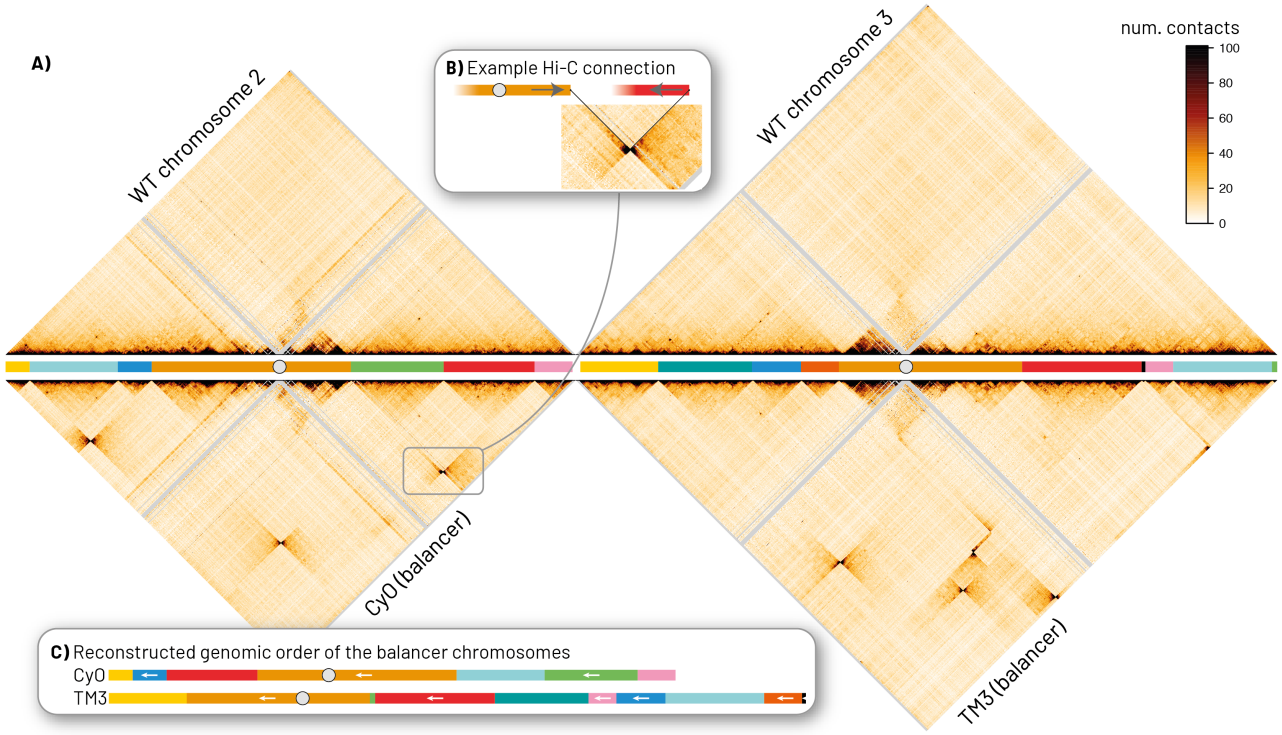


Figure 3.8: Major rearrangements of balancer chromosomes seen in Hi-C. **A:** Haplotype-resolved Hi-C contact frequency maps of chromosomes 2 and 3 in respect to the wild type reference genome are shown. Details of the data processing can be found in Appendix C.11. The bottom triangles show characteristic "bowtie-shaped" patterns as well as gaps on the diagonal that demarcate breakpoint junctions of the rearranged balancer chromosomes. The respective sections of the reference genome are color-coded, with a grey circle representing centromers. **B:** These junctions can be followed to reconstruct the relative order of the balancer chromosomes such as shown in the top panel. Vertical "bowties" represent a connection of segments in same orientation, horizontal ones segments that are inverted to one another. **C:** The fully reconstructed order of the balancer genomes is shown in the bottom panel and exactly matches previous annotation from karyotyping[1].

3.4. Results II: Global changes in gene expression

In the next step we sought to measure gene expression from both haplotypes to determine the potential impact of the observed genetic variation on the regulation of genes. To do so we first had to overcome biases related to maternally deposited mRNA present in early embryos, as described subsequently. Then I implemented robust ASE detection across multiple biological replicates (Section 3.4.2) and applied it to the data of our study as well as in the scope of two additional control experiments (Section 3.4.3).

3.4.1. Controlling for maternally deposited mRNA in early embryos

At an early stage of the project we sequenced mRNA from 4-8h embryos of the N_1 backcross (see Figure 3.3) to explore the power to detect significant ASE genes. As briefly introduced in Section 3.2.2, 25% of chromosomes 2 and 3 within the genetic pool consist of balancer chromosomes. This means that when we separate RNA-seq reads by haplotype we expect their distribution to be centered around a balancer fraction of 25%. As the left panels of Figure 3.9 display, this initial experiment was marked by a (unsatisfactory) wide distribution. More precisely, it did not resemble a binomial distribution, which is a model intrinsically applied by many ASE detection methods. Instead, it showed a large overrepresentation of wild type mRNA. Only after discussions with David Garfield we figured that, since we had crossed female wild type flies with double balancer males, the reason for this skew in the distribution could be attributed to maternally deposited mRNA present in early embryos as well as unfertilized eggs remaining in the sequenced pool.

maternally deposited mRNA *Before maternal-to-zygotic transition zygotes or early embryos do not transcribe their own genes but utilize mRNA of the maternal cell that was loaded into the egg during oogenesis. After around two hours of embryonic development a large portion of maternal mRNA is actively degraded, yet some maternal transcipt can remain [Tadros et al., 2009]*

To solve this issue, Yad Ghavi-Helm initiated a new experiment with modified conditions: First, she collected embryos only in a later, 6-8h time window when the impact of maternal mRNA is expected to be weaker. Secondly, she manually sorted out unfertilized eggs from the collection of embryos. And thirdly, she derived two different crosses, one with a paternal double balancer line (as initially), termed N_1^{pat}, and another one with a maternal double balancer line, N_1^{mat}. The potential presence of maternally deposited mRNA will then cause a shift in exactly opposite directions in both these lines, which can be accounted for during ASE calling. These considerations are nicely reflected in the histograms on the right of Figure 3.9. More narrow peaks indicate a much lower fraction of maternal transcripts, and a shift from left to right between N_1^{pat} to N_1^{mat} point to remaining maternal transcripts that can be identified as such.

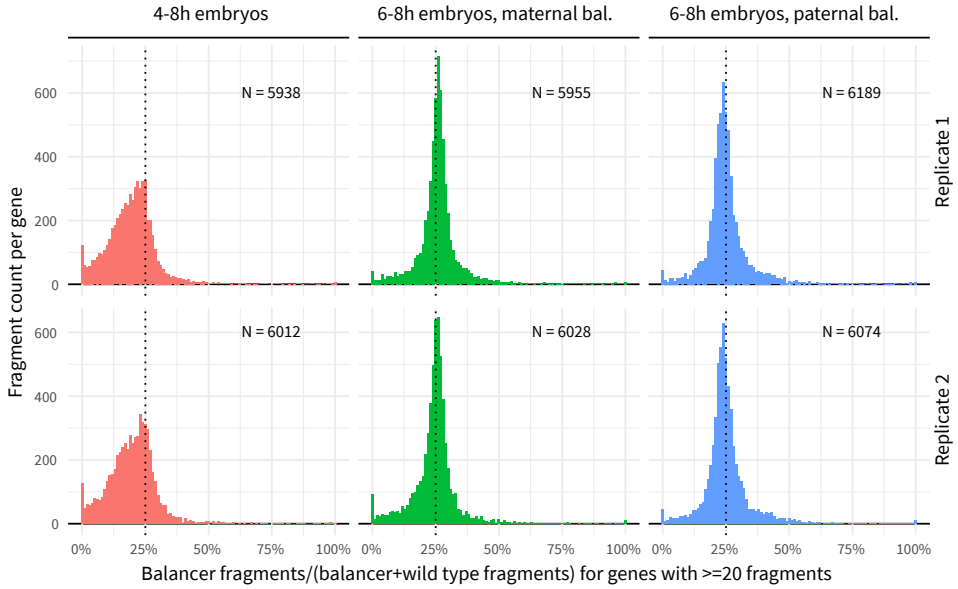


Figure 3.9: Allelic mRNA ratio per gene. Histograms of the fraction per gene of balancer RNA-seq fragments among the fragments that can be assigned to one of the haplotypes for three different RNA-seq experiments with two replicates each. The expected fraction of 25% is marked by a dotted line. Counts were derived as described in Section 3.4.2.

3.4.2. Allele-specific expression detection

The basic idea of measuring ASE is to separate the RNA-seq signal for a given gene into its two alleles and to test the resulting ratio against a null hypothesis of both alleles being represented equally. This is typically done by mapping RNA-seq reads to a common reference and inspecting sites of tagging variants (i.e. SNVs), although some approaches also rely on mapping to personalized genomes to overcome mapping biases [Rozowsky et al., 2014] and future extensions of this idea are pointing towards reference graphs [Dilthey et al., 2015; Marschall et al., 2016; Novak et al., 2017]. Accurate unbiased ASE detection requires many different considerations, as nicely elaborated by Castel et al. [2015], and there is an abundance of software available to perform the different steps involved [Skelly et al., 2011; Mayba et al., 2014; Harvey et al., 2014; Pirinen et al., 2015; Romanel et al., 2015; van de Geijn et al., 2015; Liu et al., 2016]. However, due the particular requirements of our study, they are mostly not applicable here: First, we needed to test ASE against an unusual fraction of 25%, whereas some models (reasonably) expect a 50% null hypothesis. Second, we would like to incorporate multiple rep-

reference graphs
Representation of different alleles or haplotypes simultaneously in a graph structure. A diploid genome, but also the haplotypes of a whole populations could be represented in such a data strucutre to avoid a reference bias. However, such approaches, which are being actively researched, require new methods for processing and interpretation of sequencing data

Data set	Total reads	Balancer	WT	Disaccordant
N_1^{pat} 1. replicate	51,560,368	6.151%	22.489%	0.136%
N_1^{pat} 2. replicate	52,618,060	5.762%	21.471%	0.139%
N_1^{mat} 1. replicate	54,521,194	8.030%	20.879%	0.142%
N_1^{mat} 2. replicate	48,179,068	7.834%	20.352%	0.153%

Table 3.3: Haplotype separation of RNA-seq data. Fraction of RNA-seq read pairs that could be assigned to either one of the haplotypes balancer or wild type (WT) ot that contained conflicting variants (disaccordant). The difference to 100% in each row comes from read pairs that did not overlap sites of tagging SNVs.

licates to obtain robust estimates, and third, we expected the different replicates to entail different ASE ratios for some genes (i.e. the ones with maternal mRNA), a situation that is unique to our study. I hence implemented an approach to ASE detection that captures these features.

As a first step I mapped RNA-seq read pairs to DM6 using STAR [Dobin et al., 2013]. I then separated read pairs into haplotypes using a simple Python script based on PYSAM[3]. This approach considers all SNVs that a read pair overlaps and can hence also detect if observed alleles within the fragment are disaccordant—a fact that we deemed especially important for the separation of Hi-C data later on (Section 3.6.1). This approach also differs from the SNV-centric approaches of other tools [Castel et al., 2015, for example] that yield allelic read counts at each variant. Furthermore, I preferred to derive a single sum of fragments for each allele of a whole gene, instead of considering the sites of variation within a gene separately. This comes with caveat that it might mask potential allele-specific alternative exon usage, which, for instance, was recently shown to occur at sites of CTCF binding [Ruiz-Velasco, Kumar, et al., 2017] and which can be identified using profound statistical tests [Skelly et al., 2011, for instance]. The advantage, though, is that our method does not require an additional model to integrate across sites of variation within a gene (such as presented by Mayba et al. [2014]) and can be expected to be more robust to local fluctuations in coverage than single-SNV analyses. On average we could separate 28% of RNA-seq read pairs (read lengths 2·144 bp) using this approach, as summarized in Table 3.3.

Then, each gene should be statistically tested against the null hypothesis. The commonly used Binomial test is known to inflate small p-values [Harvey et al., 2014] and further cannot integrate our replicate design. I thus used DESEQ2, a

[3]Found online at `https://github.com/pysam-developers/pysam`

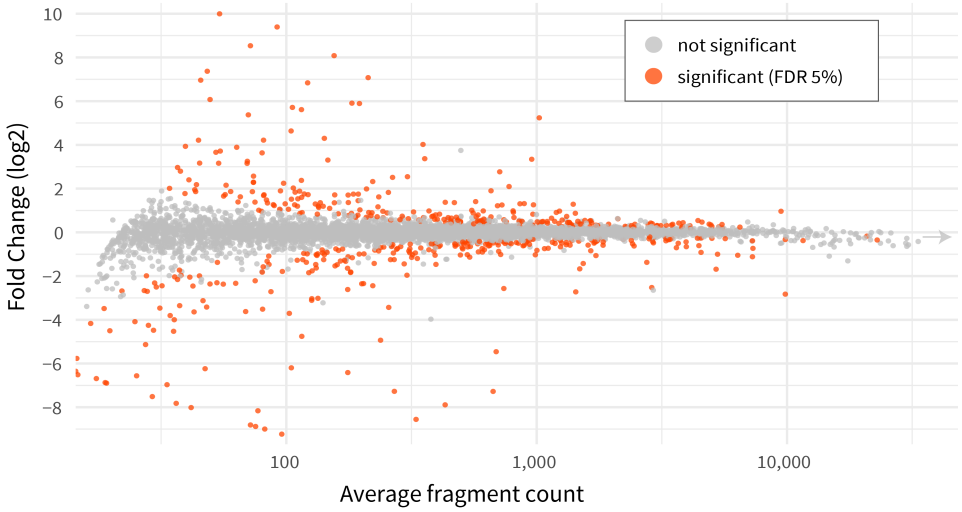


Figure 3.10: ASE analysis based on DESeq2. The x-axis shows the average fragment count per gene across four replicates (only haplotype-assigned read pairs go into this number). The y-axis shows $\log_2$ ratio of balancer over wild type counts. A single outlier to the right was trimmed.

well-established tool for differential RNA-seq analysis, to test for ASE. Given the haplotype-specific counts for each gene, DESeq2 is able to detect divergence from the 25% fraction across multiple replicates and can further cope with a potential shift in ASE ratio between the replicates caused by maternally deposited mRNA, which will be reflected in the p-values. It also performs multiple testing correction at a controlled FDR (more detail in Appendix C.7).

Among 5357 genes on chromosome 2 and 3 for which there was sufficient haplotype-resolved coverage we detected 512 significant ASE genes (9.6%) at a FDR of 5%. The median fold change was 1.8 with a fairly symmetrical distribution in both directions (see Figure 3.10). At a minimum fold change of at least 1.5, 343 genes remain (6.4%). These percentages are similar to the fraction of ASE genes in F_1 crosses of two distinct DGRP lines, where 5-10% ASE genes are reported (Furlong lab, unpublished work).

3.4.3. RNA-seq control experiments

We carefully considered some of the assumptions made in the ASE analysis and decided to carry out two important control experiments. A first question was whether balanced chromosomes contained a higher fraction of ASE genes than unbalanced chromosomes such as chromosome X. This could not be answered in

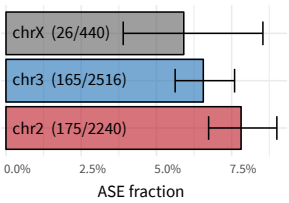


Figure 3.11: ASE fraction per chromosome. Balancer fraction of allele-specific RNA-seq fragments on chromosomes 2, 3, and X in a adult, female flies from an F_1^f line. Error bars indicate 99% confidence intervals from a binomial test.

our N_1 backcrosses, where chromosome X had eventually become homozygous due to the maternal wild type line. Instead, we performed RNA-seq on adult, handpicked female flies from the F_1 generation, named F_1^f, in two replicates. In this sample, chromosome X is paired with its (non-rearranged, but diverged) homologue at a ratio of 50%, too, and can thus be compared to chromosomes 2 and 3. The two chromosome X homologues contain less, but still sufficient distinguishing variation for ASE analysis (see Table 3.1). I executed ASE analysis as described in Section 3.4.2 on F_1^f RNA-seq data and found 301/4940 (6.1%) significant ASE genes on chromosomes 2 and 3 and 24/463 (5.2%) on chromosome X. The lower fractions compared to embryonic N_1 samples are of no concern, as gene expression is expected to be highly tissue- and developmental stage-dependent. Importantly though, the fraction of ASE genes on balancer chromosomes is not significantly higher than on chromosome X (p=0.47, Fisher's exact test), as shown in Figure 3.11.

Furthermore, we wanted to test whether the existence of multiple different genotypes in the N_1 generation could cause potential *trans* effects, e.g. whether genes on chromosome 3 would depend on the presence of CyO (i.e. balancer chromosome 2), and vice versa. In order to test this, Yad Ghavi-Helm generated three new F_1 generations of adult flies with different genotypes: one line (F_1^{CyO}) with only chromosome 2 balanced, another one (F_1^{TM3}) with only chromosome 3 balanced, and a last one with double balancer configuration (F_1). The latter differs from F_1^f by including both sexes and it was sequenced together with F_1^{CyO} and F_1^{TM3} to prevent batch effects. Note that in these single-balancer lines, haplotype separation is only possible on the balanced chromosomes themselves, for which SNVs had been previously mapped.

I then determined ASE genes on chromosome 2 using two replicates of F_1^{CyO} and two replicates of F_1 as described previously. Respectively, I did the same for chromosome 3 using F_1^{TM3} and F_1. Together I obtained 478/6356 (7.5%) significant ASE genes across both balancer chromosomes – a number that is fairly in accordance to the results of the previously sequenced F_1^f generation. Moreover, the balancer-to-wild type ratios of genes from these two samples are in high correlation (Pearson's r^2 = 0.824). Relevantly, I was then able to test the interaction term of ASE and genetic background using DESeq2: I asked, for instance, for which genes on chromosome 2 the ASE ratio significantly changed between F_1 (containing TM3) and F_1^{CyO} (not containing TM3). Again I executed this analysis separately on both chromosomes and found a significant interaction for 77/5981 genes (1.3%) at an FDR of 5% and a minimum fold change of 1.5. In order to put this number into context, I also estimated the differences in ASE ratio between

the two adult double balancer samples (i.e. F_1 vs. female-only F_1^f). These samples also exhibit high correlation (Pearson's r^2 = 0.848) and I detected 20/5382 (0.4%) genes for which the balancer/wild type ratio is significantly influenced by the genetic background (also at an FDR of 5% and a minimum fold change of 1.5). Requiring a fold change of at least two, these rates decrease to 39/5981 (0.65%) and 11/5382 (0.2%). These results are visualized in Figure C.4.

We concluded from this analysis that there is indeed a *trans* effect of the genetic background on the haplotype ratios of genes. However, this effect is small, maybe negligible, and even between the genetically seemingly identical lines such as F_1 and F_1^f significant changes are observed. We together reasoned that the analysis of a double balancer line (in contrast to a single balancer line) adds another variable to our study design that should be kept in mind, yet that this strategy is still preferable due to the sheer amount of chromosomal rearrangements that can be studied.

3.5. Results III: The interplay between SVs and differentially expressed genes

After characterizing SVs and detecting differentially expressed genes, we were interested in their relationship. In the upcoming section I describe how we explored their correlation and which SVs we believe are causal to ASE. Later on, Section 3.6 will specifically cover the role of chromatin conformation.

3.5.1. Genes affected by large rearrangements

During the characterization of the large chromosomal rearrangements I observed that they often happen to disrupt genes. In fact, 11 out of 15 breakpoints do so, affecting genes such as *Src42A* (Draper-Shark-mediated signalling immune response pathway), *GlyP* (carbohydrate metabolism pathway) and *p53* (tumour suppressor). The deletion present at breakpoint *chr3R:20.3 Mb* even spans a complete gene, which is consequently lost in the balancer chromosome. A full list of breakpoints and disrupted genes is given in Table C.1. The majority of these genes consequently show up in the list of ASE genes and are likely non-functional or at least severely truncated on the balancer chromosomes.

These frequent gene knockouts are perhaps surprising, as it is expected that selective forces during the creation of balancer chromosomes would rule out breakpoints disrupting genes. It can thus be assumed that these particular genes are

not dosage-dependent. Balancer chromosomes are indeed known to tolerate recessive mutation due to the reduced influence of natural selection [Araye et al., 2013]. Nevertheless, the frequency of gene knockouts without apparent phenotypic consequences should be acknowledged here.

3.5.2. Positional clustering of ASE genes

Following the enhancer adoption hypothesis, we reasoned that ASE genes would be preferably located around the rearrangement junctions on the balancer chromosomes. To assess this, I specifically tested for an enrichment of ASE genes around breakpoints in contrast to around randomly chosen genomic positions (Figure C.5). Surprisingly, only a single breakpoint, namely *chr3R:20.31 Mb*, appeared to be surrounded by more ASE genes than expected by random chance: 5 out of 6 neighbouring genes, which span a region of around $\pm$ 50 kb, were significant. By means of integrated visualization (Section 3.7) of this locus I noted that two genes were highly up-regulated left of the breakpoint, another gene down-regulated right of the breakpoint, and an additional gene disrupted by the breakpoint itself. Notably, we considered this breakpoint as the best candidate for a potential chromatin structure-related effect on gene regulation. However, closer inspection revealed that the ASE signal of these genes was likely caused by chimeric transcription across the breakpoint junction. We observed such chimeric, likely non-functional expression happening on an appreciable set of other genes, too, as I elaborate in Section 3.5.4.

I then turned to a more unbiased approach and searched the set of ASE genes for positional clustering anywhere in the genome. As Figure 3.12 demonstrates, ASE genes turned out to be located as far away from another as random control genes, with the exception of a small set of genes in a distance of approximately 3 kb to one another. These, as the integrated visualization again quickly revealed, mostly belong to the large duplication on CyO (see Section 3.3.2 and Figure 3.6) and not to one of the chromosomal rearrangements. We thus concluded that there is no other region in the genomes with an enrichment in significantly misregulated genes; instead allele-specific expression is similarly distributed as other expressed genes. In addition, these analyses (e.g. the CyO duplication) hinted at mechanisms unrelated to chromatin conformation that could create ASE signal, which are further explored subsequently.

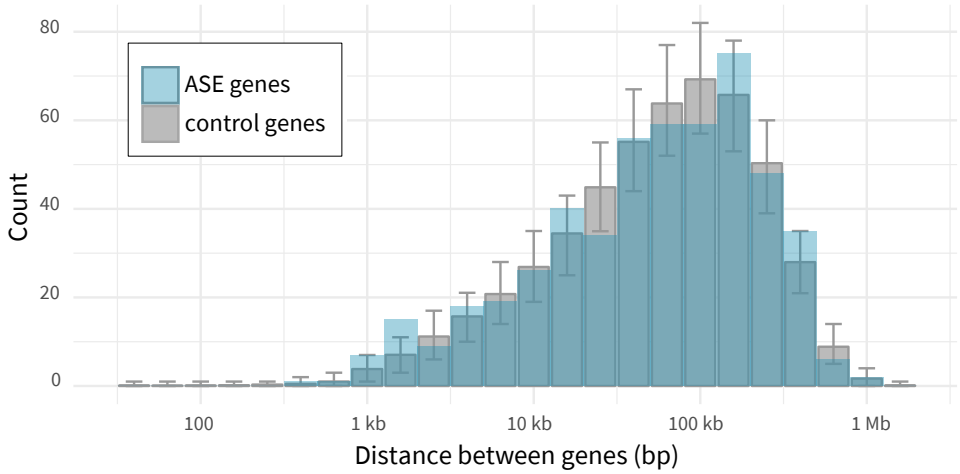


Figure 3.12: Distance between neighboring ASE genes. This is a histogram of distances between neighboring significant ASE genes (blue) as well as between a random set of control genes (grey), which were sampled from expressed, but non-ASE genes 500 times. Errorbars indicate the 5% and 95% quantiles of random sub-sampling.

3.5.3. ASE signal related to changes in copy number

Next, I tried to understand how much of the ASE signal is caused by CNVs. I therefor inspected significant ASE genes that have at least one of their exons (incl. $3'$ or $5'$ untranslated regions) affected by a CNV. This overlap yielded 73 genes, which is a significantly higher fraction among the ASE genes (14%) than among expressed, but non-ASE genes (292 cases, 6%; p-value $< 10^{-}9$, Fisher's exact test). Figure 3.13 shows the effect CNVs have on the expression of ASE genes. As expected, a clear trend towards a dosage effect can be seen, i.e. that a deletion within a balancer gene decreases and a duplication increases the balancer expression (seen by a positive log fold change), and vice versa for wild-type-specific CNVs. Duplications, which are typically much larger in size than deletions (Figures 3.4 and 3.5), show a clearer impact on transcript levels, as they often duplicate whole genes. Yet it cannot be expected that a higher copy number neccessarily leads to an increase in transcribed RNA. For example, a partial duplication could disrupt a gene in a way that its expression is decreases. Accordingly, the impact of deletions is less, as they often affect only single exons. The specific explanation for each gene could, if need to, be unravelled in case-by-case fashion.

We derive as a conclusion from this analysis that up to 14% of significant ASE genes could be explained directly by CNVs affecting their exons—ergo, chromatin conformation-related are unlikely to play a causal role in the altered expression

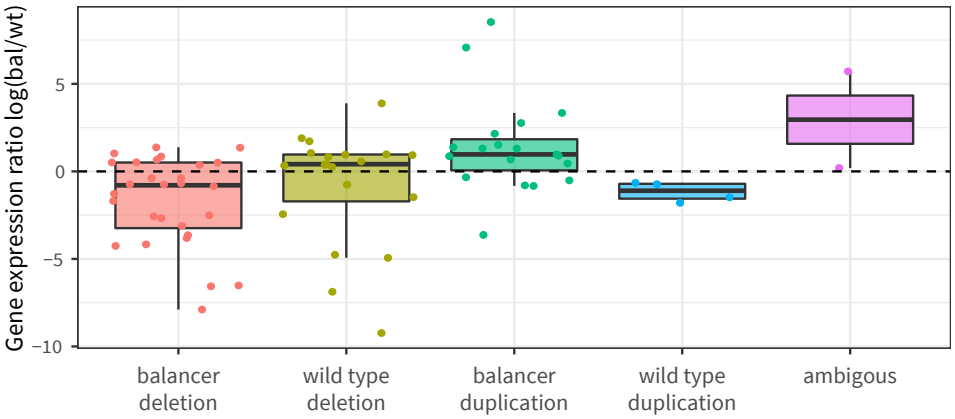


Figure 3.13: Log fold change of ASE genes overlapping CNVs. Expression level log fold change of 73 significant ASE genes that overlap CNVs. Two genes overlap multiple different CNVs (*ambiguous*). A positive log fold change means higher expression in the balancer haplotype and vice versa.

of these genes.

3.5.4. Mobile element insertions can give rise to strong ASE signals

In a thorough manual data inspection, Yad Ghavi-Helm found a sizeable number of genes with an RNA-seq signal that did not start at their expected transcription start site. We further found that this expression was often only present on one of the two haplotypes, and that the RNA-seq coverage started at a seemingly random position. Interestingly, despite starting in an intronic or even intergenic region, this signal would often pick up a characteristic splicing pattern (high signal within exons, no or low signal along introns) from the end of the first traversed exon. Figure 3.14 depicts an example gene with strong balancer-specific gene expression.

We speculated that this might be driven by the insertion of mobile elements in the positions where the RNA-seq signal started. Hence I developed a computational pipeline that extracts DNA-sequencing reads of a given genomic region, computes an assembly of these reads using SAMtools and SPAdes [Bankevich et al., 2012], and compares the assembled contigs to a database of common transposable elements in *Drosophila*. The comparison is done via read mapping and confirmed visually using the dotplot functionality of Maze developed during previous work (Section 2.4). Using this approach, I found transposable element insertions in 42 loci. Based on the position and orientation of the mobile element (i.e.

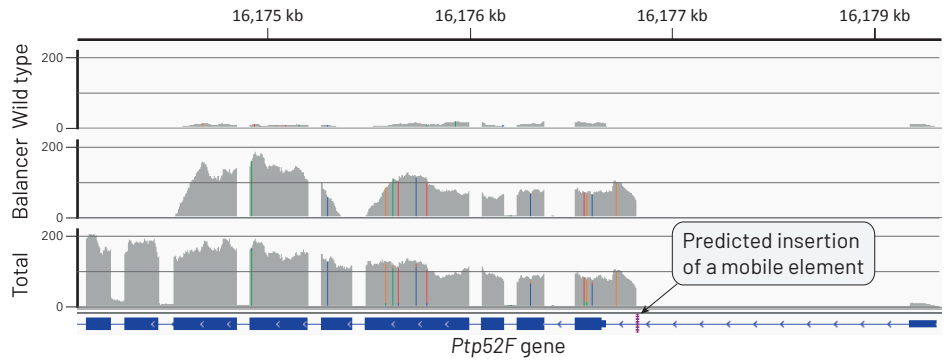


Figure 3.14: Example of a MEI driving ectopic expression of a gene. RNA-seq tracks along the gene *Ptp52F* on chromosome 2R showing the total RNA-seq reads (total read coverage in absolute numbers; bottom panel) as well as the portion of RNA-seq reads that could be resolved to the wild type (top panel) and balancer (middle panel) haplotypes. Colored vertical lines represent SNVs used for haplotype-separation. RNA-seq tracks were visualized using IGB.

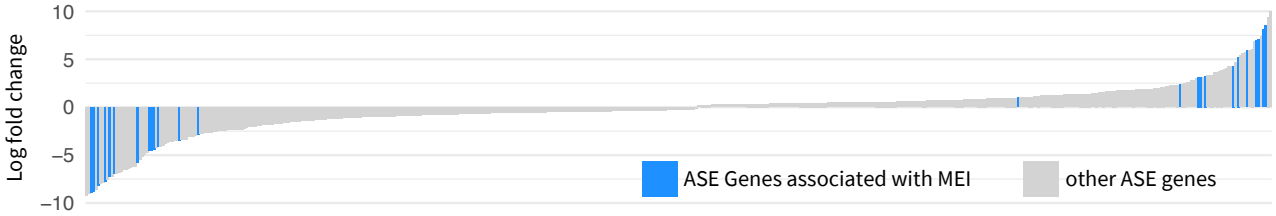


Figure 3.15: ASE genes associated to MEI. List of significant ASE genes (x axis) ordered by log fold change (balancer/wild type, y axis). Genes that are very likely dis-regulated due to an MEI driving their chimeric expression are highlighted in blue.

mobile elements contain their own transcription start site) I evaluated whether the aberrant ASE signal of the gene could be caused by chimeric transcription from the mobile element's promoter. This is the case in 37 loci (Table C.2), of which 25 were expressed high enough to be tested for ASE.

Astonishingly, all 25 genes were also called to show significant ASE, partially with extreme fold changes as shown in Figure 3.15. In the aftermath, the high fold changes can likely be explained by the activation of silenced (or very weakly expressed) genes. The promoter intrinsic the the mobile element then allows expression only from the affected homologue, which can be orders of magnitudes larger than the expression from second copy. MEIs could in principle also cause chimeric expression of higher expressed genes, the additional mRNA is probably just not detecteble. At last, it should be kept in mind that this analysis centered

on differentially expressed genes with aberrant transcription start sites, hence the full impact of MEI is still under-explored.

3.6. Results IV: Changes in chromatin conformation

Here, I summarize our findings in respect to the chromatin conformation of balancer and wild type chromosomes. Most analyses described below were performed by Aleksander Jankowski unless clearly marked by "I". These investigations were ongoing at the time of writing and might hence be subject to change prior to final publication.

3.6.1. Differences in chromatin conformation between wild type and balancer chromosomes

To study chromatin organization, we performed Hi-C on nuclei from 4-8 h embryos of the N_1 generation in two biological replicates. We sequenced the Hi-C library in as many as 24 lanes on an Illumina machine to obtain high resolution contact frequency maps. We then annotated read pairs by haplotype in the same way RNA-seq data was annotated beforehand (Section 3.4.2). Moreover, a strict filter procedure was designed following recommendations from Ramírez et al. [2018], which is described in the appendix (see Appendix C.11 and Table C.3 for details on the effect of different filters). Finally, 117 million pairwise contacts remained on the wild type chromosomes and 35 million on the balancer chromosomes to build Hi-C maps in 5 kb resolution. It is interesting to note that more Hi-C read pairs could be separated into haplotypes relative to RNA-seq read pairs, despite shorter read lengths (38% vs. 28%, approximately). This is a consequence of the large "insert sizes" of Hi-C data, which make it more likely for a pair to bridge genomic regions with low SNV density [Edge et al., 2017].

At first we visually compared the contact maps of matching genomic regions of balancer and wild type chromosomes using, among others, the visualization tools described in Section 3.7. Except for in proximity of the breakpoints of large rearrangements, at which contact frequencies of the two differently arranged chromosomes cannot simply be compared, we observed a striking similarity between the balancer and wild type chromosomes. As it appears, the rearrangement on the balancer chromosomes affects the three-dimensional chromatin organization only locally—anywhere else, the three-dimensional organization remains largely unaffected. In order to quantify this, we calculated the TAD separation score as a one-dimensional feature of the local compactness [Ramírez et al., 2015; Ramírez

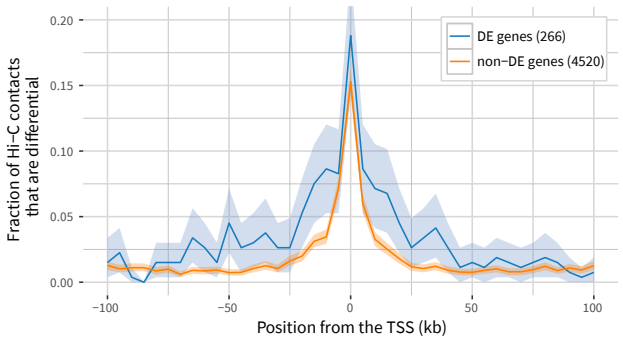


Figure 3.16: Differential Hi-C contact from gene promoters. This shows differential Hi-C contacts between gene promoters (at 0 kb; all genes are aligned and oriented with the gene and downstream sequence to the right) and all other 5 kb bins within ±100 kb. On the y-axis, the fraction of significantly differential Hi-C contacts among all Hi-C contacts is shown. The signal is averaged across ASE and non-ASE genes and plotted here including 95% confidence interavls.

et al., 2018]. In fact we calculated score profiles for seven window sizes[4] and averaged them at each genomic position. We found a high correlation ($r^2 = 0.93$) between balancer and wild type haplotypes, confirming our visual impression.

Despite the similarity on the large scale, single pairwise contacts can in fact differ between the haplotypes. We tested for significant differential pairwise contacts using DESeq2 by providing haplotype-resolved fragment counts in 5 kb bins utilizing both Hi-C replicates. Among 509,217 tested pairwise connections that had sufficient read support, we found a total of 48,206 significant differences. Despite attempts to normalize for distances in the altered genomic order, 1,388 (2.9%) of the detected significant differences spanned across a breakpoint of the large chromosomal rearrangements—these pairs were excluded from the further analysis.

We were then interested in whether changes in chromatin architecture are associated with differential gene expression. To study this relationship, we zoomed in on the differential Hi-C contacts at gene promoters (i.e. at the 5 kb bin surrounding the promoter) and averaged across many genes. Figure 3.16 shows the significant contacts from the promoter (center) to all 5 kb-bins in the surrounding region. Genes are aligned and oriented according to their transcription start site and the average fraction of differential contacts among all contacts are shown on the y-axis. Interestingly, we found that, on average, there were more differential Hi-C contacts reaching to the promoters of ASE genes than to promoters of

[4] 50 kb, 60 kb, 80 kb, 100 kb, 130 kb, 160 kb, and 195 kb

non-ASE genes. This difference is pronounced within 50 kb around the gene and notably vanishes at a distance of 100 kb.

This data suggests that chromatin conformation might indeed play a role in regulating gene expression. However, it is unclear whether the conformational changes are cause or consequence of altered gene expression and how much this association varies between single loci.

3.6.2. TAD structure around breakpoints

One of the crucial questions of the study was how chromatin conformation, especially TAD structure, would change around breakpoints of large chromosomal rearrangements. Based on Hi-C data, TADs can be predicted using one of several available algorithms – however, given the differences in the predictions from these methods, this is still a major challenge in the field [Forcato et al., 2017, notably figure 3]. We utilized TAD calling based on the TAD separation score [Ramírez et al., 2018] to predict TADs only on the contact map of the wild type chromosomes. This resulted in 880 domain calls with a median size of 125 kb, which is a slightly lower segmentation than reported by Sexton et al. [2012] (1,170 domains with a median size of 62 kb).

We then inspected the TADs around the 15 breakpoints of large chromosomal rearrangements. Most TAD callers, including the method we applied here, report TAD boundaries as infinitesimal points instead of larger intervals, which we think does not capture the reality well. In order to distinguish intra-TAD space from boundaries, I thus required a minimum distance of 20% (of TAD size) to both ends of the interval in order to consider a TAD disruption. According to this computational analysis, 12/15 breakpoints fall into TADs rather than into boundaries (Table 3.4). Since TAD calls did not always match our visual impression, Table 3.4 also includes a manual annotation, in which we classified 10/15 breakpoints as breaking TADs.

Aforementioned computational criteria can also be applied to TAD calls from balancer Hi-C data in respect to the balancer genomic order, in order to evaluate formation of new TADs around the junction points. According to computational analysis (again using the 20%-distance criterion), 8/15 of these junction points appear to reside within newly formed TADs. A manual inspection also yielded 8/15 cases of TAD formation, yet in two examples my classification differed from the computational analysis.

Despite inaccuracies in computationally defining TADs, the numbers reported here make clear that breakpoints in fact disrupt TAD structure of the wild type

Breakpoint coordinates			TAD	Distance	Evaluation
chr2L	2,137,067	2,137,075	180 kb	22,925	+
chr2L	12,704,649	12,704,657	335 kb	~~14,649~~	−
chr2L	9,805,567	9,805,575	200 kb	94,425	++
chr2R	14,067,771	14,067,782	215 kb	107,218	++
chr2R	6,012,459	6,012,739	40 kb	7,261	−−
chr2R	21,971,918	21,972,072	205 kb	101,918	+
chr3L	6,925,034	6,926,125	230 kb	53,875	++
chr3R	9,943,831	9,944,040	115 kb	40,960	+
chr3L	15,150,269	15,150,272	120 kb	~~5,269~~	−−
chr3R	23,050,763	23,050,764	380 kb	79,236	+
chr3L	19,386,273	19,388,151	310 kb	126,273	+
chr3R	20,637,930	20,637,930	275 kb	82,930	−
chr3L	22,637,876	22,637,952	310 kb	82,048	+
chr3R	31,653,695	31,653,707	135 kb	43,695	++
chr3R	20,308,200	20,325,700	260 kb	~~13,200~~	−−

Table 3.4: Overview of TAD calls at breakpoint positions. TAD annotation at breakpoints of large chromosomal rearrangements. TADs were called only based on wild type Hi-C data using the approach of Ramírez et al. [2018]. Column *TAD* contains the size of the overlapping TAD call and *Distance* the distance of breakpoints to the TAD boundaries. Only in three cases (marked by ~~Distance~~) breakpoints were closer to TAD boundaries by more than 20% of TAD size—these cases supposedly do not interrupt TAD structure. The *Evaluation* column contains a manual assessment of TADs based on inspection of Hi-C maps. + means that a TAD is likely interrupted, ++ it is clearly interrupted, and − (−−) that breakpoints likely (clearly) fall into TAD boundaries.

chromosomes. Moreover, in the new genomic order of the balancer chromosome, new TADs were formed across the junctions in many cases. This is for example the case at locus *chr2R:14.0 Mb*, which is visualized in Figure 3.17 (and explained in more detail subsequently). Domains that were not affected by breakpoints, on the other hand, did not seem to change at all (Section 3.6.1). Interestingly, when we inspected newly formed TADs spanning breakpoint junctions, we observed that they often expanded up to neighboring domains, effectively re-using previous TAD boundaries. This has not been formally tested, though. However, as already noted earlier (Section 3.5.2) TAD disruptions or formations appear not to have a notable effect on gene expression. ASE genes are not enriched around breakpoints, but rather spread evenly across the genome, and no evidence for the presence of an enhancer hijacking-like mechanism could be found in this study.

3.7. Integrated visualization of genomic loci

At last, this section covers a methodological aspect that was essential in deriving many of the aforementioned results, especially of section Section 3.6.2.

With a variety of data available in our study including RNA-seq and Hi-C, we faced the challenge of how to explore these datasets around loci of interest. Genome Browsers, which were specifically designed for this task, can show multiple tracks of different data in respect to the same genomic coordinates [Freese et al., 2016; Thorvaldsdottir et al., 2013; Gramates et al., 2017]. In our particular case existing solutions were not satisfying, though. First of all, we would like to include Hi-C maps into our figures. Second, we needed all data to be represented in the two different genome assemblies resulting from balancer and wild type chromosomes. Only recently some tools (published or at pre-print stage) addressed the first point [Ramírez et al., 2018; Kerpedjiev et al., 2017], yet not allowing the flexibility we required. This is why Aleksander Jankowski and I utilized the R-package ggBio [Yin et al., 2012] to create our own tailored visualization.

Figure 3.17 shows an excerpt of one of the plots we generated. The top panel of the figure highlights (a) the position of the genomic locus within the respective chromosomal assembly (here the wild type reference genome dm6) with colors that roughly correspond to the initial illustration of balancer chromosomes (Figure 3.8). Below, we incorporated different Hi-C maps. In Figure 3.17, there is one Hi-C map created only on (b) wild type data and one only on (c) balancer Hi-C data. Around the breakpoint of one of the large rearrangements, for example at *chr2R:14.07 Mb* as shown here, these haplotype-resolved Hi-C maps quite remarkably display the missing connections in the balancer chromosome. If these signals were plotted in respect to the genomic order of the balancer chromosome, this gap would be visible in the wild type track respectively (see Figures C.6 and C.7 for exactly this). Optionally, the plot can include additional Hi-C maps, for example generated from the complete data or showing differential contacts, as in Figure 3.7.

The middle part of the figure may contain arbitrary one-dimensional genomic features. The example figure includes (d) deletion calls on balancer and wild type haplotype, (e) DNase-I hypersensitivity tracks (Furlong lab, unpublished data), and strand-specific RNA-seq tracks including (f) all RNA-seq reads, only (g) wild type-specific RNA-seq reads and only (h) balancer-specific RNA-seq reads (positive: coverage on plus strand; negative: minus strand). Other signals that could be displayed are SNV positions, other SV classes, insulation scores, or predicted TAD boundaries.

The bottom part contains (i) gene annotations and maps the genomic coordinates of genes to a table listing additional information about these genes, including their names. Shown here are the (k) total expression level, the (j) log fold change of RNA levels in balancer compared to wild type haplotypes, and whether this change is significant according to our ASE analysis (orange color for significance). Such a breakpoint-centered plot is accompanied by two additional plots that represent the same locus in the genomic order of the balancer chromosomes, which can be found in the appendix (Figures C.6 and C.7).

We generated these plots for all loci of interest, notably for the breakpoints of the large balacer rearrangements. These plots were highly useful to confirm statistical analyses and to gain insights on the mechanisms acting in a locus. We are currently considering to make this software available to the public.

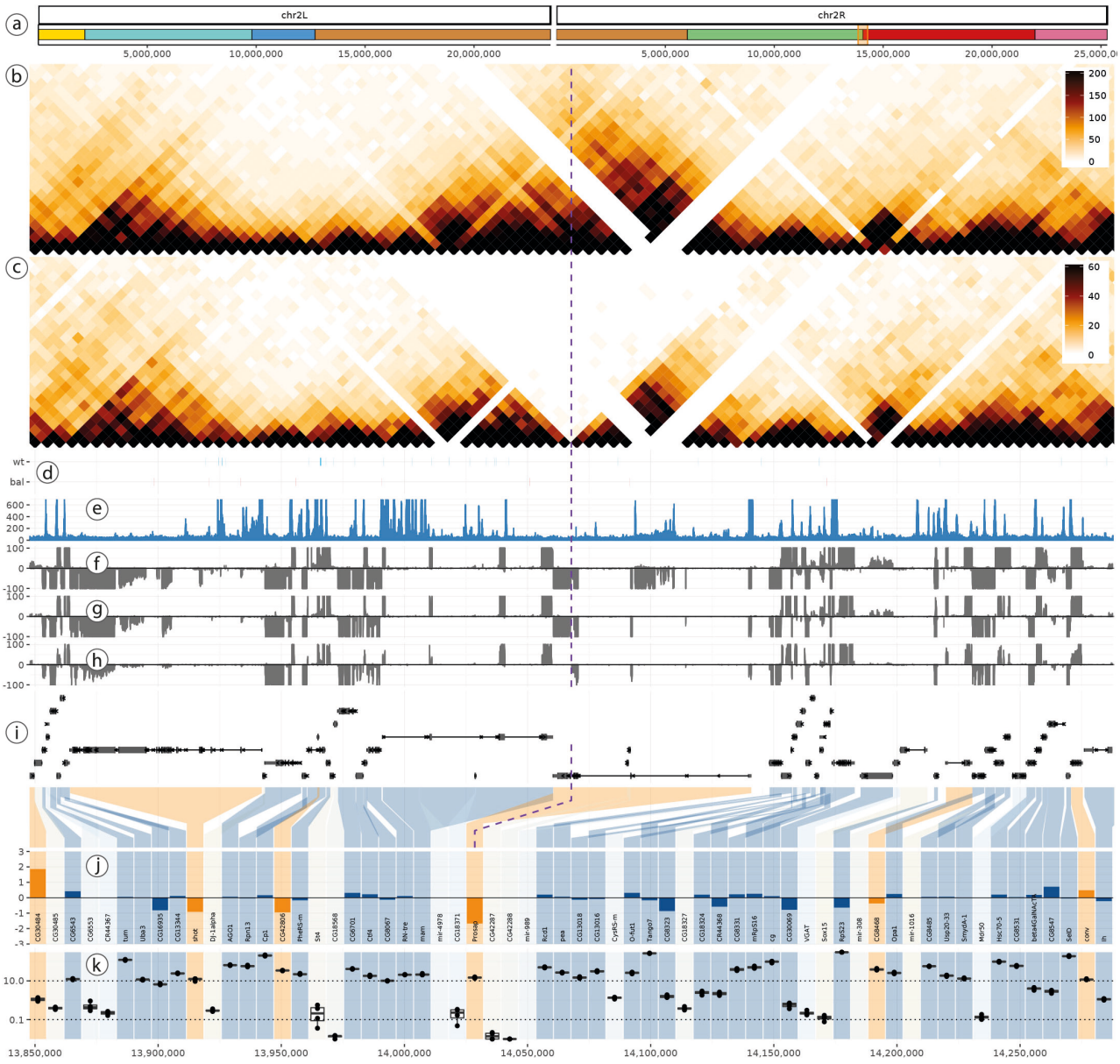


Figure 3.17: Integrated visualization around breakpoint *2R:14.1 Mb*. Visualized through our custom plot functions based on GGBIO. The figure is explained in the main text.

3.8. Conclusions

In this study, we aimed at testing the influence of SVs on chromatin organization and gene expression. Particularly, we wanted to assess whether an enhancer hijacking mechanism, which affects gene regulation by restructuring chromatin in 3D, occurs in a phenotypically healthy organism and how frequent it might be.

Based on the study design and experiments of Yad Ghavi-Helm, I deeply characterized balancer chromosomes of *Drosophila melanogaster* in terms of variation, gene expression and chromatin conformation. I unraveled the exact positions of the cytogenetically annotated rearrangements, which were partly found concurrently by two studies [Miller et al., 2016; Miller et al., 2018] and which are in perfect agreement to the results stated therein. I further characterized the balancer and wild type chromosomes in respect to SNVs, small indels, deletions and duplications, the latter two of which were validated experimentally or computationally using independent signatures. Based on haplotype-tagging SNVs, I developed a methodology to detect allele-specific expression across four biological replicates and including a correction for maternally deposited mRNA. I find that genes with significant differential expression between both haplotypes are distributed rather equally across the genome and not enriched at breakpoints.

Utilizing Hi-C data, Aleksander Jankowski found that global chromatin structure, notably the patterning into topologically associating domains, is generally not different between wild type and balancer chromosomes, despite their tremendously different genomic order. This supports the current notion in the field that TADs are established by their boundaries and are hence a local feature of chromatin. Nevertheless, exactly *at* the breakpoints, local chromatin structure necessarily changes, as I exemplified for a locus in Figure 3.17.

Based on computational TAD predictions (Aleksander Jankowski) and manual inspection, I found that in 10-12 out of 15 cases TADs of the wild type chromosome were disrupted and in approximately 8 cases, new TADs were formed within the balancer chromosomes. Since only few ASE genes are found within these TADs and we generally could not detect any enrichment for ASE genes close to breakpoints, we do not think that these changes in chromatin structure affect gene expression in our model. This is in stark contrast to previous studies (Section 3.1.3), where typically a single TAD disruption is reported to ectopically express genes in non-physiological amounts.

A possible explanation for this apparent discrepancy is natural selection. Just as in our study TAD disruptions had undergone selection to guarantee viability of the organism, the alterations in chromatin structure analyzed in many of these

aforementioned studies had been selected for a pathological phenotype such as cancer or Mendelian diseases.

Selective pressure might indeed have prevented misregulation via enhancer hijacking-like mechanisms. However, at the same time a great amount of variability with consequences on gene expression was apparently tolerated: For instance, copy number variants, one of which duplicates 33 genes at once, are abundant and believed to alter expression of 73 genes significantly. Rearrangement breakpoints directly disrupt genes in 11/15 cases, including *p53*. And while these changes typically alter expression ratios in the range of not more than two- to three-fold, there is a relevant number of mobile element insertions that drive expression of at least 25 genes. This typically leads to drastic ASE signals in the order of dozens to hundred times fold change. However, the resulting chimeric transcripts lack exons or contain intergenic sequence, so it is unclear whether they are functional.

Eventually, hundreds of genes remain that are significantly differentially expressed but for which we have no explanation at hand. By correlating these genes with significant differences in three-dimensional chromatin interactions, we found that indeed chromatin structure appears to be preferably altered around ASE genes (Figure 3.16) and might consequently play a role in mediating transcriptional regulation. Recently, Yanjian Li et al. [2018] described a clear correlation between changes in intra-TAD density and up- or down-regulation of gene expression during differentiation of leukemia cell lines. It was suggested that both arise simultaneously as consequence of an altered epigenetic state of the TAD, as for example Le Dily et al. [2014] observed.

Their observation is different from enhancer hijacking, but epigenetic mechanisms could principally also play a role in suppressing enhancer hijacking. To give an example, it is easily conceivable that among the eight described examples of neo-TAD formation at least one active enhancer elements was juxtaopsed to potential target genes. Ectopic expression could then be suppressed through histone modifications, for example. Alternatively, these potential position effects could also be buffered by other genetic variants, such as one of the 760 thousand SNVs.

In conclusion, this study finds that genomic rearrangements can vastly alter chromatin architecture, but that this does not necessarily translates to functional consequences. We do not observe any signs of enhancer hijacking when TADs are broken, which could be a consequence of selective pressure and which demonstrates the extraordinary robustness of biological systems. Subtle changes in gene expression, to which our approach is sensitive to, were not found to be enriched around disrupted TADs either, suggesting that enhancer hijacking-related mech-

anisms operate on a all-or-nothing basis. Additional biological mechanism such as epigenetics might play a role in buffering the effects of rearrangements, yet analyzing this was beyond the scope of this study. Future research will be necessary to fully understand the impact of chromatin architecture on gene regulation.

Structrual Variant Detection in Single Cells

4

Here, I present the current state of a project with the objective to develop a computational method for SV detection based on the Strand-seq technology. I have been developing this method together with Jan Korbel and Ashley Sanders as well as with our collaborators David Porubský, Maryam Ghareghani, and Tobias Marschall from the Max-Planck Institute for Informatics, Saarbrücken. Ashley Sanders and Jan Korbel conceived the general concept of SV discovery using the signals provided by Strand-seq data. Tobias Marschall and David Porubský contributed analyses related to phasing, general improvements to the overall work flow and fruitful discussion on many aspects. Maryam Ghareghani developed the Bayesian classification approach described below. Also work from Venla Kinanen on sister chromatid exchange events was included into this method. The main implementation and all other analyses, including figures, are my own. At last, I would like to thank Balca Mardin, who provided the cell lines utilized for demonstration purposes, and Peter Lansdorp's lab, who carried out the Strand-seq experiments on these cell lines.

4.1. Introduction: Structural variants in the context of somatic mosaicism

As briefly introduced earlier, somatic mosaicism refers to the presence of genetically distinct subpopulations of cells within an individual [Youssoufian et al., 2002]. The relatively high susceptibility of dividing cells to mutagenesis suggests that nearly all of the cells in our body harbor variation, ranging from SNVs to large SVs [I. M. Campbell et al., 2015]. Fortunately, the majority of these accumulated variants appear not to affect our health. Certain cases of somatic mosaicism are even part of our physiological development, such as the programmed rearrangements of T-cell receptor genes, or the tendency of liver cells to develop

polyploidy [Forsberg et al., 2017; Davoli et al., 2011].

However, mosaicism has also been linked to disease and it is becoming increasingly clear that structural variants play a major role in this context [Forsberg et al., 2017]. The prime example is cancer, which arises from the clonal expansion of a cell that carries beneficial (driver) mutations as well as passenger mutations. But other diseases, such as type 2 diabetes mellitus or Alzheimer, have been linked to the presence of mosaic mutation, too [Forsberg et al., 2017]. Moreover, the amount of mosaicism has been shown to increase with age. For example, Forsberg et al. [2012] observed Megabase-range aberrations in people at the age of 60 and above, but not in younger subjects. Especially the blood compartment, but also skin, appear to be affected by somatic mutation—as, for instance, was shown in the blood of a 115-year old woman, which harbored a manifold-higher amount of mosaic variants than other tissues [Holstege et al., 2014; Forsberg et al., 2017].

In order to study somatic mosaicism within a seemingly homogeneous tissue, three major approaches had been tested in the past: the deep analyses of bulk samples, clonal expansion of single cells, and single-cell genomics techniques. In bulk analysis, variants of low frequency can be difficult to detect. Spencer et al. [2014] estimated that, in order to detect SNVs variants with a frequency of 2.5% within the cell population, a coverage of at least 500 x would be required. Interestingly, the detection of aneuploidy (or CNVs of chromosomal scale) with similar frequencies can be achieved with much lower coverage by collecting evidence across many SNV sites. This was demonstrated in cancer data using microarrays and MPS data [Van Loo et al., 2010; Carter et al., 2012; Roller et al., 2016]. However, SV breakpoints are difficult to assess in this way, as typical paired-end read calling methods require a coverage of at least 10–20 x to detect breakpoints.

Clonal expansion of single cells is a way to achieve higher coverage. These techniques are based on the the culturing of single cells in the laboratory, which are then let grown to obtain populations of genotypically identical (clonal) cells. Using such techniques, the heterogeneity within fibroblasts and fibroblast-derived induced pluripotent stem cells, among other things, could be studied [Saini et al., 2016; Abyzov et al., 2012]. It is a laborious approach, though, as it requires the culturing and subsequent sequencing of each clone. Instead, techniques that directly measure single cells have gained more and more attention over the last years.

Single-cell studies have uncovered vast amounts of somatic copy number alterations in cancer [Navin et al., 2011; Demeulemeester et al., 2016]. They also revealed unforeseen amounts of somatic mosaicism in brain tissue, including in post-mitotic neurons, which were analyzed on the level of SNVs, CNVs, and MEIs [Lodato et al., 2015; Cai et al., 2014; Evrony et al., 2012]. These single-cell meth-

ods typically rely on whole-genome amplification to increase the amount of DNA available for sequencing. However, the amplification step has been noted to introduce biases that hamper detection of CNVs below several Megabases in size [Deleye et al., 2017]. While single-cell CNV detection has been constantly improving since, allowing much higher resolution [Garvin et al., 2015; Gao et al., 2016; Bakker et al., 2016; Knouse et al., 2016], the detection of copy-neutral SVs has been lagging behind.

whole-genome amplification *Method to amplify the entire DNA in a nucleus prior to single-cell sequencing. Multiple technologies were put forward for this task, such as MDA or MALBAC [Dean et al., 2002; Zong et al., 2012], witch have been compared recently for their usability in CNV detection [Deleye et al., 2017]*

These limitations made it difficult to robustly discern structural variation, especially copy number-neutral and complex events such as inter-chromosomal rearrangments or inversions, in the context of cellular heterogeneity. Consequently, the role of somatic SVs remains comparatively underexplored to other forms of genetic variation. In this chapter, I present a novel principle to study SVs on the single-cell level. Based on the single-cell sequencing method Strand-seq, we developed a computational approach for the detection of varous SV classes, including ones that were previously difficult to ascertain. Below, I explain this method in detail and give an outlook on the future research that can and will be done using this promising novel approach.

4.2. Results I: A novel method for single-cell SV detection

Strand-seq generates sequencing reads that are all of the same directionality when they stem from the same homologue. As explained in Section 1.3.4, this is achieved by labeling and degrading the non-template strand of actively replicating cells [Falconer et al., 2012; Sanders et al., 2017]. When the resulting sequencing reads are mapped to a reference assembly, they map either to the Watson (W) or Crick (C) strand. The presence of two homologous chromosomes then leads to the the observation of WW, WC, or CC chromosomes. I call those (WW, WC, CC) their *inherited strand states*. Here, I show how these unique characteristics of Strand-seq reveal the presence of seven different SV classes and demonstrate our computational method called MOSAICATCHER that realizes this idea.

4.2.1. Single-cell Strand-seq libraries

Figure 4.1 depicts Strand-seq data from a single cell of a retinal pigmented epithelium (RPE)-1 wild type cell line (courtesy by Balca Mardin and Peter Lansdorp). This cell-wide overview plot was generated using MOSAICATCHER and is typically

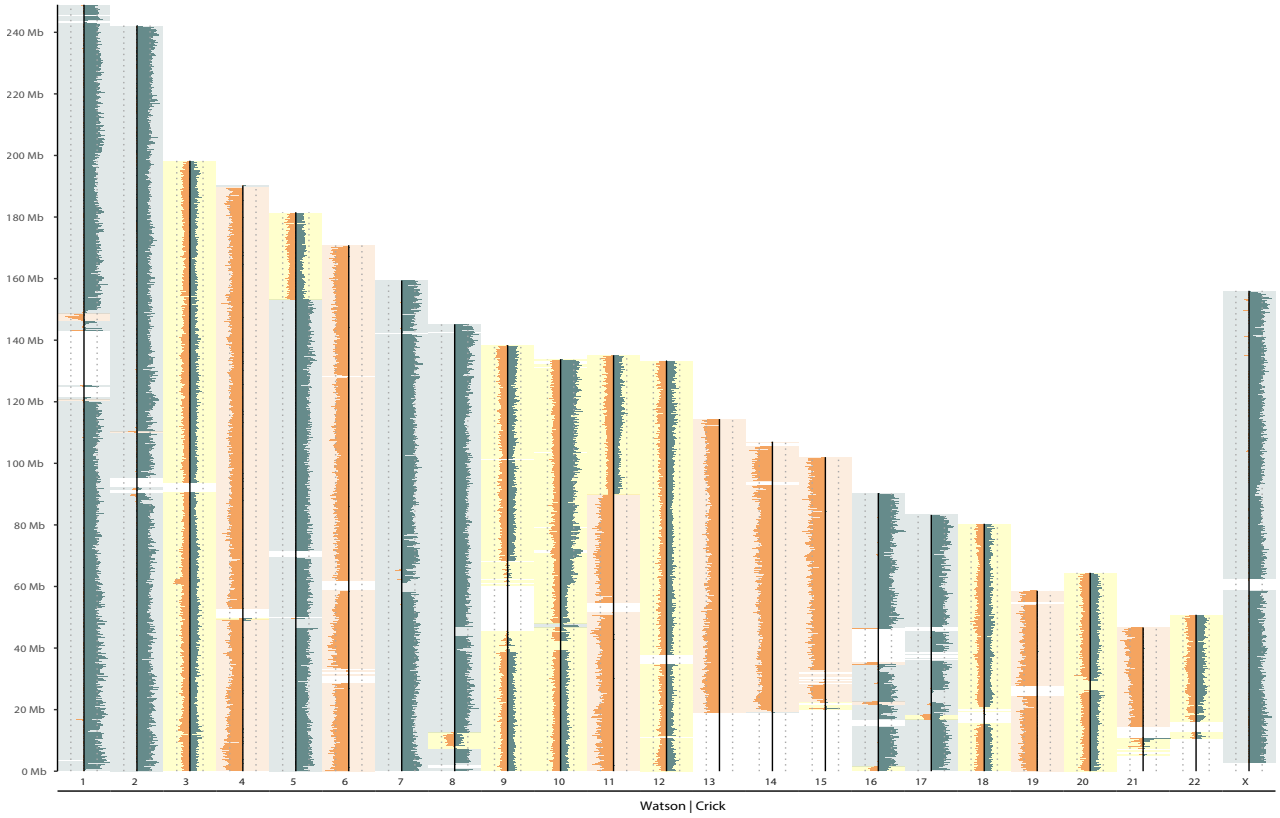


Figure 4.1: Example of a single cell Strand-seq library. Here, a single cell Strand-seq library of an RPE-I wild type cell line is displayed using the plot function of MOSAICATCHER (Section 4.2.3). Each vertical panel shows binned read counts of one chromosome, with the Watson strand on the left, in orange, and the Crick strand on the right, in blue. In the cell shown here, data reads were binned at 200 kb-bin contains, leading to a median count of 76 reads per bin depicted by dotted lines. Some regions, e.g. the centromere of chromosome 1 or the sub-telomeric regions of chromosomes 13–15, are not coverded by reads because of mappability issues; these bins are excluded from further analyses. The estimated strand inheritance state of each bin is color-coded in the background: blue for CC, orange for WW and yellow for WC. Chromosomes 5 and 11 carry SCEs, which are visible by a change in the strand inheritance state that continues to the end of the chromosomes.

the starting point for an analysis of Strand-seq data. The procedure leading to this plot is outlined briefly in Section 4.2.3.

This overview plot allows an initial judgment on the success, quality, and depth of a Strand-seq library. In the cell shown here, reads aligning to either W or C strand were binned at 200 kb resolution. The cell was sequenced at ample depth (median of 76 reads per bin), shows the Strand-seq characteristic strand inherit-

ance patterns per chromosome (WW, WC, CC) and is of high quality, which can be judged from the low number of reads on the opposite strand in WW or CC cells and a fairly even coverage distribution. Experimental parameters influencing the quality of Strand-seq libraries are discussed in detail by Sanders et al. [2017].

Figure 4.1 further gives a first impression of genomic rearrangements in this cell. For example, the region of WW reads at around 148 Mb on chromosome 1 reveals an inversion of this locus in respect to the reference assembly. The most prominent alterations are visible on chromosomes 5 and 11, though. Here, the strand inheritance states change at one position (specifically, at 155 Mb on 5 and 90 Mb on 11) and remain consistent to the end of the chromosome from there. These positions mark *sister chromatid exchange (SCE)* events, which are reciprocal exchanges between two identical sister chromatids and which can be specifically measured using Strand-seq [Falconer et al., 2012]. SCEs occur randomly across the genome and must not be mistaken with other classes of SVs for our purpose.

4.2.2. Three signals within Strand-seq data are distinctive of SVs

Strand-seq libraries reveal the presence of large structural variation. This had been shown for inversions in the past by Sanders et al. [2016]. Here, we conceptually identified seven different SV classes that can be revealed in Strand-seq data, five of which are principally discernible within a single cell (deletion, duplication, inversion, inverted duplication, and LOH), and two that become apparent across a population of cells (aneuploidy and translocation). These SV classes can be distinguished using three independent signals: (1) the normalized read coverage of a locus, which corresponds to a diploid state (2N) in a non-affected locus. (2) The strand ratio, i.e. the number of W reads over the number of C reads. Alternatively, a fraction can be used here, for example the Watson fraction W/(W+C). And (3), haplotype information. When whole-chromosome haplotypes are known for sites of SNVs, sequencing reads overlapping such variants can be queried for both their original homologue *and* their strand direction [Porubsky et al., 2016]. These alleles can be used to reason about potential rearrangements. In the following report, I refer to the two homologues (or their haplotypes) as h_1 and h_2. Below, I explain how we can utilize the combination of these signals to detect, genotype and phase the seven SV classes. Figure 4.2 contains examples of five different SV classes that we identified in RPE cell lines.

Deletions and duplications CNVs alter the total read coverage of an affected locus. A deletion decreases the coverage from 2N to 1N, i.e. by a factor of two,

	WC chromosome				WW chromosome			
			Haplotype				Haplotype	
	Cov.	W.f.	*W*	*C*	Cov.	W.f.	*W*	*C*
Reference allele	2N	50%	h_1	h_2	2N	100%	$h_1 + h_2$	-
Deletion of h_1	1N	0%	-	h_2	1N	100%	h_2	-
Deletion (homozygous)	0N	-	-	-	0N	-	-	-
Duplication of h_1	3N	66%	$2h_1$	h_2	3N	100%	$2h_1 + h_2$	-
Duplication (homozygous)	4N	50%	$2h_1$	$2h_2$	4N	100%	$2h_1 + 2h_2$	-
Inversion of h_1	2N	0%	-	$h_1 + h_2$	2N	50%	h_2	h_1
Inversion (homozygous)	2N	50%	h_2	h_1	2N	0%	-	$h_1 + h_2$
Inverted duplication (h_1)	3N	33%	h_1	$h_1 + h_2$	3N	66%	$h_1 + h_2$	h_1

Table 4.1: Distinct signatures of focal SVs in Strand-seq data. SVs can be identified based on three separate signatures of Strand-seq data: the total read coverage (*Cov.*), the strand ratio (here shown as Watson fraction; *W.f.*), and the presenece of haplotype-tagging SNVs on each strand (*Haplotpye*). The table shows how focal SV types can be inferred from these signals. This is different for WC chromosomes than for WW or CC chromosomes. For the sake of simpliciy, the table only shows the WW case and assumes heterozygous variants to affect haplotype h_1. Entries in orange are SV classes that cannot be distinguished unambigously using only coverage and Watson fraction. Specifically, homozygous inversions remain hidden in WC cells and an inverted duplication of h_1 cannot be distinguished from a duplication of h_2. Entries marked in blue cannot be phased based on coverage and Watson fraction alone. However, all these cases can be disentangled when sufficient haplotype-resolved SNVs are available or by integrating information across several cells that share an SV.

and is hence typically easier to detect than higher copy number states—the same has been observed for other read depth-based SV callers. Duplications, which increase copy number to 3N, alter the read depth only by a factor of 1.5 compared to the reference state. Homozygous duplications increase the copy number even to 4N. Homozygous deletions are marked by a complete absence of reads—they thus provide the strongest change in read depth, with the only caveat that they resemble regions of low mappability such as the centromere on chromosome 1. Homozygous deletions are hence best studied in the presence of a control sample not carrying the deletion.

In contrast to classic read depth analysis, Strand-seq additionally provides strand information. For example, a heterozygous deletion in a WC chromosome will only lack reads on one of the strands. Similarly, a heterozygous duplication will increase coverage only on one strand. In a WW or CC cell, strand information does not add supportive evidence, but when phased SNVs are available, the alleles

of only one homologue will be deleted/duplicated. Table 4.1 summarizes in detail how these signals allow to differentiate the focal SV classes that I cover here.

Copy-neutral variants Inversions do not change copy number, but become visible as a change in strand ratio. A heterozygous inversion in a WW cell, for example, switches the observed strand state to WC within the inverted locus. In a WC cell, it becomes WW or CC. Again, in the WC cell the inversion can be phased trivially based on strand directionality, whereas in the WW or CC case further SNV information has to be consulted for phasing. A homozygous inversion is a special case: It changes a WW chromosome into CC, making the locus clearly stand out, but it cannot be observed within a WC cell. This is because both alleles change their strand state, leading again to a WC region. This case can be disentangled with the help of phased SNVs (Table 4.1).

In a copy-neutral loss of heterozygosity, one homologue is "overwritten" by the other homologue, leading to extended regions of homozygousity. To reveal LOH, the alleles of SNVs must be assessed. We recently noted sporadic regions of LOH within cells of a lymphoblastoid cell line (data not shown). This analysis as well as other work on haplotype information was spearheaded by David Porubský and will hence not be covered in this dissertation.

Complex variants Complex variants can partly be discovered in Strand-seq data, too. We chose inverted duplications as a particular candidate, as this class has been found to be abundant both on the small (Chapter 3) as well as on the large scale [Chaisson et al., 2017]. As Table 4.1 shows, inverted duplications are marked by an increase in copy number as well as a flip in strand orientation. In a WC cell, such an event cannot be distinguished from a normal duplication unless haplotype information is available. Figure 4.2 includes an example of an inverted duplication within RPE-1 wild type cells.

Ploidy alterations In a Strand-seq experiment, each homologue of a chromosome is sequenced in either W or C orientation. This fundamental property remains valid even in the presence of more than two homologues. In a tetraploid cell, for example, a chromosome is inherited in any one of five states: WWWW, WWWC, WWCC, WCCC, or CCCC. By looking across a population of cells, Strand-seq can disclose the ploidy of each chromosome, assuming it is not heterogeneous (Figure 4.2). A usual way to estimate ploidy from MPS data is to assess the BAF of SNVs within a bulk sample—in a tetraploid chromosome, some SNVs will be present at ratios of 25 or 75%. Interestingly, Strand-seq detects ploidy even

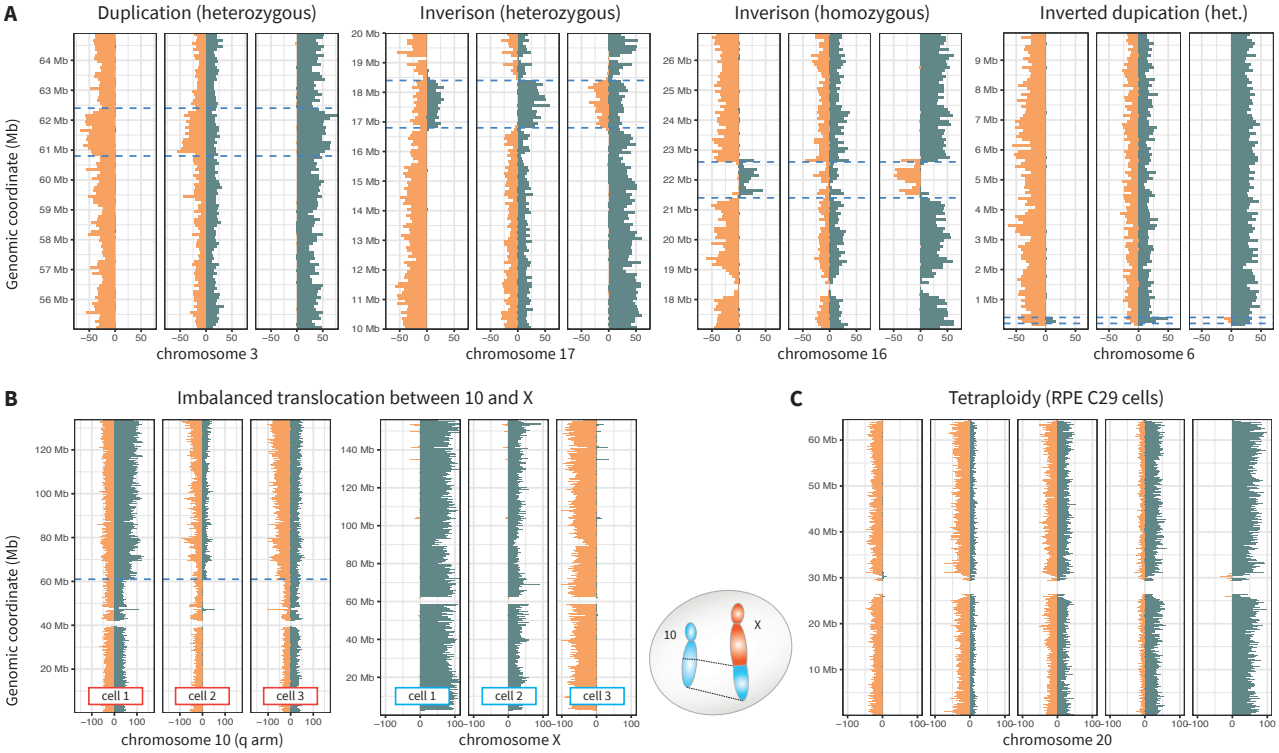


Figure 4.2: Examples of SVs in RPE-1 cells. Strand-specific count data in 100 kb bins (Watson orange, Crick blue) are shown in several chromosomal regions of RPE cells. **A:** Four loci harboring SVs are shown across three cells each. The affected loci (marked by dashed lines) show the characteristic changes in read coverage and strand state that were described in Table 4.1. **B:** A copy number gain of the q-arm of chromosome 10 is visible. The strand state of the extra copy correlates with the strand inheritance pattern of chromosome X (same cells for both chromosomes shown). This observation is best explained by an imbalanced translocation, as shown in the schematic to the right. **C:** Five chromosomes of cells from the tetraploid RPE C29 cell line (courtesy by Balca Mardin and Peter Lansdorp) were selected to represent the five possible strand inheritance patterns in a tetraploid cell: WWWW, WWWC, WWCC, WCCC, and CCCC.

in the complete absence of homologue-specific variants, for example in cells with multiple copies of the same homologue such as the CHM1 cell line [Steinberg et al., 2014]. To the best of our knowledge, this cannot be achieved by any other sequencing method.

Translocations Lastly, also translocations can be revealed using Strand-seq. A reciprocal translocation alters the strand inheritance state of two chromosomes simultaneously: An exchange between a WW and a CC chromosome, for in-

stance, would switch the strand inheritance pattern to WC in both these chromosomes. However, such changes in strand state can initially not be distinguished from randomly occurring SCE events. When a strong correlation of strand state changes between two chromosomes is observed, though, this is indicative of a translocation event. In the RPE-1 wild type cell line (Figure 4.2), we report an *imbalanced* translocation between chromosomes 10 and X, which had been noted beforehand[1]. Here, chromosome 10 shows an increase in copy number on the q-arm that seems not to be consistently in the same strand across cells. However, the strand inheritance state of the extra copy perfectly matches the strand inheritance pattern of chromosome X. This suggests that the extra copy of chromosome 10 is linked physically to chromosome X. By correlating strand states across chromosomes in this way, translocations can be discovered. The very same idea has been used to assign unmapped contigs of the reference assembly to chromosomes [Hills et al., 2013].

4.2.3. Automated SV detection with MOSAICATCHER

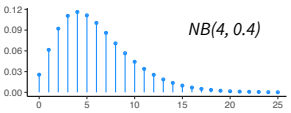
In order to capture this spectrum of SV classes, my collaborators and I designed a computational approach for SV calling from Strand-seq data. This approach is implemented in a tool called MOSAICATCHER, a first version of which has recently been developed. The core principle consists of the three steps (1) binning, (2) segmentation, and (3) classification and is explained below. My code for the first two steps is available online at `https://github.com/friendsofstrandseq/mosaicatcher`. The third part is being maintained by Maryam Ghareghani and can be found at `https://github.com/friendsofstrandseq/MaRyam`. At last, the combined workflow is available at `https://github.com/friendsofstrandseq/pipeline`.

Binning Strand-seq data is extremely sparse: the best libraries currently produced contain around 300 reads per Megabases [Chaisson et al., 2017; Sanders et al., 2017]. This is related to the fact that the Strand-seq protocol does not include whole-genome amplification. Incidentally, this also makes Strand-seq excempt from related biases in coverage. In order to work with sparse data, I applied a binning scheme that summarizes sequencing reads in windows of a given size, e.g. 50 kb. Alternatively, these bins can have variable sizes (dynamic-width bins) to accommodate for regions of low mappability. The transformation to binned

[1] In the description of the commercially available "hTERT RPE-1" cell line, a derivative of X is recognized, but chromosome 10 is not mentioned. Source: `https://www.lgcstandards-atcc.org/Products/All/CRL-4000.aspx`

counts allowed us to apply a statistical model to estimate the expected number of reads per bin—specifically, we utilized a negative binomial (NB) distribution, as I describe later. Prior to binning, sequencing reads are filtered for low mapping quality (by default, a minimum score of 10), supplementary alignments, and PCR duplicates. Moreover, I only counted each read pair once to avoid double-counting of sequenced fragments. I utilized the HTSlib library within my implementation to efficiently read and filter sequencing data.

NB distribution



Statistical distribution that describes the number of successes in a Bernoulli experiment before a given number of failures occurs. It is commonly used to describe counts in biological data, e.g. by DESeq2 [Love et al., 2014]

Given the strand-specific binned counts, I implemented a plot function to generate overview plots as shown in Figure 4.1. At first, bins that consistently appear as outliers in all cells are masked: this is typically the case in centromeric or telomeric regions with zero read coverage. Further, I designed a classifier to distinguish WW, WC, and CC states for each bin. This classifier predicts the class of each bin based on its strand-specific read counts, but it should also smoothen out fluctuations in single bins. To achieve this, I implemented a Hidden Markov Model (HMM) with a multivariate NB emission distribution. The transition probability of the HMM is chosen according the expected number of SCEs within a cell (e.g. 10 transitions across all chromosomes). Its output is used as background color in aforementioned overview plots and guides a researcher in detecting SCEs and other rearrangements in a single cell.

Segmentation The second step towards SV calling is to detect the boundaries of potential SVs. This had been done in the past for the detection of heritable (germ line) inversions by merging data of all cells in a strand-aware fashion and detecting boundaries within the merged signal [Sanders et al., 2016]. However, this approach has the disadvantage that it can mask subclonal variants, especially the ones at low allele frequency that we are particularly interested in. Boundaries have also been determined within single cells separately, e.g. via BreakPointR[2] or based on a HMM similar to our aforementioned approach [Bakker et al., 2016]. However, this leads to the subsequent challenge of forming consensus boundaries across all cells.

Instead, I explored multivariate segmentation algorithms that consider all cells simultaneously, yet still recognize them as individual cells. This is further elaborated in Section 4.2.4. Notably, the segmentation algorithm is expected to provide potential SV breakpoints across all cells, which are then tested in the subsequent step. In order increase sensitivity, we allow the segmentation to predict slightly too many boundaries that we resolve during classification.

[2]Unpublishd method by David Porubský. `https://github.com/daewoooo/BreakPointR`

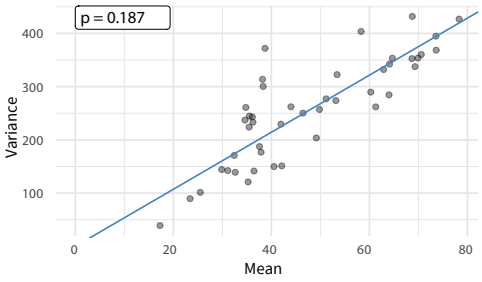


Figure 4.3: Mean variance relationship of binned read counts. Each dot represents one cell of the RPE-1 wild type cell line. Shown here are the mean number of reads per 200 kb-bin vs. their variance. In a theoretical NB distribution, mean and variance show a perfectly linear relationship with a slope p. In real data, we estimate p by linear regression without intercept (blue line).

Classification The third and final step is to test segments for the presence of SVs (theory and implementation contributed by Maryam Ghareghani; figures from me). In order to classify segments into either SV or non-SV states, we employ a Bayesian model based on negative binomial distributions. Specifically, we model the read coverage in each strand by a NB distribution, which is adjusted to capture the expected number of reads within a given region.

Both strands combined yield a joint distribution for all possible strand combinations, i.e. WW, WC, or CC, but also including abnormal states such as C, WWC, CCCC, and so on. Each of these joint strand states can then be interpreted in respect to the expected strand state in the chromosome and cell. For example, a WC state would signify a heterozygous inversion if it occurred in a WW chromosome, but no SV (or a homozygous inversion!) if it occurred in a WC chromosome. Figure 4.4 gives a detailed example of how read counts are modeled in a joint NB distribution. The dispersion parameter of the NB distribution, p, is estimated across all cells, as Figure 4.3 explains. The second NB parameter controls the number of expected reads within a locus and must hence be scaled to the size of the tested region. In line with our intuition, NB distributions yield a clearer separation for higher counts, i.e. in larger genomic regions intervals than for small SVs.

As we do not have a clear understanding of the processes that cause background reads (i.e. reads from a homologue that map to opposite, non-tamplate strand), we cannot accurately model them. We hence added a factor $\alpha \approx 5\%$ to our model to formulate a NB distribution that captures this case. For instance, in a WW regions with e expected total reads, the expected number of C reads would be modeled by $\alpha \cdot e$ and the expected number of W reads by $(1 - \alpha) \cdot e$. In the

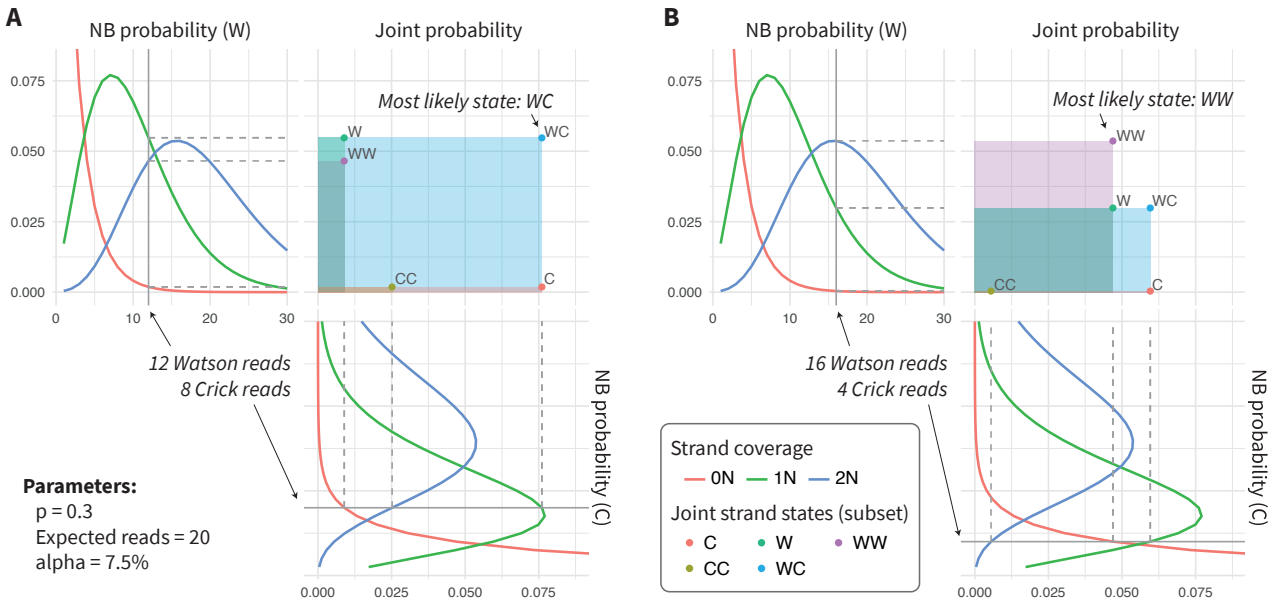


Figure 4.4: Graphical example of the negative binomial model. The subplots on the top left and bottom right (in both panels) show the probability density functions of three NB distributions. These distributions describe the probability of a given number of reads for the possible copy number states 0N, 1N, and 2N. In this example, the expected number of reads in a 2N state is 20 (blue curves). Both strands are modeled by separate NB distributions, that, when combined, yield a joint probability (area of the rectangles; top right subpanels). Here, only the joint states W, C, WW, WC, CC are shown for the sake of simplicity (other possible states would be WWW, WWC, WWCC, and so on). **A:** 12 W and 8 C reads were observed, for which the joint strand state WC is by far the most likely one. **B:** 16 W and 4 C were observed, which are best explained by a joint WW state, yet the difference to the runner up (WC) is not big. For SV calling, these joint states have to be interpreted in respect to the strand inheritance state of the chromosome: for example in a WC cell, example **A** would be rejected (WC, i.e. reference, is the most likley state) but example **B** would be considered for a heterozygous inversion (WW state).

end, the estimated NB probabilities for all SV classes are considered across cells to make a final SV call.

Additional steps Together with Maryam Ghareghani, Tobias Marschall and David Porubský, we set up an automated workflow with the aim to process Strand-seq data all the way from raw sequencing files to final list of SV predictions. This pipeline, which is based on the workflow engine Snakemake [Köster et al., 2012], involves many other relevant steps in addition to the tripartite calling procedure explained above, two of which shall be mentioned here. First of all, the dominant strand state has to be determined for each cell and chromosome in order to

guide the SV classification. This includes the detection of SCE events. In order to achieve this, Venla Kinanen implemented a heuristic method to estimate strand states and SCEs based on the output of the strand state HMM of MOSAICATCHER. This performed well when it was tested against experts' opinions on real-world Strand-seq data.

Secondly, haplotype information must be annotated. David Porubský hence incorporated functionality to annotate haplotypes in WC chromosomes based on the STRANDPHASER [Porubsky et al., 2016] tool. Interestingly, the phasing of SNVs can be performed from Strand-seq data alone [Porubsky et al., 2016], which was also built into the the workflow. With sufficient coverage, Strand-seq data can even be used for SNV detection, which is a required input to phasing. Together, this workflow supplies all input for subsequent SV calling on Strand-seq data without the requirement for additional data sets.

4.2.4. A multivariate segmentation algorithm to find SV breakpoints

The objective of the segmentation step is to group the genome into consecutive regions of the same copy number. In Strand-seq data, I apply this principle to both strands separately. It can be formulated as a problem of placing k breakpoints in such a way, that the resulting consecutive segments best represent a single copy number state each—"best", in this case, refers to the squarred (Gaussian) error, which shall be minimized. This problem gained much attention with the availability of comparative genomic hybridization arrays to map CNVs. Several methods were put forward for this task and a popular one is circular binary segmentation [Olshen et al., 2004; Venkatraman et al., 2007]. Today, these techniques are also commonly applied to MPS data. The number of data points (hybridization loci in arrays, or genomic bins in MPS experiments) critically impact the runtime of such methods. Circular binary segmentation is efficient in this respect, but it uses a heuristic approach (e.g. placing one breakpoint after the other) that does not guarantee to find an optimal solution. In contrast, the TILINGARRAY [Huber et al., 2006] package uses a dynamic programming algorithm to find the optimal solution according to the squared error criterion. As this algorithm does not scale well to deep MPS data, other approaches have been proposed, such as the Group fused LASSO formulation [Bleakley et al., 2011].

As Strand-seq data is inherently sparse, problem size is not a limiting factor in our application. For example with 50 kb bins, chromosome 1 still only contains ~5000 data points. I hence designed and implemented an algorithm for Strand-

seq segmentation that is based on the TILINGARRAY principle[3]. The TILINGARRAY algorithm internally uses a cost matrix that defines how expensive (in terms of variance) each consecutive segment is. A dynamic programming algorithm then choses the optimal combination of breakpoints. The calculation of the cost matrix in TILINGARRAY assumes the changes in signal across all replicates to be the same, i.e. an increase at the same locus in all samples.

For Strand-seq data, I needed to relax this criterion to capture inversions—where one strand is increased and another decreased—and mosaic variation. I hence re-defined said cost matrix to allow individual jumps within each strand. Notably, the positions of the breakpoints are still consistent throughout all cells. I implemented this algorithm using C++ as a central part of MOSAICATCHER.

4.3. Results II: Strand-seq simulations to explore the limits of MOSAICATCHER

A principal challenge during the implementation of MOSAICATCHER was to measure its performance. Initial results on RPE-1 cells looked very promising, but did not allow a systematic investigation of the limitations of our method. In order to make this possible, I developed a framework to simulate Strand-seq data. Furthermore, with this framework I can introduce SVs into the virtual cells in a fully controlled manner. To give an example, Figure 4.5 shows different simulated SV classes at varying sizes. Naturally, such simulations represent idealized conditions, yet they were designed in a way to closely reflect essential properties of real-life data. These simulations allow us to test the correctness of our method during its development. More importantly though, they make it possible to explore the theoretical limitations of MOSAICATCHER, e.g. in terms of the smallest SV size or lowest variant frequency that can be detected reliably. Due to the simulations we can focus our efforts on continously refining the methodology in order to push these limits.

4.3.1. Development of a versatile simulation framework

Binned read counts of Strand-seq libraries approximately follow a negative binomial distribution. This impression is supported by the near linear relationship of

[3]Original from `http://bioconductor.org/packages/release/bioc/html/tilingArray.html`; implemented independently (including major changes) in MOSAICATCHER (see the file `segmentation.hpp` at `https://github.com/friendsofstrandseq/mosaicatcher`)

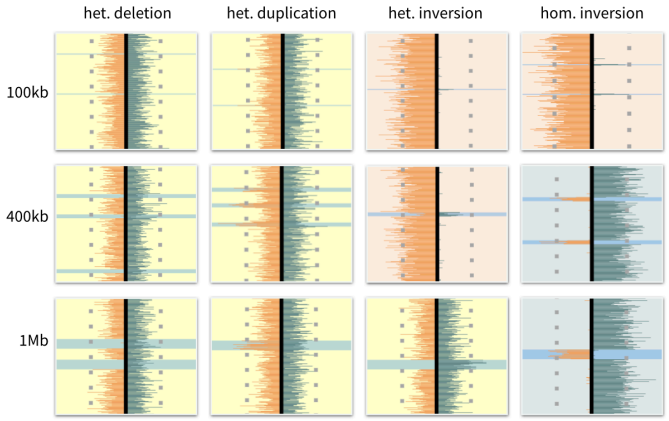


Figure 4.5: Examples of simulated SVs in different sizes. Cells were simulated in 50 kb bins with a median of 20 reads per bin. Then, SVs were inserted at sizes of 100 kb, 400 kb, and 1 Mb and are highlighted by a light blue background.

mean and variance that we observed (Figure 4.3). Hence, in order to computationally generate Strand-seq data, I devised a tool to sample binned read counts from a negative binomial distribution. The NB parameters were chosen in a way to closely resemble real data, but can be changed by the user.

The simulation consists of three steps. At first, a basic read coverage of each homologue is generated via sampling from a NB distribution. To do that, the bin size, the NB parameter p and the expected number of reads per bin have to be specified (see Table 4.2). By default, a set of 24 chromosomes with the typical sizes of human chromosomes 1–22, X and Y are generated for a user-defined number of cells.

In the next step, SVs are introduced into these homologues. The variants must be predefined in a configuration that lists the position, size, SV type and variant frequency of each SV; however, I included routines to generate this file by randomly distributing SVs of varying size across the genome and simultaneously avoiding overlapping loci. Currently, the simulations allow four different SV classes, namely deletions, duplications, inversions (each of them either heterozygous or homozygous), and inverted duplications. These variants are then applied to the h_1 homologue (or to both) by changing the strand-specific coverage.

Internally, each homologue is composed of two read counts per bin: the reads sequenced in the orientation the homologue will later inherit, called *forward* for now, and reads sequenced in *reverse* orientation, which are all initialized to zero. An inversion, for example, is incorporated by flipping reads from the forward

Parameter	Default	Description
n	100	Number of cells
w	50,000	Window size
c/C	10/20	Expected coverage. Sampled uniformely from $[c, C]$
p	0.5	NB parameter p
a	0.1	Fraction of non-zero background bins
s	4	Expected number of SCE events per cell
z	0.1	Fraction of reads that can be phased

Table 4.2: Major parameters controlling Strand-seq simulations. These parameters can be specified during Strand-seq simulations. The coverage is specified as the mean number of reads per bin and sampled uniformely form the allowed range for each cell. The parameter a controls the fraction of bins that allow background reads: for those bins, the number of background reads is sampled from a geometric distribution; a is hence only losely related to α (from the classification).

to the reverse orientation in a specific region. A duplication, on the other hand, doubles the forward read counts and does not alter the reverse counts. Importantly, the coordinates of an SV do not need to align with the boundaries of bins—this would be a very unrealistic requirement. Bins that are only partially affected by a SV will also only have a fraction of their read counts altered. Variants with frequency $f < 1$ are incorporated only into subset of cells, chosen with a probability f for each cell. The user-defined variant frequency f hence represents only the *expected* fraction of cells carrying the SV. This reflects the situation within a cell population that underlies a random sampling of cells during a Strand-seq experiment.

At last, cells are rendered into Strand-seq libraries. Initially, I add spurious background reads (in reverse) to a subset of bins to reflect imperfect experimental conditions. Then, each homologue is randomly inherited as either W or C strand, with the reverse reads (containing SVs and spurious reads) on the opposite strand. I further allow the strand state of a homologue to switch at any position with a very low probability—this creates SCE events. The final W and C counts lose the information about their original homologue, as forward reads from a W homologue are merged with reverse reads from a C homologue (just like when reads from Strand-seq experiment are mapped to the reference genome). However, prior to the merging I allow a fraction of reads in each bin (typically 10%, but this is chosen from a binomial distribution within each bin) to be phased, i.e. to virtually overlap haplotype-specific SNVs. Hence, we know both strand orientation

as well as homologue for this subset of reads, which is written to a separate table.

4.3.2. Performance of the segmentation algorithm

The simulation framework offers an excellent opportunity to benchmark MOSAICATCHER. Our tripartite calling procedure is based on first finding potential breakpoints, and then testing the resulting segments for SVs—these parts can hence be evaluated independently. Here, I benchmarked the segmentation algorithm introduced in Section 4.2.4 using different classes of simulated SVs. I was especially interested in how well breakpoint detection performed for low SV sizes.

I simulated a number of Strand-seq experiments under different conditions. In each simulation, 200 cells were generated in 50 kb bins with an expected number of 20 reads per bin. I then distributed one of three SV classes, namely deletion, duplication, or inversion, of a fixed size across chromosome 1. Next, this procedure was repeated for the SV sizes 100 kb, 400 kb, 1 Mb, and 2 Mb. As mentioned previously, the random placement of SV boundaries is independent of the binning scheme. A 100 kb event, for example, hence rarely overlaps exactly two bins, but rather spans one bin completely and two neighboring bins partially.

After that, I applied the segmentation algorithm of MOSAICATCHER with increasing numbers of allowed segments (which is a parameter of the algorithm) to all these data sets. Too few segments are bound to miss some SV breakpoints, whereas too many segments are likely to add false predictions. Based on the number of correctly identified breakpoints and the number of spurious predictions, I calculated receiver operator characteristics. The results are shown in Figure 4.6.

These initial results show that the segmentation algorithm performs well in detecting SV breakpoints in simulations. Large SV, in the range of 1 Mb and above, can be identified exhausitvely for all three SV classes. The number of required breakpoints depends on the number of SVs present within the cell, which is of course not known *a priori*. This is why the receiver operator characteristics plots show a rectangular behavior: with an increasing number of allowed segments more and more SV breakpoints are successfully detected, until they were all found. Then, more segments add only false positive breakpoints. Smaller SVs are expectedly more difficult to capture, yet even among the 100 kb events, around three quarter of the breakpoints could be detected while generating less than 20 false positive predictions. At a size of 400 kb, again all breakpoints were found with slightly lower specificity compared to large SVs. This means that, in order to capture small SV (in the range of 100-400 kb), we need to allow the segmentation to create more segments than expected (oversegmentation), which then have to

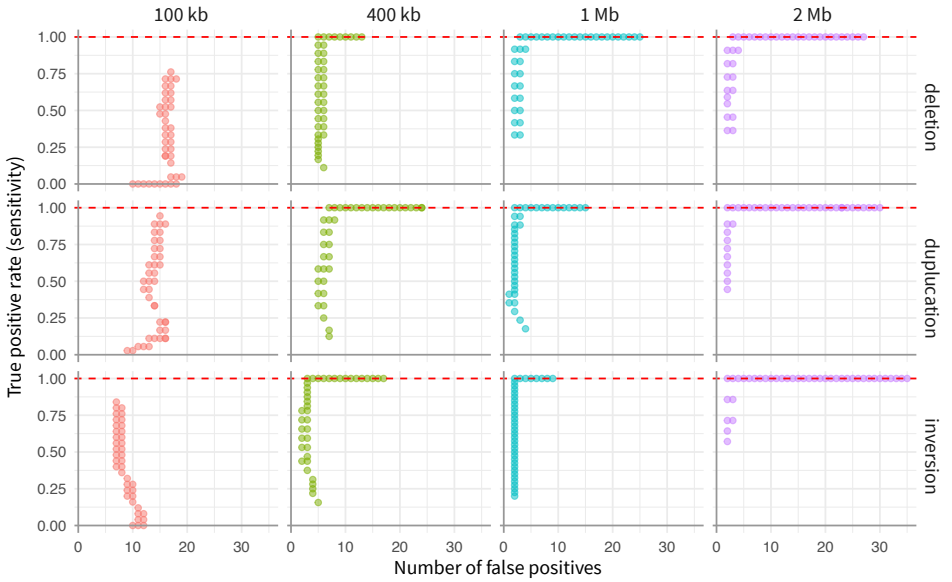


Figure 4.6: Performance of segmentation on simulated SVs. Here, I simulated a random number (around 20-50 per simulation) of each of three heterozygous SV classes of different sizes on chromosome 1. The underlying simulated Strand-seq data contains 200 cells, binned at 50 kb resolution with a mean of 20 reads per bin. Then, for each of these data sets, I calculated receiver operator characteristics by measuring the fraction of correctly identified SV breakpoints (sensitivity; correctness means that the segment boundary is not more than 3 bin sizes away from the simulated breakpoint) vs. the number of spurious breakpoints. Each dot represents these values for a specific number of allowed segments in the range from 10-50. SVs are present in 100% of the cells.

be subjected to statistical testing in the classification step.

4.4. Conclusions and outlook

4.4.1. Summary of findings

Strand-seq is an emerging sequencing technique with the unique ability to resolve single homologues by sequencing only their template strands. While it had been used in the past for SCE mapping [Falconer et al., 2012], phasing [Porubsky et al., 2016], and inversion discovery [Sanders et al., 2017], we demonstrated here that Strand-seq data reveals the presence of at least seven different SV classes. Based on an integration of three separate signals of Strand-seq data, these SV classes can be detected, genotyped and assigned to a haplotype. To give substance to this claim, I highlighted examples of five different SV classes that we identified in recently sequenced retinal pigmented epithelium (RPE) cell lines. Moreover, as Strand-seq operates on single cells, these SVs can be analyzed even if they are present only in a small fraction of the total cell population. This enables us to study SVs in the context of somatic mosaicism. Previously, these studies had been severly limited in their ability to assess SVs other than CNVs or struggled with lower variant frequencies. Strand-seq based SV calling will thus open whole new possibilities to study the yet underexplored role of SVs in cellular heterogeneity.

I then presented our computational method called MOSAICATCHER that was designed to achieve fully automated SV calling from Strand-seq. MOSAICATCHER operates by grouping sparse read data into bins, segmenting the genome to find the boundaries of potential SVs and subsequently classifying these segments. I designed and implemented a multivariate segmentation algorithm based on a quadratic error term that captures two important characteristics: It recognizes single cells separately—instead of merging their signals—but still defines breakpoints across the population of cells. This approach intuitively captures large variation present in few cells or small variants present in many cells. The SV classification utilizes negative binomial distributions—for which I provide evidence that they approximate real data well—to calculate probabilities for each SV class based on the number of strand-resolved reads within the locus. Due to our Bayesian inference strategy, we are able to supply prior knowledge and can easily integrate SV predictions across the population of cells. While this method is currently under development, we have already seen positive evidence—for example in RPE cells and publicly avialble data sets—that this approach performs well.

In order to test the limitations of MOSAICATCHER, I designed a versatile framework for the simulation of Strand-seq data. This set of tools can computationally generate Strand-seq libraries containing up to four classes of focal SVs of arbit-

rary sizes. Read counts are modeled from the same NB distributed that was fitted to real data previously. I then used this framework to explore the capabilities of the segmentation algorithm with respect to small SV sizes. The results suggest that even at an SV size of 100 kb, which corresponds to only two data points in our binning scheme, the majority of breakpoints can be correctly detected. SVs larger than 1 Mb are found robustly. The capabilities of MOSAICATCHER will have to be tested further using simulations and real-life data, but these initial results suggests that the lower detection limit of MOSAICATCHER will likely be in the size range of 100-500 kb.

4.4.2. Open challenges

Despite good progress in the development of MOSAICATCHER, we are facing several challenges that yet have to be overcome. First of all, our current approach covers the detection of focal SVs, but not yet of translocations and aneuploidy. As their detection works principally different than for focal SVs, I have been developing independent methods for them in parallel. These approaches, once they have been refined and thouroughly tested, shall later be integrated into a single software tool. A second point is that, in its current state, MOSAICATCHER does not utilize the information from phased SNVs, although they can already be annotated using STRANDPHASER. Consequently, SVs can only be assigned to homologues on WC chromosomes, where homologues are separated by their strand.

Third, the segmentation algorithm of MOSAICATCHER currently requires the number of segments as an input parameter, which is not known a priori and for which we need to find an appropriate way of estimation. Others have used the Bayesian Information criterion in similar situations [Huber et al., 2006], but it did not seem to work well in our scenario. Moreover, we do not yet know how sensitive the segmentation is towards low variant frequencies—in case it performs poorly, we will revisit single-cell segmentation again. Also, the segments can only be called at the bin coordinates (typically every 50 kb), which only sporadically align with SV breakpoints. In order to improve this, we are experimenting with a shifted binning scheme. We further thought of a subsequent refinement procedure that directly accesses the read mapping positions in SV-carrying cells to estimate breakpoints at higher accuracy.

At last, the classification step was work in progress at the time of writing. In its current state, it accurately classified segments in single cells, but it did not yet integrate this information across the cell population. Single outliers in read counts, which occur both in real as well as in simulated data, can lead to a certain

SV state receiving the highest likelihood. This can falsely trick an individual classifier into rejecting the null hypothesis of a reference state. To tackle this issue, we plan to introduce a size-adjusted prior distribution that assumes a reference state in most positions. Further, we are considering to restrict the model to biallelic SVs—i.e. only variant class at the same time at each locus. This way, single outlier cells, such as a spurious triplication call in a duplication locus, could be corrected. A related challenge originates from the unevenly distributed coverage of Strand-seq data. In addition to mappability and GC content, Strand-seq coverage can be affected by nucleosome positioning (due to MNase digestion), which is tissue-specific—this aspect is further explored by Hyobin Jeong in our laboratory. Because of that, a one-fits-all solution (such as the widely-used GC correction) might belie expectations of a proper normalization. We are currently testing methods for normalization based on control samples, or methods that operate within the data itself, such factor analysis [Stegle et al., 2012] or a specifically designed expectation-maximization algorithm.

4.4.3. Future directions

Our immediate future steps towards the publication of MOSAICATCHER are the completion of the method, addressing some of the points mentioned above. We then plan on demonstrating its performance in cell line data. Besides the RPE-1 wild type cell lines I showed here, we can further apply MOSAICATCHER to a set of previously published, high-quality Strand-seq libraries (Chaisson2017). This data includes more than 1000 cells across 9 individuals from family offspring trios and is likely the richest Strand-seq data set available to date. Moreover, SVs have been mapped comprehensively based on a combination of techniques in these samples, making it a perfect test scenario for MOSAICATCHER. The cell line data is expected to be very homogenous, but we already observed sporadic stretches of LOH (unpublished work by David Porubský). Moreover, we have Strand-seq libraries of multiple hyperploid cells, including the aforementioned RPE C29 line, to explore the identification of aneuploidy with Strand-seq data.

Moreover, applications of MOSAICATCHER beyond the simple demonstration of feasibility are already planned. For example, Ashley Sanders recently sequenced clones derived from fibroblast, which had been previously reported to harbor somatic mosaicism including large structural variation [Saini et al., 2016]. Another project, led by Karen Grimes, aims at studying cellular heterogeneity within the blood in the context of ageing. Previous studies had shown outstanding degrees of somatic mosaicism, from point mutations to aneuploidy, in the hematopoietic

lineage [Razzaghian et al., 2010; Holstege et al., 2014]. Also, cancer is another prime target to be analyzed using Strand-seq and MOSAICATCHER. Finally, there is a lot of excitement in the community these days about building atlases of cell types in humans and model organisms. Given the relevance of mosaic SVs in human disease, MOSAICATCHER may even pave the way for single-cell projects like the Human Cell Atlas [Regev et al., 2017] to extend their studies genetic mosaicism and its consequences on health and disease.

Conclusions and Discussion

5

The goal of this dissertation was to explore the potential of emerging DNA sequencing technologies to discover, characterize and validate structural variation. These technologies had brought large improvements to related fields such as chromosome phasing and *de novo* assembly, but are particularly suited for the purpose of SV characterization, too. Throughout the projects described herein, I presented three concrete approaches of how these techniques advance the detection of difficult SV types. I was able to scrutinize SVs to an extent that had not been possible beforehand based on standard MPS experiments. Importantly, the SVs unraveled by my work led to novel insights into the complexity and functional role of SVs. I further developed new computational approaches to detect and analyze SVs based on these techniques, proved their utility, and made them available to the community.

While each chapter contains a conclusion on its own, I here summarize again the major findings of my work and put them into the context of recent developments within the community.

5.1. Complex inversions in the human genome

Inversions are a SV class of outstanding relevance for human disease [Feuk, 2010], yet they are especially difficult to detect and they eluded ascertainment also in the 1000 Genomes Project. As I show in Chapter 2, I was able to validate hundreds of inversion loci by using targeted long-read sequencing data from both PacBio and ONT MinION platforms. This revealed that more than 80% of the inversion loci predicted from MPS indeed carried an inversion signature. Strikingly, this verification had previously not been possible via PCR experiments. This solved my first research goal and a principal challenge of the overall study [Sudmant, Rausch, et al., 2015].

Moreover, I then found that the majority of predicted loci contained not simple inversions, but complex variants containing inverted sequence. I categorized them into five major classes, which included inverted duplications as the most frequent event. These insights had only been possible due to the ability of long-read techniques to span complete loci around predicted inversions. My analyses critically relied on the visualization tool Maze, which I developed simultaneously and which I made available to the public (`https://github.com/dellytools/maze`).

The unforeseen amount of complex variation resulting from my work and the work of others was one of the key lessons learned from the 1000 Genomes Project's SV study. The function and origin of these complex sv classes remained uncharted, though. I thus carefully analyzed the breakpoints of complex SVs with the goal to infer the mechanisms they originated from. The evidence I found was not distinctive of any precise mechanism that might have formed these SVs, but it suggested that several of the seemingly very different classes might originate from the same mutagenic process, with slight evidence for replication-based mechanisms such as MMBIR.

Intrigued by the unforeseen amount of complex variation revealed in the 1000 Genomes Project, others continued to study this SV class in human genomes [Chaisson et al., 2014; Collins et al., 2017]. Using the emerging 10X Genomics technology and mate pair sequencing, Collins et al. [2017] even extended the five classes that I reported to a total of 16 different complex SV classes (which they call cxSV), more than 80% of which contained inverted sequence. This further emphasizes the that this phenomenon was previously underappreciated, as I predicted. They also note that these complex events might have been created by a replicative mechanism such as MMBIR.

My work and the subsequent finding of Collins et al. [2017] underline the prevalence of complex inverted rearrangments—leading to the notion of the "morbin" human genome. Whereas my work revelead complex SVs in healthy individuals, Collins et al. [2017] found them in patients with autism spectrum disorder. The functional role of these SV classes is not yet understood, but our results suggest that inverted and complex variation can and and should be detected, especially in the context of genetic studies around human disease.

Long-read sequencing on the rise

In the mean time (especially since 2014, when I started this project) an increasing number of studies were published by others that utilize long-read sequencing technologies for SV detection and related tasks. Notably the PacBio technology,

which became commercially available in 2011, has gained many users. Initially, PacBio had been used to perform targeted validation experiments like I showed here, and computational tools have been proposed since to facilitate this approach [M. Wang et al., 2015; Rudewicz et al., 2016]. The method by M. Wang et al. [2015] is even designed specifically for SV characterization and uses a breakpoint visualization approach very similar to the one I developed for Maze.

However, the applications of PacBio have long gone beyond this level. Due to increases in throughput, WGS has become possible in a more cost-effective fashion. For example, prior to 2014, Chaisson et al. [2014] still needed 243 SMRT flow cells to achieve a coverage of 40 x in a human genome, whereas the most recent developments (i.e. the PacBio Sequel system) promise a 10 x coverage from only 4 SMRT cells[1].

The capabilities of the PacBio technology have especially caused a stir in the plant genomics community, which had been affected by the limitations of short-read MPS to a special degree [Bickhart et al., 2014]. Notably, the hope is to perform *de novo* assembly of highly repetitive, or even polyploid genomes [C. Li et al., 2017]. An accurate assembly would make the discovery of SVs trivial—it could simply be done by sequence comparison. However, the problem of *de novo* assembly from PacBio data alone is not yet considered to be solved, despite a number of available software tools [Chin et al., 2013; Chin et al., 2016; Koren et al., 2017; Koren et al., 2018] and the attention of renowned scientists[2].

Nevertheless, PacBio WGS data was specifically used to study SVs in the human genome. Notably by Chaisson et al. [2014], who utilized a local assembly approach to detect insertions and deletions in a haploid CHM1 cell line. Remarkably, they found tens of thousands of SVs that had not been detected beforehand. They further observed an insertional bias (more insertions than deletions) of short tandem repeats, ALU elements and complex variation. In addition, they could close (or reduce in size) dozens of gaps that were missing in the reference genome (GRCh37). Obviously, this had not been possible based on standard MPS approaches. These results highlight shortcomings of the human reference assembly, which does not represent the human genome in its entireness. Even more though, they highlight the capabilities of long-read sequencing for SV detection.

Since then, SV detection based on PacBio data has been further improved and new software tools have been developed using read mapping or assembly approaches [Pendleton et al., 2015; Huddleston et al., 2017]. In a recent study, we utilized PacBio and other techniques for an unprecedentedly deep characteriza-

[1] https://www.pacb.com/blog/new-software-polymerase-sequel-system-boost-throughput-affordability/

[2] E.g. the efforts of Gene Myers, see https://dazzlerblog.wordpress.com/

tion of SVs in the human genome [Chaisson et al., 2017], which to a large part relied on PacBio technology.

ONT technology has seen many improvements, too, and is slowly gaining popularity. Recently, the *de novo* assembly of the genomes of yeast, *Caenorhabditis elegans*, and *Drosophila melanogaster* were demonstrated using ONT seqeuncing [Istace et al., 2017; Tyson et al., 2017; Solares et al., 2018]. What is especially interesting to note, though, is the pace of these technological improvements. In the latest study by ONT, Jain et al. [2018] used a novel protocol capable of generating sequencing reads with a N50 value of more than 100 kb (and a maximum of 880 kb). This is a length so far unachieved by PacBio, which typically yields a maximum read length below 100 kb[1].

Together, these technological improvements in long-read sequencing will facilitate studies on SVs that have been overlooked in the past—they might even, at some point in the future, make whole-homologue *de novo* assembly possible, which would directly reveal the full spectrum of SVs within an indiual's genome.

5.2. Effects of SVs on gene expression and chromatin organization

In Chapter 3, I set out to study the functional consequences of SVs in respect to gene expression and chromatin conformation. My first goal within this collaborative project was to characterize the variants present in highly rearranged balancer chromosomes. I achieved this by utilizing deep WGS and Hi-C data. Among many other aspects, I discovered the exact breakpoints of large rearrangements of the balancer chromosomes. In the meantime, others had mapped these breakpoints, too, and reassuringly, our results perfectly matched their findings [Miller et al., 2016; Miller et al., 2018].

However, through the technological advantage of Hi-C data, I could additionally detect precisely (in 2 cases) or approximately (in 1 case) the breakpoints that had been missed by these studies. In addition, I utilized haplotype-resolved Hi-C maps to validate large rearrangements including an inversion, and a duplication of 258 kb. The large duplication most likely inserted in reverse orientation next to the original copy, which I concluded from the differential contact frequencies around the affected locus. Together, these findings clearly show the benefits of Hi-C for the characterization of large SVs.

Afterwards, I implemented a test for allele-specific expression that utilizes multiple biological replicates and that corrects for effects of maternally deposited

RNA. I found that changes in expression occur almost everywhere across the genome and that they appear not to be caused by enhancer hijacking, as had been observed in previous studies (Section 3.1). Instead, SVs alter expression via alternative mechanisms such as dosage effects or chimeric expression of transcripts through mobile elements (summarized in Section 3.8). Our findings appear contrary to what has been seen in other scenarios; however, I argued that this might be a result of natural selection in both the other studies and in ours. In conclusion, balancer chromosomes show a remarkable robustness towards the huge rearrangements and other variation that they carry, and the potential effects of enhancer hijacking mechanisms appear to be buffered. I speculated that this buffering might be caused by other forms of variation, such as SNVs, or possible via changes of the epigenome.

I think that these results will complement previous studies and lead to a more holistic view on the role of chromatin architecture. The manuscript was in preparation at the time of writing this thesis.

SV characterization via Hi-C

I demonstrated in Chapter 3 how Hi-C data can be utilized for SV characterization. Naturally—and considering the popularity of Hi-C and the amount of publicly available data—this observation was made by others, too.

The prospects of Hi-C for purposes other than studying chromatin conformation have been noted early in the field of *de novo* assembly: Kaplan et al. [2013], for instance, predicted that Hi-C could facilitate assembly and assigned unplaced contigs to the human genome; Burton et al. [2013] created scaffolds of human, mouse, and *Drosophila* genomes based on Hi-C and MPS data; Selvaraj et al. [2013] successfully extended the idea to haplotyping; And recently, the mosquito *Aedes aegypti*, vector of the Zika virus, was assembled using Hi-C data [Dudchenko et al., 2017].

The core idea of Hi-C-based SV detection is the identification of characteristic alterations in contact frequencies. The presumably first SVs detected using Hi-C were translocations in cancer cell lines, which were detected during the search for *trans* interactions between chromosomes [Rickman et al., 2012]. This idea was then augmented towards the detection of arbitrary rearrangements. For example, large rearrangements in scrambled synthetic yeast genomes were recently studied based on Hi-C [Mercy et al., 2017]. Moreover, Putnam et al. [2016] explored the potential of Hi-C to identify inversions. And more translocations could be identified in cancer cell lines [Barutcu et al., 2015; Ay et al., 2015; Harewood et

al., 2017]. Eventually, Hi-C based SV detection was further extended to CNVs [Harewood et al., 2017; Yanjian Li et al., 2018].

Hence in summary, the idea of applying Hi-C for SV characterization has been commonly known beforehand. Moreover, recent efforts within the community have advanced the state of the art far way beyond the application I presented here. Nevertheless was Hi-C-based SV detection a highly important step within our study and it allowed us to gain novel insights on the relationship of chromatin conformation and gene regulation.

5.3. Structural variation detection in single cells

Finally, in Chapter 4 I present a novel method for SV detection on the single-cell level, which is currently under active development. This method termed MOSAICATCHER allows, for the very first time, the detection of multiple different SV classes based on single-cell Strand-seq data. In a first step, my collaborators and I devised a detailed scheme about how each SV type can be detected, genotyped and phased within a set of single-cell data. This conceptual work is largely based on previous experience with Strand-seq data, but I presented examples of five SV classes in a new, yet unpublished data set of retinal pigmented epithelium cells. In the next step, I conceived and implemented a framework to simulate Strand-seq data. This framework models Strand-seq data in terms of a negative binomial distribution, for which I provided evidence that it reflects well the properties of real data. The framework can then be used to simulate single-cell Strand-seq libraries of arbitrary sequencing depth and incorporating four different SV classes at any desired size and subclonal fraction. Simulations within this framework enable us to explore the theoretical limitations of MOSAICATCHER. At last, I designed and implemented an algorithm for data binning and segmentation, which covers two of three steps of our conceptual SV calling procedure. The segmentation algorithm uses the multivariate strand-specific read depth to find the boundaries of potential SVs based on a quadratic error term and I showed that it performs well in simulations. The last goal—implementing and applying this method—has not yet been reached at the time of writing. MOSAICATCHER will, once completed, greatly facilitate studies of somatic mosaicism, e.g. in the context of ageing or cancer, which had recently been severly limited—if not in respect to SVs.

5.4. Concluding remarks

Copy-number neutral and complex forms of structural variation have often been neglected in genetics studies compared to CNVs. Consequently, less is known about their prevalence and their role in health and disease. This is owed to technical limitations in their detection based on commonly used techniques such as MPS.

Here, I presented three concrete examples of how emerging technologies improve the detection and characterization of SVs. Utilizing these technologies allowed me to detect an unforeseen amount of complex inversions in the human genome, and to shed new light onto the functional impact of SVs. I further developed a computational tool for the characterization of complex rearrangements from long-read data, including the fine-mapping of their breakpoints, and a novel approach for SV detection within single cells.

Together with efforts by others in the community, these new approaches will enhance our abilities to discern such SVs both in the germ line as well as in the context of somatic mosaicism. This will eventually contribute to a deeper understanding of genetic variants and their potential functional roles.

APPENDIX

List of Software Tools

A

Blasr: PacBio read mapping. https://github.com/PacificBiosciences/blasr

bwa mem: Read mapping. https://github.com/lh3/bwa

wgs-Assembler: *De novo* assembly. Supported up to version 8.3 (May 2015). Successor Canu at http://canu.readthedocs.io

Delly: SV detection. https://github.com/dellytools/delly

DESeq2: Differential gene expression analysis. https://bioconductor.org/packages/release/bioc/html/DESeq2.html

FreeBayes: SNV/indel detection. https://github.com/ekg/freebayes

ggBio: R-package for visualization of genomic data. https://doi.org/10.18129/B9.bioc.ggbio

HTSlib: Workflow engine. https://github.com/samtools/htslib

Last: Alignments. http://last.cbrc.jp

Maze: Long read visualization. https://github.com/dellytools/maze

Mosaicatcher: Our single-cell SV caller (work in progress). https://github.com/friendsofstrandseq/mosaicatcher

MUMmer: Maximal matches/alignments. http://mummer.sourceforge.net

Primer3: Primer design. https://primer3plus.com/cgi-bin/dev/primer3plus.cgi

Quiver: PacBio software. Its successor Arrow at https://github.com/PacificBiosciences/GenomicConsensus

SAMtools: General purpose tool for MPS data. https://github.com/samtools/

Snakemake: Workflow engine. https://snakemake.readthedocs.io

SPAdes: *De novo* assembly. http://bioinf.spbau.ru/spades

STAR: RNA-seq read mapping. https://github.com/alexdobin/STAR

StrandPhaseR: Haplotype phasing from Strand-seq data. https://github.com/daewoooo/StrandPhaseR

vcflib: General purpose tool for variant files. https://github.com/vcflib/vcflib

vt: General purpose tool for variant files. https://genome.sph.umich.edu/wiki/Vt

Supplementary Information to Chapter 2

B

Table B.1: List of inversion loci with nucleotide resolution. These are the loci that could be resolved to nucleotide resolution with at least one of the three different approaches, namely Sanger sequencing, PacBio long-read assembly or Illumina short read assembly. Presented coordinates span the actual inversion.

Chr	Start	End	Sample	SVtype	Source
1	2,810,097	2,813,130	HG00240	invdup	Sanger
1	31,343,442	31,345,950	NA18626	invdup	PacBio
1	44,821,288	44,824,322	NA12717	complex	PacBio
1	61,058,347	61,067,336	pooled	proxdup	Illumina
1	92,130,537	92,133,743	NA12760	simple	PacBio
1	104,509,460	104,515,918	pooled	invdup	Illumina
1	104,577,303	104,585,534	NA19462	invdup	PacBio
1	115,678,410	115,684,369	pooled	proxdup	Illumina
1	116,117,516	116,124,785	HG00683	invdel2	PacBio
1	118,866,376	118,875,456	pooled	proxdup	Illumina
1	119,491,847	119,494,184	HG02187	invdup	PacBio
1	197,756,058	197,758,630	HG01501	simple	PacBio
1	209,934,001	209,937,071	NA19648	proxdup	Sanger
1	225,093,501	225,096,376	NA18952	simple	PacBio
1	232,812,220	232,815,483	HG01488	proxdup	Sanger
1	236,753,138	236,757,057	NA19725	invdup	PacBio
1	240,115,049	240,118,193	NA18909	invdup	PacBio
2	1,373,228	1,377,067	HG01198	proxdup	Sanger
2	38,523,662	38,527,434	NA20862	invdup	PacBio
2	51,879,685	51,882,806	NA20869	invdup	PacBio
2	61,699,968	61,704,838	NA19908	invdup	PacBio
2	66,752,907	66,755,431	NA20888	invdup	PacBio
2	72,066,684	72,069,956	NA19462	invdup	PacBio
2	72,437,947	72,443,984	HG01177	invdup	Sanger
2	77,048,694	77,054,125	NA19440	invdup	PacBio
2	88,574,293	88,578,600	NA12842	complex	PacBio
2	100,816,898	100,819,370	NA19130	simple	PacBio
2	125,765,562	125,769,408	HG00260	proxdup	Sanger
2	126,185,839	126,188,629	HG01069	proxdup	Sanger
2	129,683,821	129,687,475	NA11919	invdup	PacBio
2	131,885,572	131,888,371	HG01353	invdup	Sanger
2	133,517,218	133,520,825	NA21104	invdup	PacBio
2	153,458,952	153,462,309	NA20896	invdel	PacBio
2	176,681,072	176,689,113	NA18621	invdup	PacBio
2	183,675,967	183,680,241	HG02282	invdup	PacBio
2	187,155,554	187,158,547	HG00261	invdel	PacBio
2	216,826,684	216,829,198	NA19716	simple	PacBio
2	226,994,897	227,002,679	HG01464	proxdup	Sanger
2	230,842,479	230,846,050	NA21123	invdup	PacBio
3	10,738,894	10,741,904	NA18597	invdup	PacBio
3	12,988,206	12,991,165	HG01191	invdup	Sanger
3	18,655,279	18,658,136	NA19256	complex	PacBio
3	36,405,643	36,408,059	NA20813	simple	PacBio

Chr	Start	End	Sample	SVtype	Source
3	43,834,099	43,837,000	NA12763	invdup	Sanger
3	88,545,013	88,552,419	pooled	proxdup	Illumina
3	104,030,235	104,039,365	HG01710	complex	PacBio
3	122,889,782	122,893,062	NA19663	invdup	PacBio
3	127,495,514	127,498,977	NA18909	invdup	PacBio
3	133,687,260	133,690,830	HG02568	invdup	PacBio
3	139,978,284	139,981,411	NA12750	proxdup	Sanger
3	160,277,732	160,279,909	NA20890	invdup	PacBio
3	165,030,799	165,033,479	NA19701	invdup	PacBio
3	168,104,732	168,106,843	HG02081	invdup	PacBio
3	184,308,171	184,311,964	NA18993	complex	PacBio
3	190,832,762	190,835,088	NA18877	invdup	PacBio
3	196,276,818	196,280,716	NA19375	complex	PacBio
4	16,916,389	16,918,472	HG01610	invdel	PacBio
4	23,127,395	23,133,886	HG01105	proxdup	Sanger
4	26,485,682	26,488,207	NA20766	simple	PacBio
4	45,755,433	45,758,093	HG02508	invdel2	PacBio
4	53,703,955	53,706,001	NA19213	invdup	PacBio
4	62,473,294	62,477,236	NA12717	invdel	PacBio
4	83,860,778	83,864,528	NA19703	invdel2	PacBio
4	101,598,859	101,602,434	NA18498	invdel	PacBio
4	110,134,019	110,138,014	NA18621	simple	PacBio
4	112,006,211	112,009,195	NA20506	invdup	PacBio
4	115,289,789	115,293,034	NA19429	invdup	PacBio
4	117,788,673	117,791,591	NA19652	invdel2	PacBio
4	120,584,771	120,588,111	NA19117	invdup	PacBio
4	138,659,437	138,661,397	NA19380	simple	PacBio
4	148,548,181	148,555,509	NA12155	invdup	PacBio
4	151,163,095	151,167,180	HG00120	proxdup	Sanger
5	10,905,194	10,909,204	NA12878	invdup	PacBio
5	18,083,693	18,087,591	HG01510	invdel	PacBio
5	31,189,246	31,194,091	pooled	invdup	Illumina
5	32,884,289	32,887,629	NA20815	simple	PacBio
5	92,760,283	92,763,153	NA20298	invdup	PacBio
5	95,599,929	95,602,938	HG02008	invdup	PacBio
5	116,049,680	116,052,708	NA19457	invdup	PacBio
5	128,383,831	128,388,265	pooled	invdup	Illumina
5	142,283,542	142,286,903	NA19401	invdup	PacBio
5	143,511,483	143,516,048	NA19429	proxdup	Sanger
5	167,176,735	167,179,485	HG00132	invdup	PacBio
5	171,074,258	171,077,635	NA19404	simple	PacBio
5	171,845,749	171,848,814	NA19334	invdup	PacBio
5	174,875,526	174,878,773	HG03162	invdup	PacBio
5	174,891,963	174,895,522	NA21120	invdel	PacBio
5	178,122,463	178,126,334	NA18552	invdel	PacBio
6	341,808	345,233	HG03052	proxdup	Sanger
6	3,883,917	3,888,364	NA18519	invdel	PacBio
6	12,435,664	12,442,395	pooled	invdup	Illumina
6	22,949,565	22,952,564	HG02178	simple	PacBio
6	24,044,152	24,046,232	NA18528	simple	PacBio
6	32,314,159	32,317,262	HG00102	proxdup	Sanger
6	32,980,439	32,986,191	pooled	proxdup	Illumina
6	41,039,563	41,047,733	pooled	proxdup	Illumina
6	91,942,170	91,951,265	pooled	proxdup	Illumina
6	119,010,711	119,014,924	HG01879	proxdup	Sanger
6	119,539,622	119,543,304	NA12348	simple	PacBio
6	138,060,160	138,062,315	HG00554	invdel	PacBio
7	811,112	813,724	NA19451	invdup	PacBio
7	3,488,490	3,491,579	HG01083	invdup	PacBio
7	8,350,680	8,358,342	NA20845	invdel	PacBio
7	18,724,936	18,730,404	HG02943	invdup	PacBio
7	25,683,658	25,686,303	NA20876	invdup	PacBio
7	30,507,209	30,513,004	pooled	proxdup	Illumina
7	38,233,698	38,238,284	NA19119	invdup	PacBio
7	50,263,289	50,266,539	NA18596	invdup	PacBio
7	53,018,261	53,028,101	NA10847	invdup	Sanger
7	90,062,875	90,071,654	HG03367	invdup	PacBio
7	101,637,967	101,642,114	NA19383	invdel2	PacBio
7	117,008,274	117,015,041	HG00683	invdup	PacBio
7	127,202,018	127,206,098	NA12760	proxdup	Sanger
7	128,817,901	128,821,861	NA19213	invdup	PacBio
7	139,979,709	139,986,651	NA20521	invdel	PacBio

Chr	Start	End	Sample	SVtype	Source
7	151,009,481	151,013,242	NA19716	invdup	PacBio
8	21,307,540	21,312,322	pooled	proxdup	Illumina
8	26,348,722	26,351,695	NA19116	invdup	PacBio
8	43,285,112	43,288,081	HG00565	invdup	PacBio
8	47,500,475	47,503,744	NA12717	proxdup	Sanger
8	80,510,952	80,515,416	NA19393	complex	PacBio
8	87,669,453	87,673,747	NA19099	complex	PacBio
8	100,156,890	100,159,093	NA18596	invdup	PacBio
8	103,891,670	103,894,247	NA19372	simple	PacBio
8	110,096,523	110,100,023	NA20528	invdup	PacBio
8	117,091,416	117,096,195	HG01595	invdup	Sanger
8	122,512,509	122,516,193	NA21089	invdup	PacBio
8	133,531,442	133,535,173	HG01190	proxdup	Sanger
8	135,239,332	135,242,167	NA18621	invdup	PacBio
8	138,129,967	138,132,573	NA12812	invdup	Sanger
8	138,347,078	138,350,618	NA12878	invdup	PacBio
9	15,127,054	15,129,637	HG02946	invdup	PacBio
9	20,075,503	20,081,530	pooled	proxdup	Illumina
9	34,595,327	34,601,065	HG03052	invdup	PacBio
9	38,666,722	38,675,128	pooled	invdup	Illumina
9	79,778,313	79,781,868	NA19314	invdup	PacBio
9	87,133,944	87,139,287	pooled	invdup	Illumina
9	88,741,878	88,745,736	NA19720	invdel	PacBio
9	91,735,559	91,739,219	NA20757	invdup	PacBio
9	94,720,336	94,722,969	NA19466	simple	PacBio
9	125,378,010	125,381,471	NA19307	invdup	PacBio
9	125,484,055	125,492,389	NA19312	invdup	PacBio
9	134,394,191	134,398,761	HG02035	invdup	PacBio
10	62,063,542	62,065,680	NA19684	invdel	PacBio
10	76,503,187	76,506,726	NA19661	invdup	PacBio
10	77,858,287	77,861,697	HG03304	simple	PacBio
10	86,799,703	86,804,886	pooled	proxdup	Illumina
10	87,242,251	87,246,132	NA19920	complex	PacBio
10	97,205,777	97,209,023	NA19397	proxdup	Sanger
10	99,778,760	99,783,080	NA18636	simple	PacBio
10	107,247,393	107,251,563	NA19088	invdup	PacBio
10	115,014,501	115,018,805	pooled	proxdup	Illumina
10	115,170,837	115,172,953	NA19701	simple	PacBio
11	24,902,559	24,905,575	NA19093	invdup	PacBio
11	25,353,203	25,361,606	pooled	proxdup	Illumina
11	66,017,849	66,020,965	NA20852	invdel	PacBio
11	73,475,246	73,484,209	HG03115	invdup	PacBio
11	105,571,092	105,576,377	NA19655	complex	PacBio
11	113,803,414	113,808,661	pooled	proxdup	Illumina
11	131,064,693	131,067,736	NA19114	invdup	PacBio
12	24,294,074	24,296,700	NA19700	simple	PacBio
12	26,988,940	26,992,947	NA19920	simple	PacBio
12	38,316,933	38,320,034	HG01992	invdup	PacBio
12	45,744,570	45,752,672	pooled	proxdup	Illumina
12	51,996,881	52,000,148	NA19152	simple	PacBio
12	71,707,228	71,712,753	NA06994	invdup	PacBio
12	78,382,870	78,389,248	NA12005	invdup	Sanger
12	91,377,082	91,383,619	pooled	invdup	Illumina
12	94,337,361	94,347,985	pooled	proxdup	Illumina
12	95,333,304	95,335,398	HG02017	invdup	PacBio
12	95,951,781	95,954,935	NA19712	invdup	PacBio
12	117,588,989	117,591,059	NA18570	invdup	PacBio
12	121,338,020	121,340,751	HG03367	invdup	PacBio
13	25,801,807	25,804,767	NA18489	invdel	PacBio
13	40,172,270	40,176,559	NA12340	invdup	PacBio
13	44,679,673	44,682,428	NA19062	simple	PacBio
13	74,690,142	74,692,748	NA18870	invdup	PacBio
13	103,199,207	103,204,904	HG01085	invdup	PacBio
14	25,610,038	25,615,609	NA11994	invdel2	PacBio
14	44,689,839	44,692,672	HG00419	simple	PacBio
14	48,323,276	48,325,817	NA19719	invdup	PacBio
14	50,461,714	50,467,254	NA18567	invdup	PacBio
14	54,420,845	54,423,480	NA18559	simple	PacBio
14	59,039,358	59,042,736	NA19457	invdel2	PacBio
14	62,255,009	62,258,710	NA21091	invdup	PacBio
14	65,841,159	65,843,994	NA20525	invdel	PacBio
15	38,482,819	38,490,047	NA18563	invdup	PacBio

Chr	Start	End	Sample	SVtype	Source
15	46,052,726	46,055,399	HG01365	invdup	PacBio
15	68,406,025	68,409,036	HG01051	invdup	Sanger
16	55,917,832	55,921,409	NA19462	invdel	PacBio
16	69,179,664	69,182,488	NA18552	invdel	PacBio
16	69,759,566	69,765,483	pooled	invdup	Illumina
16	78,036,292	78,043,806	pooled	proxdup	Illumina
17	26,112,410	26,115,025	NA18916	simple	PacBio
17	33,180,764	33,183,545	HG01776	invdup	PacBio
17	40,541,067	40,546,253	NA18563	invdel	PacBio
17	46,614,626	46,618,231	HG01334	proxdup	Sanger
17	64,938,572	64,944,951	pooled	proxdup	Illumina
17	76,845,394	76,849,002	NA19319	simple	PacBio
18	13,917,343	13,920,817	NA18879	invdup	PacBio
18	39,837,841	39,841,521	HG00266	invdup	PacBio
18	69,710,794	69,713,885	HG00268	proxdup	Sanger
18	73,931,872	73,939,922	pooled	proxdup	Illumina
19	41,138,443	41,146,725	HG02938	invdup	PacBio
20	2,358,695	2,361,666	HG00246	proxdup	Sanger
20	43,963,732	43,966,725	NA19393	invdup	PacBio
20	58,306,804	58,309,850	NA18944	simple	PacBio
21	20,646,874	20,649,790	NA11832	complex	PacBio
21	43,343,620	43,352,881	NA18635	simple	PacBio
X	18,606,738	18,611,680	HG01625	invdup	PacBio

Supplementary Information to Chapter 3

C

The content of this supplementary chapter is largly taken from the supplementary methods of our unpublished manuscript and adapted when neccessary.

C.1. Commentary on cited literature

Comments on literature used in Sections 3.1.2 and 3.1.3.

- Sexton et al. [2012] were the first ones discovering TADs in *D. melanogaster*.
- Rao et al. [2014] developed an *in situ* Hi-C protocol and produced the (presumably to date) densest Hi-C contact map with a resolution of 1 kb, which led to a discovery of smaller TADs and a striking correlation of TAD boundaries with convergent CTCF motifs.
- Le Dily et al. [2014] treated breast cancer cells with hormones and observed coordinated activation or suppression of genes within the same TAD.
- Nora et al. [2012] investigated a 4.5 Mb region on chromosome X using 5C and super-resolution microscopy including several TADs. They saw a higher correlation of gene expression within the same TAD than between TADs (figure 4b). Also TADs aligned with H3K27me3 or H3K9me2 blocks (figure 2). Moreover, by studying an additional mouse line with TAD boundary deletion between the *Xist* and *Tsix* loci, they performed a first perturbation experiment showing that new TADs can be created this way.
- Dekker et al. [2015] reviewed the existence of TAD-like structures in a variety of organisms, including mammals, *Drosophila*, and *S. pombe* (see figure 1).
- Pope et al. [2014] found that "replication domain boundaries share a near one-to-one correlation with TAD boundaries".
- Shen et al. [2012] defined pairs of enhancers and promoters with correlated activity based on chromatin states and polymerase II occupancy and found that these units correlated with TADs.

- Schmitt et al. [2016] produced Hi-C maps of 21 primary human tissues and cell types and found a high conervation of TAD boundaries (figure 1 E).
- Lettice et al. [2011] reported a long-range *cis*-regulatory mutation in which ectopic gene expression arises through mis-regulation by a juxtaposed enhancer element at the *shh* locus. They were the first ones to term this mechanisms "enhancer adoption".
- Guo et al. [2015] CRISPR-engineered an inversion of a CTCF binding sites at a TAD boundary in mice. They observed a profound change in promoter-enhancer interactions across the boudary with consequential change in gene expression.

C.2. SNV calling

Both WGS and mate pair sequencing data was mapped to DM6 using BWA MEM version 0.7.15. SNV and short indel calling was performed using FREEBAYES version v0.9.21-19 with disabled population priors on the WGS data of both F_0 and F_1 samples simultaneously. The results were filtered using VCFLIB based on a quality value of at least 30, a minimum of at least two reads carrying the allele to the right and to the left end, and on the fact that the allele was seen on at least two reads mapping in each direction. We further normalized variants, removed mutli-allelic variants, and decomposed multi-nucleotide substitutions (which are reported as haplotype blocks by FREEBAYES) into SNVs using VT [Tan et al., 2015] (the sub-command DECOMPOSE_BLOCKSUB was used for decomposition). We finally remove contigs other than chromosome 2, 3, and X and obtained a total of 860,095 SNVs and small indels.

C.3. Mutational signature analysis

Starting from the set of 520,521 balancer- or wild type-specific SNVs, I removed the ones which are present in the DGRP freeze 2.0 SNV call set. Then I used the R package SOMATICSIGNATURES [Gehring et al., 2015] to count base substitutions and their contexts of the remaining 58,457 variants and plotted their relative frequencies in Figure C.1. The absence of striking differences between balancer and wild type spectra demotivated me from deeper investigations of mutational signatures.

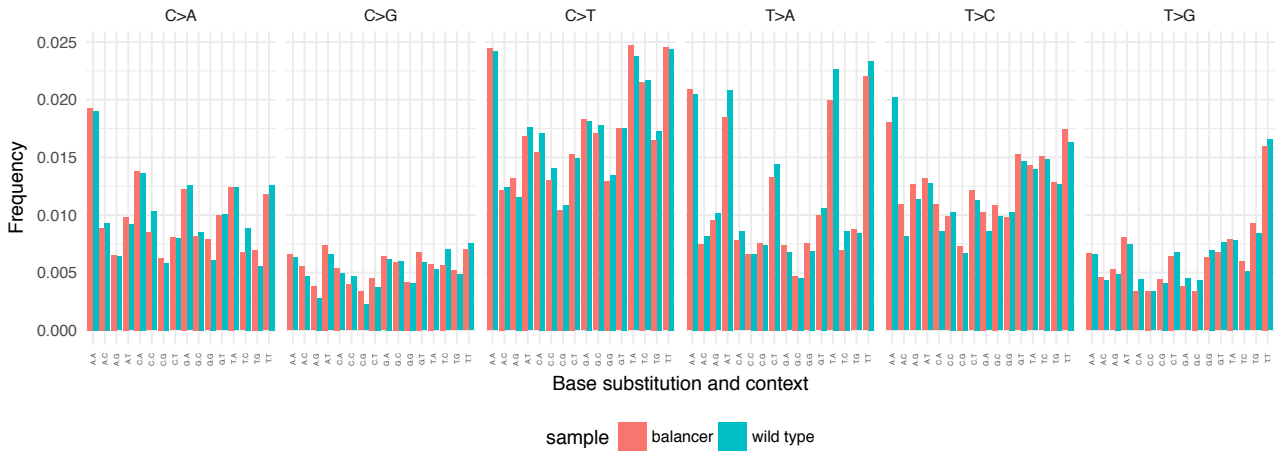


Figure C.1: SNV mutation spectrum. Frequency of the different base substitutions in their three-nucleotide context for balancer- and wild type-specific SNVs. SNVs that are found in DGRP were removed, leaving 58,457 variants.

C.4. Deletion calling

I used Delly version 0.7.2 on the WGS data of the F_0 and F_1 data simultaneously and applied an extensive filtering procedure to reduce the number of false positive calls. From the initial 10,421 deletion calls, 5,150 dropped out because they were not flagged as "QC PASS", were not on one of the main chromosomes (Chr2, Chr3, or ChrX), had a mapping quality value of less than 60 or did not match the expected genotypes (i.e. balancer-specific, wild type-specific, and common - together constituting more than 90% of the calls). Furthermore I required a minimum number of supporting read pairs for reference and alternative allele combined, namely 40 read pairs for "imprecise" Delly calls and 25 split reads for breakpoint-precise Delly calls.

Next, I developed a dynamic read depth ratio filter that was applied to deletion predictions of 160 bp or larger. To this end, the read count within the predicted deletion was normalized by the summed read count in size-matched intervals flanking the locus and these values were compared between samples. I required a minimum difference in the read depth ratio between samples with different genotypes and this threshold increases dynamically with SV size. This is motivated by the fact that for larger deletions the average read depth signal is more robust against local fluctuations in coverage. To give an example, this filter removed a number of predictions above 100 kb in size, which could be clearly identified as false positives by inspecting additional (e.g. Hi-C) data. At last I overlapped deletions with a mappability map to classify them into high (at least 50% in a uniquely mappable region) or low-confidence loci. Eventually we obtained four call sets: 3,072 calls with high-confidence and below 50 bp, 737 calls with high confidence and from 50 - 159 bp, 395 large calls with high confidence and 75 large ones with low

confidence.

As a validation Yad Ghavi-Helm performed PCR on randomly selected loci in the latter three categories. I designed primers using a lab-internal extension to Primer3 [Untergasser et al., 2012] and Yad Ghavi-Helm amplified 25 loci per category in both samples via PCR. In the size range 50-159 bp 24 out of 25 loci validated, also 24/25 loci validated for high confidence calls of 160 bp, and 25/25 loci validated for low-confidence calls, yielding an estimated FDR of 2.66%. At last we merged the set of Delly deletion calls into the set of small deletions called by FreeBayes and chose a lower size cutoff of 15 bp. During the merging process FreeBayes calls were given priority over matching Delly calls (based on 50% reciprocal overlap). The final data set (referred to as "deletions" in the main text) contains 8,340 deletions on chromosomes 2, 3 and X.

C.5. Duplication calling and filtering

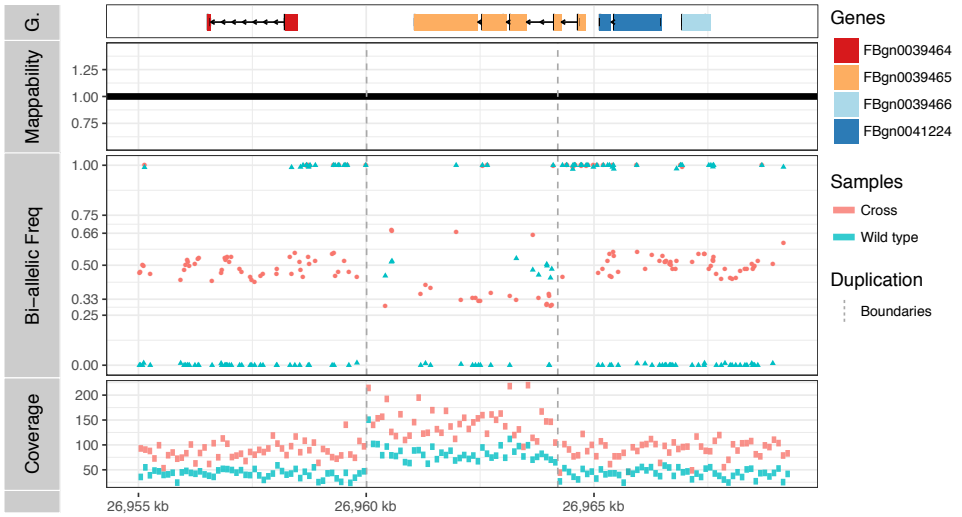


Figure C.2: Example of a wild type-specific duplication. Wild-type specific tandem duplication at locus *chr3R:26,960,008–26,964,205*. The BAF signal clearly suggests a heterozygous duplication in the F_1 cross and the increased read depth in the wild type sample identifies it as present on the wild type chromosome.

I used Delly version 0.7.5 in tandem duplication mode and supplied both mate pair and WGS libraries for F_0 and F_1 samples simultaneously. Duplication calls were initially filtered by the "quality PASS" criteria reported by Delly and by their combined genotypes, which were required to be heterozygous in the F_1 sample. We do not require homozygosity in the F_0 sample due to a known issue of the duplication classifier, which reports many homozygous tandem duplications as heterozygous. For all remaining 352 calls I generated detailed overview plots that contained multiple lines of information: a

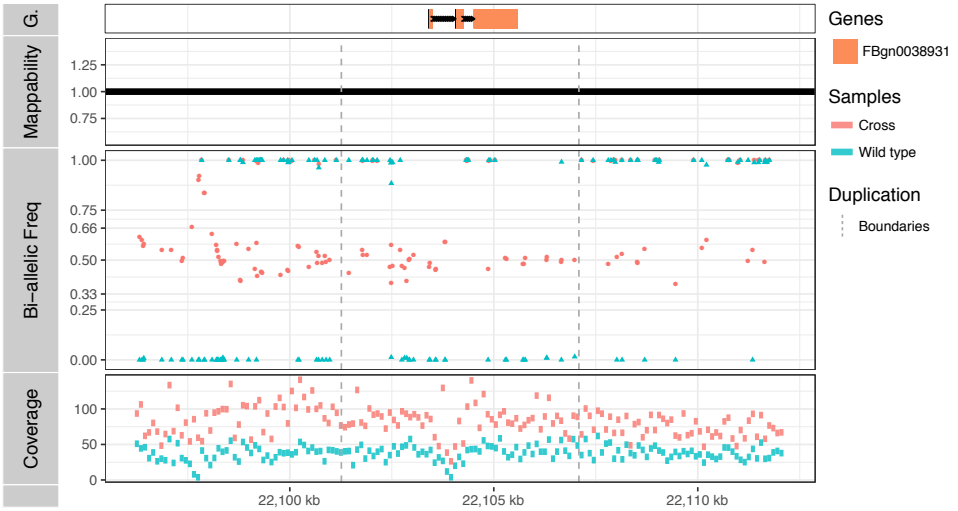


Figure C.3: Example of a false duplication prediction. Predicted duplication locus *chr3R:22,101,264–22,107,086* is not validated by the BAF signal.

total read-depth track, a mappability track, overlapping gene annotations and, importantly, a track showing the BAF measured at SNV positions. These plots allowed me to sort out false positives, leaving 122 manually curated high-quality tandem duplications. Aside from tandem duplications I further inspected the BAF ratio across the genomes and unraveled three non-tandem duplications of 4.3 kb, 10.4 kb and 258 kb size. Both sets together are summarized by "duplications" in the main text.

C.6. Breakpoints of the balancer chromosomes

Bal	Chr	5′ bp.	3′ bp.	Del/Dup	Genes affected
CyO	2L	2,137,075	2,137,067	+9	GlyP: 3′ UTR
	2L	12,704,657	12,704,649	+9	nAChRalpha6: intron
	2L	9,805,567	9,805,575	-7	CG5776, spict
	2R	14,067,771	14,067,782	-10	Src42A: 3′ UTR
	2R	6,012,739	6,012,459	+281	Prosap: intron
	2R	21,971,918	21,972,072	-153	-
TM3	3L	6,925,034	6,926,125	-1,090	-
	3R	9,943,831	9,944,040	-208	Glut4EF: exon/3′ UTR
	3L	15,150,269	15,150,272	-2	FucTA: intron
	3R	23,050,763	23,050,764	0	p53: intron
	3L	19,386,273	19,388,151	-1,877	GC32206: intron/exon
	3R	20,637,930	20,637,930	+1	Lrrk: exon
	3L	22,637,876	22,637,952	-75	CG14459: 3′ UTR
	3R	31,653,695	31,653,707	-11	kek6: 5′ UTR
	3R	*20,308,200*	*20,325,700*	*-17,500*	CG42668, CG42668
	3R	-	-	-	unknown

Table C.1: Breakpoint positions of balancer chromosomes. Breakpoint positions of the major rearrangements of both balancer chromosomes (Bal) were identified in Hi-C data and refined based on their paired-end signature uncovered by DELLY. Values in italic could not be resolved at the base pair level. Between the 3′ and 5′ ends of the breakpoints deletions or duplications can have occurred: Deletions are states as negative integers, duplications as positive integers (Del/Dup).

C.7. Detail on ASE detection using DESeq2

The main idea is explained in Section 3.4.2. Haplotype-resolved fragment counts were calculated using HTSEQ-COUNT [Anders et al., 2015]. In the main analysis we then tested genes for ASE by inserting these haplotype-specific counts of all four replicates (2x N_1^{mat}, 2x N_1^{pat}) into a matrix and supplying it to DESEQ2. Genes were filtered for a minimum number of reads (average of 50 fragments per gene per sample) and by chromosome (only 2L, 2R, 3L, and 3R were considered). DESEQ2 was tested with a design `~Replicate + Haplotype`. The resulting p-values were corrected using FDRTOOL [Strimmer, 2008].

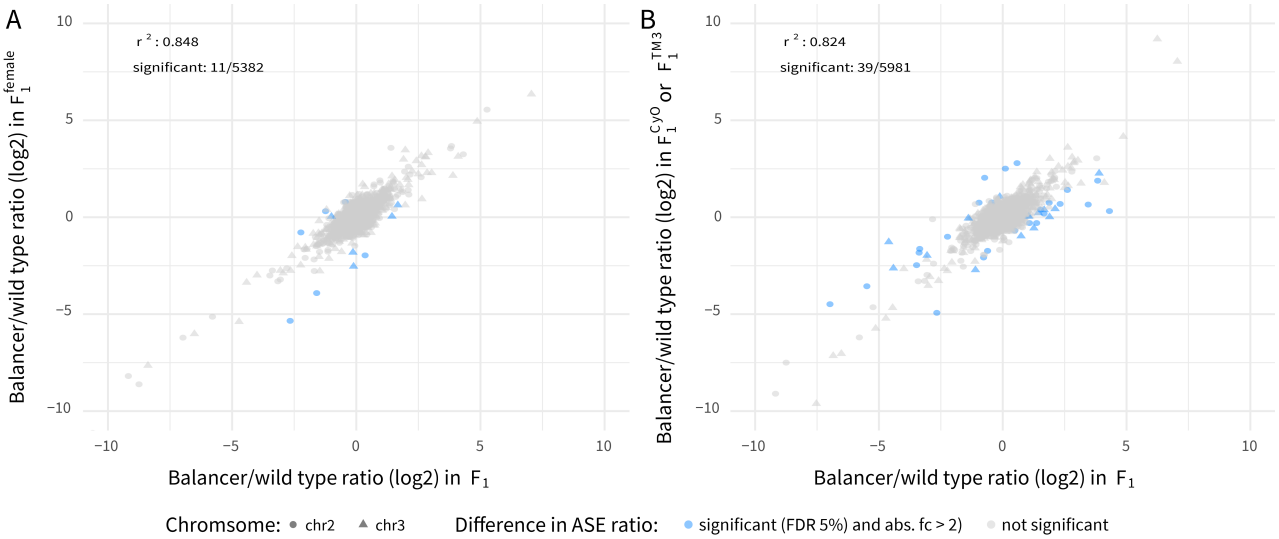


Figure C.4: Changes in ASE signal depending on the genetic background. **A:** Balancer-to-wild type ratio (log scale) of gene expression of genes on chromosomes 2 and 3 in two different F_1 samples, namely adult F_1^f, which was sequenced for Figure 3.11, and F_1, which was sequenced under different conditions. Both data sets are highly correlated (Pearson correlation of $r^2 = 0.84$), yet 0.2% of genes significantly differ between them (FDR 5%, fold change $\geqslant 2$). **B:** Balancer-to-wild type ratio (log scale) of gene expression of two different adult samples, namely F_1 on the x-axis and F_1^{CyO} (for chromosome 2) or F_1^{TM3} (chromosome 3) on the y-axis. Pearson correlation across both chromosomes is 0.824, yet for 0.65% of genes the balancer/wild type ratio significantly differs (FDR 5%, fold change $\geqslant 2$).

C.8. Further ASE-related analyses

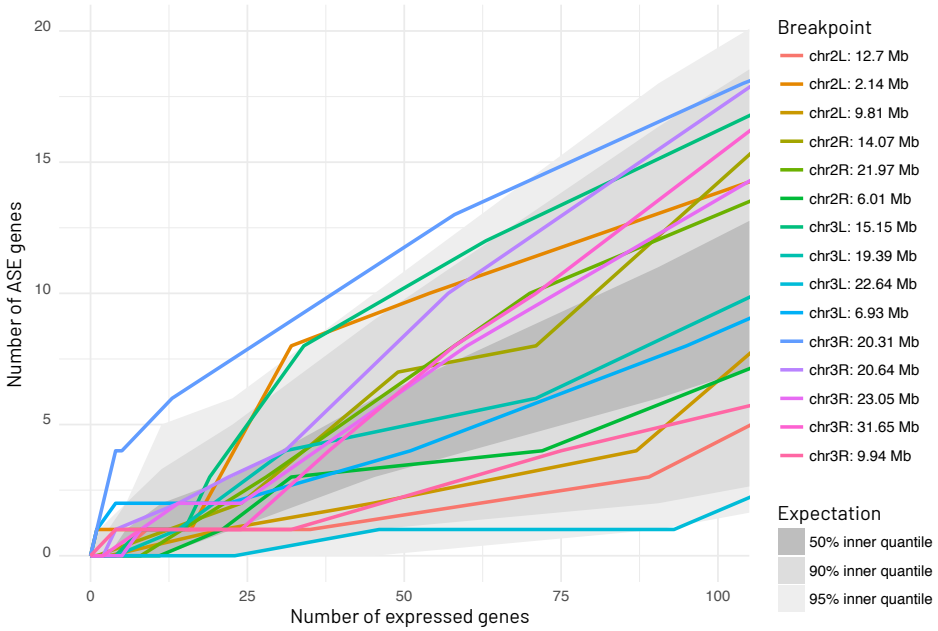


Figure C.5: Number of ASE genes around the breakpoints. This plot shows the number of significant ASE genes at given distances to one of the large rearrangments, or to random positions in the genome (quantiles of 500 random samplings are shown). Genes directly affected by breakpoints were removed.

C.9. Mobile-element analysis

Table C.2: List of identified mobile element insertions. List of insertions of mobile elements that were successfully identified as described in Section 3.5.4 and correspond to the start points of RNA-seq signal measured for the respective genes. +/-: orientation of transcription of the gene. *lfc*: log fold change of gene expression balancer/wild type. *TE family*: Most likely family of transposable element. *MEI pos*: Position of MEI relative to the gene (upstream/downstream) as well as orientation of the insertion of the MEI.

Gene ID	+/-	lfc.	TE family	MEI pos
FBgn0265959	+	5.42	gb/AY180917/roo	downstream, +
FBgn0034085	-	5.10	gb/AY180917/roo	downstream, -
FBgn0051116	-	4.91	gb/AY180917/roo	downstream, -
FBgn0015038	-	4.59	gb/X02599/copia	downstream, -
FBgn0047351	-	3.76	gb/AY180917/roo	downstream, -
FBgn0031414	+	3.74	gb/AY180917/roo	downstream, +
FBgn0052985	+	3.63	gb/V00246/FB	downstream, ?
FBgn0036224	+	3.20	gb/AY180917/roo	upstream, +
FBgn0051164	+	3.09	gb/AY180917/roo	downstream, +

FBgn0050044	+	2.80	gb/AY180917/r00	upstream, +
FBgn0033469	+	2.66	gb/AY180917/r00	upstream, +
FBgn0024150	-	2.41	gb/AY180917/r00	downstream, -
FBgn0027552	-	2.39	gb/X01472/17.6	upstream, -
FBgn0043005	-	2.34	gb/AY180917/r00	upstream, -
FBgn0038784	-	2.19	gb/AC005453/1360	upstream, -
FBgn0031219	-	2.11	gb/AY180917/r00	upstream, -
FBgn0083141	+	2.01	gb/AY180917/r00	upstream, +
FBgn0039104	-	2.01	gb/AY180917/r00	upstream, -
FBgn0031220	-	2.00	gb/AY180917/r00	upstream, -
FBgn0000527	-	2.00	gb/nnnnnnnn/412	downstream, -
FBgn0031631	-	0.95	gb/AY180917/r00	upstream, -
FBgn0039613	+	-1.25	gb/AY180917/r00	upstream, +
FBgn0036454	+	-1.32	gb/X03431/297	downstream, +
FBgn0033792	+	-1.39	gb/AY180917/r00	upstream, +
FBgn0265851	+	-1.44	gb/AY180917/r00	upstream, +
FBgn0267936	+	-1.82	gb/X02599/copia	upstream, +
FBgn0035187	-	-1.88	gb/AY180917/r00	upstream, -
FBgn0259219	+	-1.98	gb/X02599/copia	upstream, +
FBgn0250910	-	-2.10	gb/V00246/FB	downstream, ?
FBgn0026592	+	-2.44	gb/AY180917/r00	upstream, +
FBgn0266782	+	-2.60	gb/AY180917/r00	upstream, +
FBgn0011230	-	-2.78	gb/AY180917/r00	upstream, +
FBgn0035610	-	-2.84	gb/AY180917/r00	downstream, -
FBgn0267130	+	-2.81	gb/AY180917/r00	upstream, ?
			gb/AY180917/r00	downstream, +
FBgn0034467	-	-3.78	gb/AY180917/r00	upstream, -
FBgn0262020	+	-3.88	gb/AY180917/r00	upstream, +
FBgn0035638	-	-6.14	gb/X02599/copia	upstream, -

C.10. Integrated visualization

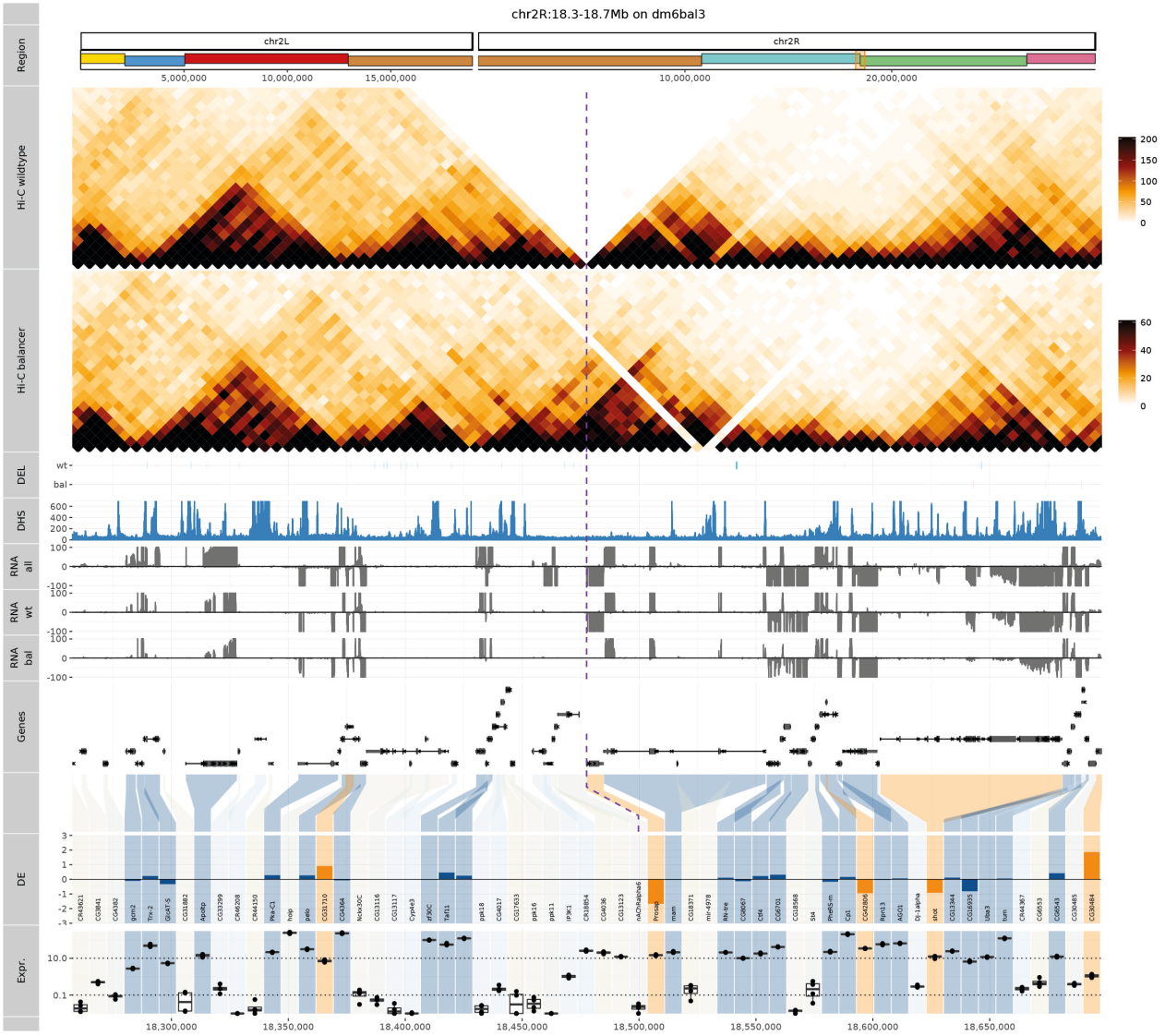


Figure C.6: Integrated visualization around breakpoint *2R:14.1 Mb*: balancer chromosome assembly (I). This figure belongs to Figure 3.17 and is explained in Section 3.7. Here, the locus belonging to the **green** genomic region is shown in respect to a custom reference genome that represents the genomic order of the balancer chromosomes. Notably, the gap in Hi-C contacts are now present in the wild type track.

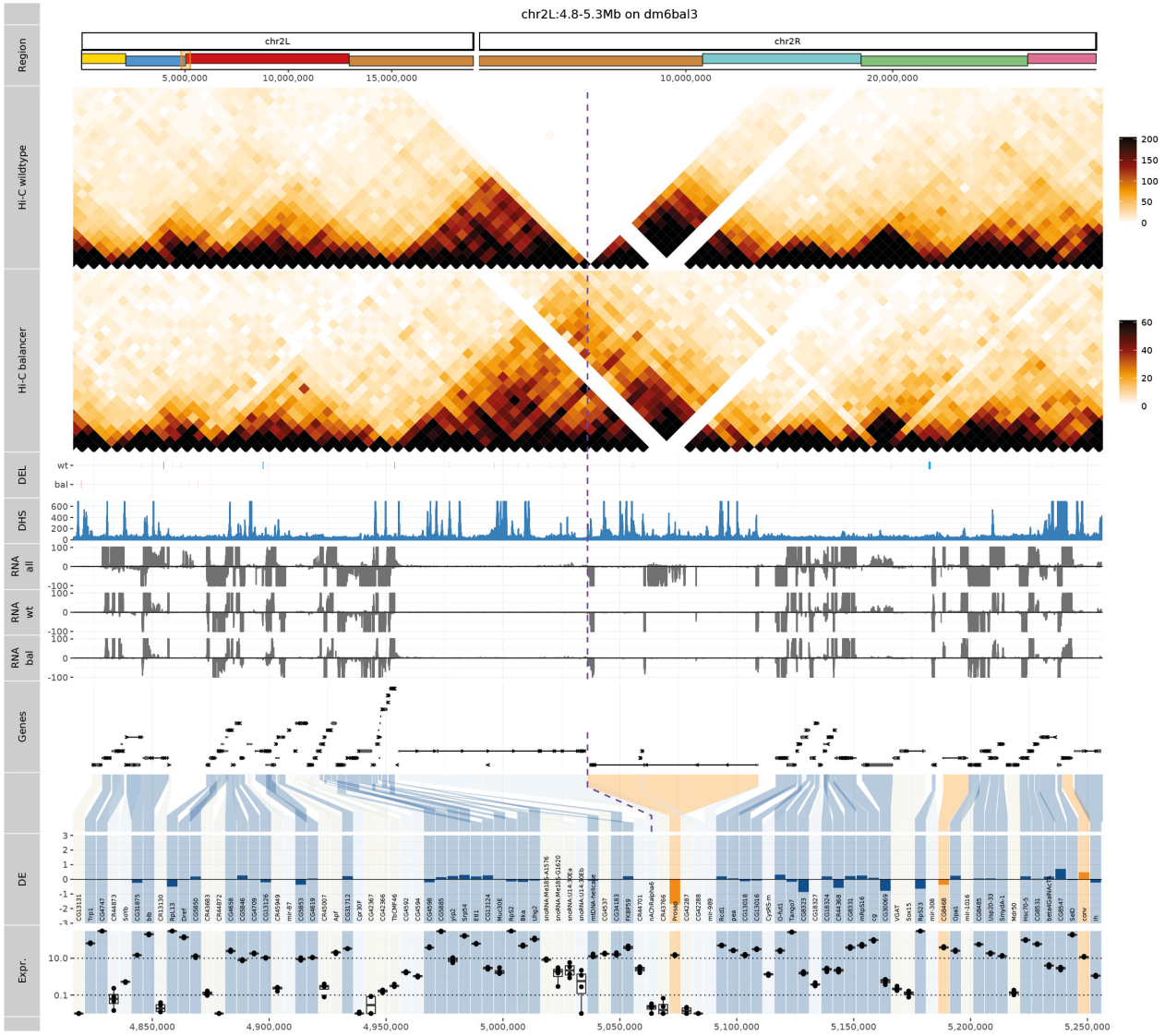


Figure C.7: Integrated visualization around breakpoint *2R:14.1 Mb*: balancer chromosome assembly (II). This figure belongs to Figure 3.17 and is explained in Section 3.7. Here, the locus belonging to the **red** genomic region is shown in respect to a custom reference genome that represents the genomic order of the balancer chromosomes. Notably, the gap in Hi-C contacts are now visible in the wild type track.

C.11. Hi-C matrix generation (Aleksander Jankowski)

Following the approach of Ramírez et al. [2018], sequencing reads obtained from the Hi-C experiment were mapped to DM6 using BWA MEM with parameters `-E50 -L0`. We performed read mapping separately for both reads of a pair to not bias the alignment towards an assumed paired-end distance or orientation. Read pairs were further processed using PAIRSAMTOOLS[1] to select only valid Hi-C molecules formed via a single ligation event. To remove PCR duplicates, we used PAIRSAMTOOLS `dedup` with option `--max-mismatch 0` to keep only a random read pair among the pairs with both reads identically mapped. Afterwards, reads were separated according to their haplotype annotations as described in Section 3.4.2. Finally, for reads with chimeric alignments only the alignment positioned at the 5′ end was retained. The effect of filtering and read separation is shown in Table C.3. Reads assigned to the balancer haplotype were lifted to the balancer pseudoassembly using CROSSMAP [H. Zhao et al., 2014]. The reads were then counted in 5 kb bins using `hicBuildMatrix` from HICEXPLORER with default parameters [Ramírez et al., 2018] to obtain contact frequency matrices per haplotype. We observed an exponential distance decay that is expected from Hi-C experiments. Finally, we applied intrinsic matrix balancing normalization using `hicCorrectMatrix` on these matrices.

Steps of data processing	Replicate 1	Replicate 2	Fraction
Total read pairs	2,261,026,305	3,742,127,293	100.0%
After duplicate removal	1,377,657,560	1,638,115,020	50.2%
Wild type haplotype	352,799,755	528,189,753	14.7%
—filtered	117,407,775	183,917,734	5.0%
Balancer haplotype	113,011,608	166,060,988	4.6%
—filtered	35,489,102	55,135,464	1.5%

Table C.3: Number reads read pairs during Hi-C analysis. Number of read pairs avaialble for a Hi-C contact map drastically reduce due to haplotype separation and filtering, as explained in Appendix C.11.

[1] `http://pairsamtools.readthedocs.io/`

Bibliography

1000 Genomes Project Consortium, The (2010). "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319, pp. 1061–1073. ISSN: 0028-0836. DOI: 10.1038/nature09534.

– (2012). "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422, pp. 56–65. ISSN: 0028-0836.

– (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74. ISSN: 0028-0836. DOI: 10.1038/nature15393.

Abyzov, Alexej, Shantao Li, Daniel Rhee Kim, Marghoob Mohiyuddin, Adrian M Stütz, Nicholas F Parrish, Xinmeng Jasmine Mu, Wyatt Clark, Ken Chen, Matthew Hurles, Jan O Korbel, Hugo Y K Lam, Charles Lee, and Mark B Gerstein (2015). "Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms." In: *Nature communications* 6, p. 7256. ISSN: 2041-1723. DOI: 10.1038/ncomms8256.

Abyzov, Alexej, Jessica Mariani, Dean Palejev, Ying Zhang, Michael Seamus Haney, Livia Tomasini, Anthony F. Ferrandino, Lior A. Rosenberg Belmaker, Anna Szekely, Michael Wilson, Arif Kocabas, Nathaniel E. Calixto, Elena L. Grigorenko, Anita Huttner, Katarzyna Chawarska, Sherman Weissman, Alexander Eckehart Urban, Mark Gerstein, and Flora M. Vaccarino (2012). "Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells". In: *Nature* 492.7429, pp. 438–442. ISSN: 0028-0836. DOI: 10.1038/nature11629.

Abyzov, Alexej, Alexander E. Urban, Michael Snyder, and Mark Gerstein (2011). "CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing". In: *Genome Research* 21, pp. 974–984. ISSN: 10889051. DOI: 10.1101/gr.114876.110.

Alkan, Can, Bradley P Coe, and Evan E Eichler (2011). "Genome structural variation discovery and genotyping." In: *Nature reviews. Genetics* 12.5, pp. 363–76. ISSN: 1471-0064. DOI: 10.1038/nrg2958.

Alkan, Can, Jeffrey M Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O Kitzman, Carl Baker, Maika Malig, Onur Mutlu, S Cenk Sahinalp, Richard A Gibbs, and Evan E Eichler (2009). "Personalized copy number and segmental duplication maps using next-generation sequencing". In: *Nature Genetics* 41.10, pp. 1061–1067. ISSN: 1061-4036. DOI: 10.1038/ng.437.

Alkan, Can, Saba Sajjadian, and Evan E Eichler (2011). "Limitations of next-generation genome sequence assembly". In: *Nature Methods* 8.1, pp. 61–65. ISSN: 1548-7091. DOI: 10.1038/nmeth.1527.

Anders, S., P. T. Pyl, and W. Huber (2015). "HTSeq–a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2, pp. 166–169. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu638.

Araye, Quenta and Kyoichi Sawamura (2013). "Genetic decay of balancer chromosomes in Drosophila melanogaster". In: *Fly* 7.3, pp. 184–186. ISSN: 1933-6934. DOI: 10.4161/fly.24466.

"Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species" (2013). In: *GigaScience* 2.1, p. 10. ISSN: 2047-217X. DOI: 10.1186/2047-217X-2-10.

Ay, Ferhat, Thanh H Vu, Michael J Zeitz, Nelle Varoquaux, Jan E Carette, Jean-Philippe Vert, Andrew R Hoffman, and William S Noble (2015). "Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C". In: *BMC Genomics* 16.1, p. 121. ISSN: 1471-2164. DOI: 10.1186/s12864-015-1236-7.

Bakker, Bjorn, Aaron Taudt, Mirjam E. Belderbos, David Porubsky, Diana C. J. Spierings, Tristan V. de Jong, Nancy Halsema, Hinke G. Kazemier, Karina Hoekstra-Wakker, Allan Bradley, Eveline S. J. M. de Bont, Anke van den Berg, Victor Guryev, Peter M. Lansdorp, Maria Colomé-Tatché, and Floris Foijer (2016). "Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies". In: *Genome Biology* 17.1, p. 115. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0971-7.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner (2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing". In: *Journal of Computational Biology* 19.5, pp. 455–477. ISSN: 1066-5277. DOI: 10.1089/cmb.2012.0021. URL: http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0021.

Bansal, Vikas, Ali Bashir, and Vineet Bafna (2007). "Evidence for large inversion polymorphisms in the human genome from HapMap data." In: *Genome research* 17.2, pp. 219–30. ISSN: 1088-9051. DOI: 10.1101/gr.5774507.

Barbe, D.F. (1975). "Imaging devices using the charge-coupled concept". In: *Proceedings of the IEEE* 63.1, pp. 38–67. ISSN: 0018-9219. DOI: 10.1109/PROC.1975.9707.

Barutcu, A. Rasim, Bryan R. Lajoie, Rachel P. McCord, Coralee E. Tye, Deli Hong, Terri L. Messier, Gillian Browne, Andre J. van Wijnen, Jane B. Lian, Janet L. Stein, Job Dekker, Anthony N. Imbalzano, and Gary S. Stein (2015). "Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells". In: *Genome Biology* 16.1, p. 214. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0768-0.

Bauman, J G, J Wiegant, P Borst, and P van Duijn (1980). "A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA." In: *Experimental cell research* 128.2, pp. 485–90. ISSN: 0014-4827.

Behjati, Sam, Gunes Gundem, David C. Wedge, Nicola D. Roberts, Patrick S. Tarpey, Susanna L. Cooke, Peter Van Loo, Ludmil B. Alexandrov, Manasa Ramakrishna, Helen Davies, Serena Nik-Zainal, Claire Hardy, Calli Latimer, Keiran M. Raine, Lucy Stebbings,

Andy Menzies, David Jones, Rebecca Shepherd, Adam P. Butler, Jon W. Teague, Mette Jorgensen, Bhavisha Khatri, Nischalan Pillay, Adam Shlien, P. Andrew Futreal, Christophe Badie, Colin S. Cooper, Rosalind A. Eeles, Douglas Easton, Christopher Foster, David E. Neal, Daniel S. Brewer, Freddie Hamdy, Yong-Jie Lu, Andrew G. Lynch, Charlie E. Massi, Anthony Ng, Hayley C. Whitaker, Yongwei Yu, Hongwei Zhang, Elizabeth Bancroft, Dan Berney, Niedzica Camacho, Cathy Corbishley, Tokhir Dadaev, Nening Dennis, Tim Dudderidge, Sandra Edwards, Cyril Fisher, Jilur Ghori, Vincent J. Gnanapragasam, Christopher Greenman, Steve Hawkins, Steven Hazell, Will Howat, Katalin Karaszi, Jonathan Kay, Zsofia Kote-Jarai, Barbara Kremeyer, Pardeep Kumar, Adam Lambert, Daniel Leongamornlert, Naomi Livni, Hayley Luxton, Lucy Matthews, Erik Mayer, Susan Merson, David Nicol, Christopher Ogden, Sarah O?Meara, Gill Pelvender, Nimish C. Shah, Simon Tavare, Sarah Thomas, Alan Thompson, Claire Verrill, Anne Warren, Jorge Zamora, Ultan McDermott, G. Steven Bova, Andrea L. Richardson, Adrienne M. Flanagan, Michael R. Stratton, and Peter J. Campbell (2016). "Mutational signatures of ionizing radiation in second malignancies". In: *Nature Communications* 7, p. 12605. ISSN: 2041-1723. DOI: 10.1038/ncomms12605.

Beroukhim, Rameen, Craig H. Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S. Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T. Mc Henry, Reid M. Pinchback, Azra H. Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S. Lawrence, Barbara A. Weir, Kumiko E. Tanaka, Derek Y. Chiang, Adam J. Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J. Kaye, Hidefumi Sasaki, Joel E. Tepper, Jonathan A. Fletcher, Josep Tabernero, José Baselga, Ming-Sound Tsao, Francesca Demichelis, Mark A. Rubin, Pasi A. Janne, Mark J. Daly, Carmelo Nucera, Ross L. Levine, Benjamin L. Ebert, Stacey Gabriel, Anil K. Rustgi, Cristina R. Antonescu, Marc Ladanyi, Anthony Letai, Levi A. Garraway, Massimo Loda, David G. Beer, Lawrence D. True, Aikou Okamoto, Scott L. Pomeroy, Samuel Singer, Todd R. Golub, Eric S. Lander, Gad Getz, William R. Sellers, and Matthew Meyerson (2010). "The landscape of somatic copy-number alteration across human cancers". In: *Nature* 463.7283, pp. 899–905. ISSN: 0028-0836. DOI: 10.1038/nature08822.

Bickhart, Derek M. and George E. Liu (2014). "The challenges and importance of structural variation detection in livestock". In: *Frontiers in Genetics* 5. ISSN: 1664-8021. DOI: 10.3389/fgene.2014.00037.

Bleakley, Kevin and Jean-Philippe Vert (2011). *The group fused Lasso for multiple change-point detection*. Tech. rep., pp. 1–25. arXiv: 1106.4199v1. URL: https://hal.archives-ouvertes.fr/hal-00602121/document.

Braslavsky, I., B. Hebert, E. Kartalov, and S. R. Quake (2003). "Sequence information can be obtained from single DNA molecules". In: *Proceedings of the National Academy of Sciences* 100.7, pp. 3960–3964. ISSN: 0027-8424. DOI: 10.1073/pnas.0230489100.

Browning, Sharon R. and Brian L. Browning (2011). "Haplotype phasing: existing methods and new developments". In: *Nature Reviews Genetics* 12.10, pp. 703–714. ISSN: 1471-0056. DOI: 10.1038/nrg3054.

Burns, Kathleen H. (2017). "Transposable elements in cancer". In: *Nature Reviews Cancer* 17.7, pp. 415–424. ISSN: 14741768. DOI: 10.1038/nrc.2017.35.

Burton, Joshua N, Andrew Adey, Rupali P Patwardhan, Ruolan Qiu, Jacob O Kitzman, and Jay Shendure (2013). "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions." In: *Nature biotechnology* 31.12, pp. 1119–25. ISSN: 1546-1696. DOI: 10.1038/nbt.2727.

Butler, T. Z., M. Pavlenok, I. M. Derrington, M. Niederweis, and J. H. Gundlach (2008). "Single-molecule DNA detection with an engineered MspA protein nanopore". In: *Proceedings of the National Academy of Sciences* 105.52, pp. 20647–20652. ISSN: 0027-8424. DOI: 10.1073/pnas.0807514106.

Cai, Xuyu, Gilad D. Evrony, Hillel S. Lehmann, Princess C. Elhosary, Bhaven K. Mehta, Annapurna Poduri, and Christopher A. Walsh (2014). "Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain". In: *Cell Reports* 8.5, pp. 1280–1289. ISSN: 22111247. DOI: 10.1016/j.celrep.2014.07.043.

Campbell, Ian M., Chad A. Shaw, Pawel Stankiewicz, and James R. Lupski (2015). "Somatic mosaicism: implications for disease and transmission genetics". In: *Trends in Genetics* 31.7, pp. 382–392. ISSN: 01689525. DOI: 10.1016/j.tig.2015.03.013.

Campbell, Lauren L and Kornelia Polyak (2007). "Breast tumor heterogeneity: cancer stem cells or clonal evolution?" In: *Cell cycle* 6.19, pp. 2332–8. ISSN: 1551-4005. DOI: 10.4161/cc.6.19.4914.

Campbell, Peter J, Gad Getz, Joshua M. Stuart, Jan O. Korbel, Lincoln D. Stein, and The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network (2017). In: *bioRxiv*. DOI: 10.1101/162784.

Campbell, Peter J, Philip J Stephens, Erin D Pleasance, Sarah O'Meara, Heng Li, Thomas Santarius, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, Claire Hardy, Jon W Teague, Andrew Menzies, Ian Goodhead, Daniel J Turner, Christopher M Clee, Michael A Quail, Antony Cox, Clive Brown, Richard Durbin, Matthew E Hurles, Paul A W Edwards, Graham R Bignell, Michael R Stratton, and P Andrew Futreal (2008). "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing". In: *Nature Genetics* 40.6, pp. 722–729. ISSN: 1061-4036. DOI: 10.1038/ng.128.

Carter, Scott L, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz (2012). "Absolute quantification of somatic DNA alterations in human cancer". In: *Nature Biotechnology* 30.5, pp. 413–421. ISSN: 1087-0156. DOI: 10.1038/nbt.2203.

Carvalho, Claudia M. B. and James R. Lupski (2016). "Mechanisms underlying structural variant formation in genomic disorders". In: *Nature Reviews Genetics* 17.4, pp. 224–238. ISSN: 1471-0056. DOI: 10.1038/nrg.2015.25.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen (2015). "Tools and best practices for data processing in allelic expression analysis". In: *Genome Biology* 16.1, p. 195. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0762-6.

Celniker, Susan E., Laura A. L. Dillon, Mark B. Gerstein, Kristin C. Gunsalus, Steven Henikoff, Gary H. Karpen, Manolis Kellis, Eric C. Lai, Jason D. Lieb, David M. MacAlpine, Gos Micklem, Fabio Piano, Michael Snyder, Lincoln Stein, Kevin P. White, and Robert H. Waterston (2009). "Unlocking the secrets of the genome". In: *Nature* 459.7249, pp. 927–930. ISSN: 0028-0836. DOI: 10.1038/459927a.

Chaisson, Mark J P, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M. Landolin, John A Stamatoyannopoulos, Michael W Hunkapiller, Jonas Korlach, and Evan E Eichler (2014). "Resolving the complexity of the human genome using single-molecule sequencing". In: *Nature*. ISSN: 0028-0836. DOI: 10.1038/nature13907.

Chaisson, Mark J P, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar Rodriguez, Li Guo, Ryan L Collins, Xian Fan, Jia Wen, Robert E Handsaker, Susan Fairley, Zev N Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M Wenger, Alex Hastie, Danny Antaki, Peter Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T Chuang, Deanna M Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, Gorkin David, Madhusudan Gujral, Victor Guryev, William Haynes-Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Patrick Marks, Karine Valud-Martinez, Sascha Meiers, Katherine M Munson, Fabio Navarro, Bradley J Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stutz, Diana C J Spierings, Alistair Ward, AnneMarie E Welsch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Goo Jun, Li Ding, Chong Lek Koh, Bing Ren, Paul Flicek, Ken Chen, Mark B Gerstein, Pui-Yan Kwok, Peter M Lansdorp, Gabor Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E Devine, Michael Talkowski, Ryan E Mills, Tobias Marschall, Jan Korbel, Evan E Eichler, and Charles Lee (2017). "Multi-platform discovery of haplotype-resolved structural variation in human genomes". In: *bioRxiv*.

Chaisson, Mark J P and Glenn Tesler (2012). "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory." In: *BMC bioinformatics* 13, p. 238. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-238.

Chen, K., L. Chen, X. Fan, J. Wallis, L. Ding, and G. Weinstock (2014). "TIGRA: A targeted iterative graph routing assembler for breakpoint assembly". In: *Genome Research* 24.2, pp. 310–317. ISSN: 1088-9051. DOI: 10.1101/gr.162883.113.

Chen, Ken, John W Wallis, Michael D McLellan, David E Larson, Joelle M Kalicki, Craig S Pohl, Sean D McGrath, Michael C Wendl, Qunyuan Zhang, Devin P Locke, Xiaoqi Shi, Robert S Fulton, Timothy J Ley, Richard K Wilson, Li Ding, and Elaine R Mardis (2009). "BreakDancer: an algorithm for high-resolution mapping of genomic structural variation." In: *Nature methods* 6.9, pp. 677–81. ISSN: 1548-7105. DOI: 10.1038/nmeth.1363.

Chiang, Colby, Alexandra J Scott, Joe R Davis, Emily K Tsang, Xin Li, Yungil Kim, Tarik Hadzic, Farhan N Damani, Liron Ganel, Stephen B Montgomery, Alexis Battle, Donald

F Conrad, and Ira M Hall (2017). "The impact of structural variation on human gene expression". In: *Nature Genetics* 49.5, pp. 692–699. ISSN: 1061-4036. DOI: 10.1038/ng.3834.

Chiang, Derek Y, Gad Getz, David B Jaffe, Michael J T O'Kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander (2009). "High-resolution mapping of copy-number alterations with massively parallel sequencing". In: *Nature Methods* 6.1, pp. 99–103. ISSN: 1548-7091. DOI: 10.1038/nmeth.1276.

Chin, Chen-Shan, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach (2013). "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data". In: *Nature Methods* 10.6, pp. 563–569. ISSN: 1548-7091. DOI: 10.1038/nmeth.2474.

Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R Cramer, Massimo Delledonne, Chongyuan Luo, Joseph R Ecker, Dario Cantu, David R Rank, and Michael C Schatz (2016). "Phased diploid genome assembly with single-molecule real-time sequencing." In: *Nature methods* 13.12, pp. 1050–1054. ISSN: 1548-7105. DOI: 10.1038/nmeth.4035.

Chong, Zechen, Jue Ruan, Min Gao, Wanding Zhou, Tenghui Chen, Xian Fan, Li Ding, Anna Y Lee, Paul Boutros, Junjie Chen, and Ken Chen (2017). "novoBreak: local assembly for breakpoint detection in cancer genomes". In: *Nature Methods* 14.1, pp. 65–67. ISSN: 1548-7091. DOI: 10.1038/nmeth.4084.

Collins, Ryan L, Harrison Brand, Claire E Redin, Carrie Hanscom, Caroline Antolik, Matthew R Stone, Joseph T Glessner, Tamara Mason, Giulia Pregno, Naghmeh Dorrani, Giorgia Mandrile, Daniela Giachino, Danielle Perrin, Cole Walsh, Michelle Cipicchio, Maura Costello, Alexei Stortchevoi, Joon-Yong An, Benjamin B Currall, Catarina M Seabra, Ashok Ragavendran, Lauren Margolin, Julian A Martinez-Agosto, Diane Lucente, Brynn Levy, Stephan J Sanders, Ronald J Wapner, Fabiola Quintero-Rivera, Wigard Kloosterman, and Michael E Talkowski (2017). "Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome". In: *Genome Biology* 18.1, p. 36. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1158-6.

Conrad, Donald F, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T Daniel Andrews, Chris Barnes, Peter Campbell, Tomas Fitzgerald, Min Hu, Chun Hwa Ihm, Kati Kristiansson, Daniel G Macarthur, Jeffrey R Macdonald, Ifejinelo Onyiah, Andy Wing Chun Pang, Sam Robson, Kathy Stirrups, Armand Valsesia, Klaudia Walter, John Wei, Chris Tyler-Smith, Nigel P Carter, Charles Lee, Stephen W Scherer, and Matthew E Hurles (2010). "Origins and functional impact of copy number variation in the human genome." In: *Nature* 464.7289, pp. 704–712. ISSN: 0028-0836. DOI: 10.1038/nature08516.

Cremer, T and C Cremer (2001). "Chromosome territories, nuclear architecture and gene regulation in mammalian cells." In: *Nature reviews. Genetics* 2.4, pp. 292–301. ISSN: 1471-0056. DOI: 10.1038/35066075.

Davoli, Teresa and Titia de Lange (2011). "The causes and consequences of polyploidy in normal development and cancer." In: *Annual review of cell and developmental biology* 27, pp. 585–610. ISSN: 1530-8995. DOI: 10.1146/annurev-cellbio-092910-154234.

De Koning, A. P. Jason, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, and David D. Pollock (2011). "Repetitive Elements May Comprise Over Two-Thirds of the Human Genome". In: *PLoS Genetics* 7.12. Ed. by Gregory P. Copenhaver, e1002384. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002384.

Deamer, David, Mark Akeson, and Daniel Branton (2016). "Three decades of nanopore sequencing". In: *Nature Biotechnology* 34.5, pp. 518–524. ISSN: 1087-0156. DOI: 10.1038/nbt.3423.

Dean, F. B., S. Hosono, L. Fang, X. Wu, A. F. Faruqi, P. Bray-Ward, Z. Sun, Q. Zong, Y. Du, J. Du, M. Driscoll, W. Song, S. F. Kingsmore, M. Egholm, and R. S. Lasken (2002). "Comprehensive human genome amplification using multiple displacement amplification". In: *Proceedings of the National Academy of Sciences* 99.8, pp. 5261–5266. ISSN: 0027-8424. DOI: 10.1073/pnas.082089499.

Dekker, Job and Edith Heard (2015). "Structural and functional diversity of Topologically Associating Domains." In: *FEBS letters* 589.20 Pt A, pp. 2877–84. ISSN: 1873-3468. DOI: 10.1016/j.febslet.2015.08.044.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner (2002). "Capturing chromosome conformation." In: *Science (New York, N.Y.)* 295.5558, pp. 1306–11. ISSN: 1095-9203. DOI: 10.1126/science.1067799.

Deleye, Lieselot, Laurentijn Tilleman, Ann-Sophie Vander Plaetsen, Senne Cornelis, Dieter Deforce, and Filip Van Nieuwerburgh (2017). "Performance of four modern whole genome amplification methods for copy number variant detection in single cells". In: *Scientific Reports* 7.1, p. 3422. ISSN: 2045-2322. DOI: 10.1038/s41598-017-03711-y.

Demeulemeester, Jonas, Parveen Kumar, Elen K. Møller, Silje Nord, David C. Wedge, April Peterson, Randi R. Mathiesen, Renathe Fjelldal, Masoud Zamani Esteki, Koen Theunis, Elia Fernandez Gallardo, A. Jason Grundstad, Elin Borgen, Lars O. Baumbusch, Anne-Lise Børresen-Dale, Kevin P. White, Vessela N. Kristensen, Peter Van Loo, Thierry Voet, and Bjørn Naume (2016). "Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing". In: *Genome Biology* 17.1, p. 250. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1109-7.

Dilthey, Alexander, Charles Cox, Zamin Iqbal, Matthew R Nelson, and Gil McVean (2015). "Improved genome inference in the MHC using a population reference graph". In: *Nature Genetics* 47.6, pp. 682–688. ISSN: 1061-4036. DOI: 10.1038/ng.3257.

Dittwald, Piotr, Tomasz Gambin, Claudia Gonzaga-Jauregui, Claudia M.B. Carvalho, James R. Lupski, Paweł Stankiewicz, and Anna Gambin (2013). "Inverted Low-Copy Repeats and Genome Instability-A Genome-Wide Analysis". In: *Human Mutation* 34.1, pp. 210–220. ISSN: 10597794. DOI: 10.1002/humu.22217.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In: *Nature* 485.7398, pp. 376–380. ISSN: 0028-0836. DOI: 10.1038/nature11082.

DNA sequencing (Wikipedia) (2018). *DNA sequencing — Wikipedia, The Free Encyclopedia*. Accessed 14 February 2018. URL: https://en.wikipedia.org/w/index.php?title=DNA_sequencing&oldid=823930239.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark J P Chaisson, and Thomas R Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner". In: 29.1, pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker (2006). "Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements". In: *Genome Research* 16.10, pp. 1299–1309. ISSN: 1088-9051. DOI: 10.1101/gr.5571506.

Doudna, J. A. and E. Charpentier (2014). "The new frontier of genome engineering with CRISPR-Cas9". In: *Science* 346.6213, pp. 1258096–1258096. ISSN: 0036-8075. DOI: 10.1126/science.1258096.

Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, Ido Machol, Eric S. Lander, Aviva Presser Aiden, and Erez Lieberman Aiden (2017). "De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds". In: *Science* 356.6333, pp. 92–95. ISSN: 0036-8075. DOI: 10.1126/science.aal3327.

Edge, Peter, Vineet Bafna, and Vikas Bansal (2017). "HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies." In: *Genome research* 27.5, pp. 801–812. ISSN: 1549-5469. DOI: 10.1101/gr.213462.116.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner (2009). "Real-Time DNA Sequencing from Single Polymerase Molecules". In: *Science* 323.5910, pp. 133–138. ISSN: 0036-8075. DOI: 10.1126/science.1162986.

Emde, A K, M H Schulz, D Weese, R Sun, M Vingron, V M Kalscheuer, S A Haas, and K Reinert (2012). "Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS". In: *Bioinformatics* 28.5, pp. 619–627.

ENCODE Project Consortium, The (2012). "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414, pp. 57–74. ISSN: 0028-0836. DOI: 10.1038/nature11247.

English, Adam C, William J Salerno, Oliver A Hampton, Claudia Gonzaga-Jauregui, Shruthi Ambreth, Deborah I Ritter, Christine R Beck, Caleb F Davis, Mahmoud Dahdouli, Singer Ma, Andrew Carroll, Narayanan Veeraraghavan, Jeremy Bruestle, Becky Drees, Alex Hastie, Ernest T Lam, Simon White, Pamela Mishra, Min Wang, Yi Han, Feng Zhang, Pawel Stankiewicz, David A Wheeler, Jeffrey G Reid, Donna M Muzny, Jeffrey Rogers, Aniko Sabo, Kim C Worley, James R Lupski, Eric Boerwinkle, and Richard A

Gibbs (2015). "Assessing structural variation in a personal genome—towards a human reference diploid genome". In: *BMC Genomics* 16.1, p. 286. ISSN: 1471-2164. DOI: 10.1186/s12864-015-1479-3.

Evrony, Gilad D., Xuyu Cai, Eunjung Lee, L. Benjamin Hills, Princess C. Elhosary, Hillel S. Lehmann, J.J. Parker, Kutay D. Atabay, Edward C. Gilmore, Annapurna Poduri, Peter J. Park, and Christopher A. Walsh (2012). "Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain". In: *Cell* 151.3, pp. 483–496. ISSN: 00928674. DOI: 10.1016/j.cell.2012.09.035.

Falconer, Ester, Mark Hills, Ulrike Naumann, Steven S S Poon, Elizabeth a Chavez, Ashley D Sanders, Yongjun Zhao, Martin Hirst, and Peter M Lansdorp (2012). "DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution". In: *Nature Methods* 9.11, pp. 1107–1112. ISSN: 1548-7091. DOI: 10.1038/nmeth.2206.

Fanciulli, Manuela, Penny J Norsworthy, Enrico Petretto, Rong Dong, Lorraine Harper, Lavanya Kamesh, Joanne M Heward, Stephen C L Gough, Adam de Smith, Alexandra I F Blakemore, Philippe Froguel, Catherine J Owen, Simon H S Pearce, Luis Teixeira, Loic Guillevin, Deborah S Cunninghame Graham, Charles D Pusey, H Terence Cook, Timothy J Vyse, and Timothy J Aitman (2007). "FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity." In: *Nature genetics* 39.6, pp. 721–3. ISSN: 1061-4036. DOI: 10.1038/ng2046.

Faust, Gregory G and Ira M Hall (2012). "YAHA: fast and flexible long-read alignment with optimal breakpoint detection." In: *Bioinformatics (Oxford, England)* 28.19, pp. 2417–24. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts456.

Fehrmann, Rudolf S N, Juha M Karjalainen, Małgorzata Krajewska, Harm-Jan Westra, David Maloney, Anton Simeonov, Tune H Pers, Joel N Hirschhorn, Ritsert C Jansen, Erik A Schultes, Herman H H B M van Haagen, Elisabeth G E de Vries, Gerard J te Meerman, Cisca Wijmenga, Marcel A T M van Vugt, and Lude Franke (2015). "Gene expression analysis identifies global gene dosage sensitivity in cancer". In: *Nature Genetics* 47.2, pp. 115–125. ISSN: 1061-4036. DOI: 10.1038/ng.3173.

Feuk, Lars (2010). "Inversion variants in the human genome: role in disease and genome architecture." In: *Genome medicine* 2.2, p. 11. ISSN: 1756-994X. DOI: 10.1186/gm132.

Fitch, W (1969). "Locating gaps in amino acid sequences to optimize the homology between two proteins." In: *Biochemical Genetics* 3, pp. 99–108.

Flusberg, Benjamin A, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner (2010). "Direct detection of DNA methylation during single-molecule, real-time sequencing". In: *Nature Methods* 7.6, pp. 461–465. ISSN: 1548-7091. DOI: 10.1038/nmeth.1459.

Forbes, S. A., N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal (2011). "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer". In: *Nucleic Acids Research* 39.Database, pp. D945–D950. ISSN: 0305-1048. DOI: 10.1093/nar/gkq929. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq929.

Forcato, Mattia, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato (2017). "Comparison of computational methods for Hi-C data analysis". In: *Nature Methods* 14.7, pp. 679–685. ISSN: 1548-7091. DOI: 10.1038/nmeth.4325.

Forsberg, Lars A., David Gisselsson, and Jan P. Dumanski (2017). "Mosaicism in health and disease — clones picking up speed". In: *Nature Reviews Genetics* 18.2, pp. 128–142. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.145.

Forsberg, Lars A., Chiara Rasi, Hamid R. Razzaghian, Geeta Pakalapati, Lindsay Waite, Krista Stanton Thilbeault, Anna Ronowicz, Nathan E. Wineinger, Hemant K. Tiwari, Dorret Boomsma, Maxwell P. Westerman, Jennifer R. Harris, Robert Lyle, Magnus Essand, Fredrik Eriksson, Themistocles L. Assimes, Carlos Iribarren, Eric Strachan, Terrance P. O'Hanlon, Lisa G. Rider, Frederick W. Miller, Vilmantas Giedraitis, Lars Lannfelt, Martin Ingelsson, Arkadiusz Piotrowski, Nancy L. Pedersen, Devin Absher, and Jan P. Dumanski (2012). "Age-Related Somatic Structural Changes in the Nuclear Genome of Human Blood Cells". In: *The American Journal of Human Genetics* 90.2, pp. 217–228. ISSN: 00029297. DOI: 10.1016/j.ajhg.2011.12.009. URL: http://linkinghub.elsevier.com/retrieve/pii/S0002929711005441.

Fox, Edward J, Kate S Reid-Bayliss, Mary J Emond, and Lawrence A Loeb (2014). "Accuracy of Next Generation Sequencing Platforms." In: *Next generation, sequencing & applications* 1. ISSN: 2469-9853. DOI: 10.4172/jngsa.1000106.

Freese, Nowlan H, David C Norris, and Ann E Loraine (2016). "Integrated genome browser: visual analytics platform for genomics." In: *Bioinformatics (Oxford, England)* 32.14, pp. 2089–95. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw069.

Frith, Martin C and Risa Kawaguchi (2015). "Split-alignment of genomes finds orthologies more accurately". In: *Genome Biology* 16.1, p. 106. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0670-9.

Fritsch, Edward F., Richard M. Lawn, and Tom Maniatis (1979). "Characterisation of deletions which affect the expression of fetal globin genes in man". In: *Nature* 279.5714, pp. 598–603. ISSN: 0028-0836. DOI: 10.1038/279598a0.

Furlong, Eileen E.M., David Profitt, and Matthew P. Scott (2001). "Automated sorting of live transgenic embryos". In: *Nature Biotechnology* 19.2, pp. 153–156. ISSN: 10870156. DOI: 10.1038/84422.

Gao, Ruli, Alexander Davis, Thomas O McDonald, Emi Sei, Xiuqing Shi, Yong Wang, Pei-Ching Tsai, Anna Casasent, Jill Waters, Hong Zhang, Funda Meric-Bernstam, Franziska Michor, and Nicholas E Navin (2016). "Punctuated copy number evolution and clonal stasis in triple-negative breast cancer". In: *Nature Genetics* 48.10, pp. 1119–1130. ISSN: 1061-4036. DOI: 10.1038/ng.3641.

Garrison, Erik and Gabor Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: p. 9. arXiv: 1207.3907.

Garvin, Tyler, Robert Aboukhalil, Jude Kendall, Timour Baslan, Gurinder S Atwal, James Hicks, Michael Wigler, and Michael C Schatz (2015). "Interactive analysis and assessment of single-cell copy-number variations". In: *Nature Methods* 12.11, pp. 1058–1060. ISSN: 1548-7091. DOI: 10.1038/nmeth.3578.

Gehring, Julian S., Bernd Fischer, Michael Lawrence, and Wolfgang Huber (2015). "SomaticSignatures: inferring mutational signatures from single-nucleotide variants: Fig. 1." In: *Bioinformatics*, btv408. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv408. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv408.

Ghavi-Helm, Yad, Felix A Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E M Furlong (2014). "Enhancer loops appear stable during development and are associated with paused polymerase". In: *Nature* 512.7512, pp. 96–100. ISSN: 0028-0836. DOI: 10.1038/nature13417.

Gibbs, A J and G A McIntyre (1970). "The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences." In: *European Journal of Biochemistry* 16.1, pp. 1–11. ISSN: 0014-2956.

Gibcus, Johan H. and Job Dekker (2013). "The Hierarchy of the 3D Genome". In: *Molecular Cell* 49.5, pp. 773–782. ISSN: 10972765. DOI: 10.1016/j.molcel.2013.02.011.

Gordon, David J., Benjamin Resio, and David Pellman (2012). "Causes and consequences of aneuploidy in cancer". In: *Nature Reviews Genetics* 13.3, pp. 189–203. ISSN: 1471-0056. DOI: 10.1038/nrg3123.

Gorkin, David U., Danny Leung, and Bing Ren (2014). "The 3D Genome in Transcriptional Regulation and Pluripotency". In: *Cell Stem Cell* 14.6, pp. 762–775. ISSN: 19345909. DOI: 10.1016/j.stem.2014.05.017.

Gramates, L. Sian, Steven J. Marygold, Gilberto dos Santos, Jose-Maria Urbano, Giulia Antonazzo, Beverley B. Matthews, Alix J. Rey, Christopher J. Tabone, Madeline A. Crosby, David B. Emmert, Kathleen Falls, Joshua L. Goodman, Yanhui Hu, Laura Ponting, Andrew J. Schroeder, Victor B. Strelets, Jim Thurmond, and Pinglei Zhou (2017). "FlyBase at 25: looking to the future". In: *Nucleic Acids Research* 45.D1, pp. D663–D671. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1016.

Guo, Ya, Quan Xu, Daniele Canzio, Jia Shou, Jinhuan Li, David U. Gorkin, Inkyung Jung, Haiyang Wu, Yanan Zhai, Yuanxiao Tang, Yichao Lu, Yonghu Wu, Zhilian Jia, Wei Li, Michael Q. Zhang, Bing Ren, Adrian R. Krainer, Tom Maniatis, and Qiang Wu (2015). "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function". In: *Cell* 162.4, pp. 900–910. ISSN: 00928674. DOI: 10.1016/j.cell.2015.07.038.

Hajirasouliha, Iman, Fereydoun Hormozdiari, Can Alkan, Jeffrey M. Kidd, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp (2010). "Detection and characterization of novel sequence insertions using paired-end next-generation sequencing". In: *Bioinformatics* 26.10, pp. 1277–1283. ISSN: 14602059. DOI: 10.1093/bioinformatics/btq152.

Handsaker, Robert E, Vanessa Van Doren, Jennifer R Berman, Giulio Genovese, Seva Kashin, Linda M Boettger, and Steven a McCarroll (2015). "Large multiallelic copy number variations in humans". In: *Nature Genetics* 47.3, pp. 296–303. ISSN: 1061-4036. DOI: 10.1038/ng.3200.

Harewood, Louise, Kamal Kishore, Matthew D. Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V. Peter Collins, and Peter Fraser (2017). "Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number

variation in human tumours". In: *Genome Biology* 18.1, p. 125. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1253-8.

Harvey, C., G. a. Moyebrailean, O. Davis, X. Wen, F. Luca, and R. Pique-Regi (2014). *QuASAR: Quantitative Allele Specific Analysis of Reads*. Tech. rep. December 2014, p. 007492. DOI: 10.1101/007492.

Hastings, P J, Grzegorz Ira, and James R Lupski (2009). "A microhomology-mediated break-induced replication model for the origin of human copy number variation." In: *PLoS genetics* 5.1, e1000327. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000327.

Hastings, P J, James R Lupski, Susan M Rosenberg, and Grzegorz Ira (2009). "Mechanisms of change in gene copy number." In: *Nature reviews. Genetics* 10.8, pp. 551–64. ISSN: 1471-0064. DOI: 10.1038/nrg2593.

Haubold, Bernhard and Thomas Wiehe (2006). "How repetitive are genomes?" In: *BMC Bioinformatics* 7.1, p. 541. ISSN: 14712105. DOI: 10.1186/1471-2105-7-541.

Heather, James M. and Benjamin Chain (2016). "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1, pp. 1–8. ISSN: 08887543. DOI: 10.1016/j.ygeno.2015.11.003.

Hehir-Kwa, Jayne Y., Tobias Marschall, Wigard P. Kloosterman, Laurent C. Francioli, Jasmijn A. Baaijens, Louis J. Dijkstra, Abdel Abdellaoui, Vyacheslav Koval, Djie Tjwan Thung, René Wardenaar, Ivo Renkens, Bradley P. Coe, Patrick Deelen, Joep de Ligt, Eric-Wubbo Lameijer, Freerk van Dijk, Fereydoun Hormozdiari, Jasper A. Bovenberg, Anton J. M. de Craen, Marian Beekman, Albert Hofman, Gonneke Willemsen, Bruce Wolffenbuttel, Mathieu Platteel, Yuanping Du, Ruoyan Chen, Hongzhi Cao, Rui Cao, Yushen Sun, Jeremy Sujie Cao, Pieter B. T. Neerincx, Martijn Dijkstra, George Byelas, Alexandros Kanterakis, Jan Bot, Martijn Vermaat, Jeroen F. J. Laros, Johan T. den Dunnen, Peter de Knijff, Lennart C. Karssen, Elisa M. van Leeuwen, Najaf Amin, Fernando Rivadeneira, Karol Estrada, Jouke-Jan Hottenga, V. Mathijs Kattenberg, David van Enckevort, Hailiang Mei, Mark Santcroos, Barbera D. C. van Schaik, Robert E. Handsaker, Steven A. McCarroll, Arthur Ko, Peter H Sudmant, Isaac J. Nijman, André G. Uitterlinden, Cornelia M. van Duijn, Evan E. Eichler, Paul I. W. de Bakker, Morris A. Swertz, Cisca Wijmenga, Gert-Jan B. van Ommen, P. Eline Slagboom, Dorret I. Boomsma, Alexander Schönhuth, Kai Ye, and Victor Guryev (2016). "A high-quality human reference panel reveals the complexity and distribution of genomic structural variants". In: *Nature Communications* 7, p. 12989. ISSN: 2041-1723. DOI: 10.1038/ncomms12989.

Hills, Mark, Kieran O'Neill, Ester Falconer, Ryan Brinkman, and Peter M Lansdorp (2013). "BAIT: Organizing genomes and mapping rearrangements in single cells". In: *Genome Medicine* 5.9, p. 82. ISSN: 1756-994X. DOI: 10.1186/gm486.

Hiltemann, Saskia, Guido Jenster, Jan Trapman, Peter van der Spek, and Andrew Stubbs (2015). "Discriminating somatic and germline mutations in tumor DNA samples without matching normals". In: *Genome Research* 25.9, pp. 1382–1390. ISSN: 1088-9051. DOI: 10.1101/gr.183053.114.

Hnisz, D., A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young (2016). "Activation of proto-oncogenes by

disruption of chromosome neighborhoods". In: *Science* 351.6280, pp. 1454–1458. ISSN: 0036-8075. DOI: 10.1126/science.aad9024.

Hnisz, Denes, Daniel S. Day, and Richard A. Young (2016). "Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control". In: *Cell* 167.5, pp. 1188–1200. ISSN: 00928674. DOI: 10.1016/j.cell.2016.10.024.

Holstege, H., W. Pfeiffer, D. Sie, M. Hulsman, T. J. Nicholas, C. C. Lee, T. Ross, J. Lin, M. A. Miller, B. Ylstra, H. Meijers-Heijboer, M. H. Brugman, F. J. T. Staal, G. Holstege, M. J. T. Reinders, T. T. Harkins, S. Levy, and E. A. Sistermans (2014). "Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis". In: *Genome Research* 24.5, pp. 733–742. ISSN: 1088-9051. DOI: 10.1101/gr.162131.113.

Holtgrewe, Manuel, Leon Kuchenbecker, and Knut Reinert (2015). "Methods for the detection and assembly of novel sequence in high-throughput sequencing data". In: *Bioinformatics* 31.12, pp. 1904–1912. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv051.

Hommelsheim, Carl Maximilian, Lamprinos Frantzeskakis, Mengmeng Huang, and Bekir Ülker (2015). "PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications". In: *Scientific Reports* 4.1, p. 5052. ISSN: 2045-2322. DOI: 10.1038/srep05052.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing". In: *Nature Genetics* 44.8, pp. 955–959. ISSN: 1061-4036. DOI: 10.1038/ng.2354.

Huang, W., A. Massouras, Y. Inoue, J. Peiffer, M. Ramia, A. M. Tarone, L. Turlapati, T. Zichner, D. Zhu, R. F. Lyman, M. M. Magwire, K. Blankenburg, M. A. Carbone, K. Chang, L. L. Ellis, S. Fernandez, Y. Han, G. Highnam, C. E. Hjelmen, J. R. Jack, M. Javaid, J. Jayaseelan, D. Kalra, S. Lee, L. Lewis, M. Munidasa, F. Ongeri, S. Patel, L. Perales, A. Perez, L. Pu, S. M. Rollmann, R. Ruth, N. Saada, C. Warner, A. Williams, Y.-Q. Wu, A. Yamamoto, Y. Zhang, Y. Zhu, R. R. H. Anholt, J. O. Korbel, D. Mittelman, D. M. Muzny, R. A. Gibbs, A. Barbadilla, J. S. Johnston, E. A. Stone, S. Richards, B. Deplancke, and T. F. C. Mackay (2014). "Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines". In: *Genome Research* 24.7, pp. 1193–1208. ISSN: 1088-9051. DOI: 10.1101/gr.171546.113.

Huber, Wolfgang, Joern Toedling, and Lars M. Steinmetz (2006). "Transcript mapping with high-density oligonucleotide tiling arrays". In: *Bioinformatics* 22.16, pp. 1963–1970. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btl289.

Huddleston, John, Mark J P Chaisson, Karyn Meltz Steinberg, Wes Warren, Kendra Hoekzema, David Gordon, Tina A Graves-Lindsay, Katherine M Munson, Zev N Kronenberg, Laura Vives, Paul Peluso, Matthew Boitano, Chen-Shin Chin, Jonas Korlach, Richard K Wilson, and Evan E Eichler (2017). "Discovery and genotyping of structural variation from long-read haploid genome sequence data." In: *Genome research* 27.5, pp. 677–685. ISSN: 1549-5469. DOI: 10.1101/gr.214007.116.

Iafrate, A John, Lars Feuk, Miguel N Rivera, Marc L Listewnik, Patricia K Donahoe, Ying Qi, Stephen W Scherer, and Charles Lee (2004). "Detection of large-scale variation in

the human genome." In: *Nature genetics* 36.9, pp. 949–51. ISSN: 1061-4036. DOI: 10.1038/ng1416.

Ibn-Salem, Jonas, Sebastian Köhler, Michael I Love, Ho-Ryun Chung, Ni Huang, Matthew E Hurles, Melissa Haendel, Nicole L Washington, Damian Smedley, Christopher J Mungall, Suzanna E Lewis, Claus-Eric Ott, Sebastian Bauer, Paul N Schofield, Stefan Mundlos, Malte Spielmann, and Peter N Robinson (2014). "Deletions of chromosomal regulatory boundaries are associated with congenital disease". In: *Genome Biology* 15.9, p. 423. ISSN: 1465-6906.

International HapMap 3 Consortium, The (2010). "Integrating common and rare genetic variation in diverse human populations". In: *Nature* 467.7311, pp. 52–58. ISSN: 0028-0836. DOI: 10.1038/nature09298.

International HapMap Consortium, The (2005). "A haplotype map of the human genome." In: *Nature* 437.10, pp. 1299–1320. ISSN: 1476-4687. DOI: 10.1038/nature04226.

– (2007). "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164, pp. 851–861. ISSN: 0028-0836. DOI: 10.1038/nature06258.

International Human Genome Sequencing Consortium, The (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062. URL: http://www.nature.com/doifinder/10.1038/35057062.

Istace, Benjamin, Anne Friedrich, Léo D'Agata, Sébastien Faye, Emilie Payen, Odette Beluche, Claudia Caradec, Sabrina Davidas, Corinne Cruaud, Gianni Liti, Arnaud Lemainque, Stefan Engelen, Patrick Wincker, Joseph Schacherer, and Jean-Marc Aury (2017). "de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer". In: *GigaScience* 6.2, pp. 1–13. ISSN: 2047-217X. DOI: 10.1093/gigascience/giw018.

Jain, Miten, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson (2015). "Improved data analysis for the MinION nanopore sequencer". In: *Nature Methods* February. ISSN: 1548-7091. DOI: 10.1038/nmeth.3290.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". In: *Nature Biotechnology*. ISSN: 1087-0156. DOI: 10.1038/nbt.4060.

Ji, W, X Y Zhang, G S Warshamana, G Z Qu, and M Ehrlich (1994). "Effect of internal direct and inverted Alu repeat sequences on PCR." In: *PCR methods and applications* 4.2, pp. 109–16. ISSN: 1054-9803.

Ji, Xiong, Daniel B. Dadon, Benjamin E. Powell, Zi Peng Fan, Diego Borges-Rivera, Sigal Shachar, Abraham S. Weintraub, Denes Hnisz, Gianluca Pegoraro, Tong Ihn Lee, Tom Misteli, Rudolf Jaenisch, and Richard A. Young (2016). "3D Chromosome Regulatory Landscape of Human Pluripotent Cells". In: *Cell Stem Cell* 18.2, pp. 262–275. ISSN: 19345909. DOI: 10.1016/j.stem.2015.11.007.

Kaplan, Noam and Job Dekker (2013). "High-throughput genome scaffolding from in vivo DNA interaction frequency." In: *Nature biotechnology* 31.12, pp. 1143–7. ISSN: 1546-1696. DOI: 10.1038/nbt.2768.

Kerpedjiev, Peter, Nezar Abdennur, Fritz Lekschas, Chuck McCallum, Kasper Dinkla, Hendrik Strobelt, Jacob M Luber, Scott B Ouellette, Alaleh Azhir, Nikhil Kumar, Jeewon Hwang, Soohyun Lee, Burak H Alver, Hanspeter Pfister, Leonid A Mirny, Peter J Park, and Nils Gehlenborg (2017). "HiGlass: Web-based Visual Exploration and Analysis of Genome Interaction Maps". In: *bioRxiv*.

Kidd, J M, G M Cooper, W F Donahue, H S Hayden, N Sampas, T Graves, N Hansen, B Teague, C Alkan, F Antonacci, E Haugen, T Zerr, N A Yamada, P Tsang, T L Newman, E Tuzun, Z Cheng, H M Ebling, N Tusneem, R David, W Gillett, K A Phelps, M Weaver, D Saranga, A Brand, W Tao, E Gustafson, K McKernan, L Chen, M Malig, J D Smith, J M Korn, S A McCarroll, D A Altshuler, D A Peiffer, M Dorschner, J Stamatoyannopoulos, D Schwartz, D A Nickerson, J C Mullikin, R K Wilson, L Bruhn, M V Olson, R Kaul, D R Smith, and E E Eichler (2008). "Mapping and sequencing of structural variation from eight human genomes". In: *Nature* 453.7191, pp. 56–64. ISSN: 1476-4687. DOI: 10.1038/nature06862. arXiv: NIHMS150003.

Kielbasa, Szymon M, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith (2011). "Adaptive seeds tame genomic sequence comparison." In: *Genome Research* 21.3, pp. 487–493. DOI: 10.1101/gr.113985.110.

Kleinjan, Dirk A. and Veronica van Heyningen (2005). "Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease". In: *The American Journal of Human Genetics* 76.1, pp. 8–32. ISSN: 00029297. DOI: 10.1086/426833.

Knouse, Kristin A., Jie Wu, and Angelika Amon (2016). "Assessment of megabase-scale somatic copy number variation using single-cell sequencing". In: *Genome Research* 26.3, pp. 376–384. ISSN: 1088-9051. DOI: 10.1101/gr.198937.115.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. E. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder (2007). "Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome". In: *Science* 318.5849, pp. 420–426. ISSN: 0036-8075. DOI: 10.1126/science.1149504.

Koren, Sergey, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy P L Smith, and Adam Phillippy (2018). "Complete assembly of parental haplotypes with trio binning". In: *bioRxiv*. URL: http://biorxiv.org/content/early/2018/02/26/271486.abstract.

Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation". In: *Genome Research* 27.5, pp. 722–736. ISSN: 1088-9051. DOI: 10.1101/gr.215087.116.

Köster, J and S Rahmann (2012). "Snakemake–a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19, pp. 2520–2522. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts480.

Krijger, Peter Hugo Lodewijk and Wouter de Laat (2016). "Regulation of disease-associated gene expression in the 3D genome". In: *Nature Reviews Molecular Cell Biology*. ISSN: 1471-0072. DOI: 10.1038/nrm.2016.138.

Kurtz, Stefan, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg (2004). "Versatile and open software for comparing large genomes." In: *Genome biology* 5.2, R12. ISSN: 1474-760X. DOI: 10.1186/gb-2004-5-2-r12.

Lam, Hugo Y K, Xinmeng Jasmine Mu, Adrian M Stütz, Andrea Tanzer, Philip D Cayting, Michael Snyder, Philip M Kim, Jan O Korbel, and Mark B Gerstein (2010). "Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library". In: *Nature Biotechnology* 28.1, pp. 47–55. ISSN: 1087-0156. DOI: 10.1038/nbt.1600.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." In: *Genome biology* 10.3, R25. ISSN: 1465-6914. DOI: 10.1186/gb-2009-10-3-r25.

Layer, Ryan M, Colby Chiang, Aaron R Quinlan, and Ira M Hall (2014). "LUMPY: a probabilistic framework for structural variant discovery". In: *Genome Biology* 15.6, R84. ISSN: 1465-6906. DOI: 10.1186/gb-2014-15-6-r84.

Le Dily, François, Davide Baù, Andy Pohl, Guillermo P. Vicent, François Serra, Daniel Soronellas, Giancarlo Castellano, Roni H.G. Wright, Cecilia Ballare, Guillaume Filion, Marc A. Marti-Renom, and Miguel Beato (2014). "Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation". In: *Genes & Development* 28.19, pp. 2151–2162. ISSN: 0890-9369. DOI: 10.1101/gad.241422.114.

Lejeune, J, M Gautier, and R Turpin (1959). "Study of somatic chromosomes from 9 mongoloid children". In: *Comptes rendus hebdomadaires des seances de l'Academie des sciences* 248.11, pp. 1721–2. ISSN: 0001-4036.

Lettice, Laura A., Sarah Daniels, Elizabeth Sweeney, Shanmugasundaram Venkataraman, Paul S. Devenney, Philippe Gautier, Harris Morrison, Judy Fantes, Robert E. Hill, and David R. FitzPatrick (2011). "Enhancer-adoption as a mechanism of human developmental disease". In: *Human Mutation* 32.12, pp. 1492–1499. ISSN: 10597794. DOI: 10.1002/humu.21615.

Li, Changsheng, Feng Lin, Dong An, Wenqin Wang, and Ruidong Huang (2017). "Genome Sequencing and Assembly by Long Reads in Plants." In: *Genes* 9.1. ISSN: 2073-4425. DOI: 10.3390/genes9010006.

Li, Guoliang, Liuyang Cai, Huidan Chang, Ping Hong, Qiangwei Zhou, Ekaterina V Kulakova, Nikolay A Kolchanov, and Yijun Ruan (2014). "Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application". In: *BMC Genomics* 15.Suppl 12, S11. ISSN: 1471-2164. DOI: 10.1186/1471-2164-15-S12-S11.

Li, Heng (2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: p. 3. arXiv: 1303.3997. URL: http://arxiv.org/abs/1303.3997.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14, pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup (2009). "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics (Oxford, England)* 25.16, pp. 2078–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352.

Li, Ruifeng, Yifang Liu, Tingting Li, and Cheng Li (2016). "3Disease Browser: A Web server for integrating 3D genome and disease-associated chromosome rearrangement data". In: *Scientific Reports* 6.1, p. 34651. ISSN: 2045-2322. DOI: 10.1038/srep34651.

Li, Yanjian, Yi He, Zhengyu Liang, Yang Wang, Fengling Chen, Mohamed Nadhir Djekidel, Guipeng Li, Xu Zhang, Shuqin Xiang, Zejun Wang, Juntao Gao, Michael Q. Zhang, and Yang Chen (2018). "Alterations of specific chromatin conformation affect ATRA-induced leukemia cell differentiation". In: *Cell Death & Disease* 9.2, p. 200. ISSN: 2041-4889. DOI: 10.1038/s41419-017-0173-6.

Li, Yingrui, Hancheng Zheng, Ruibang Luo, Honglong Wu, Hongmei Zhu, Ruiqiang Li, Hongzhi Cao, Boxin Wu, Shujia Huang, Haojing Shao, Hanzhou Ma, Fan Zhang, Shuijian Feng, Wei Zhang, Hongli Du, Geng Tian, Jingxiang Li, Xiuqing Zhang, Songgang Li, Lars Bolund, Karsten Kristiansen, Adam J de Smith, Alexandra I F Blakemore, Lachlan J M Coin, Huanming Yang, Jian Wang, and Jun Wang (2011). "Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly". In: *Nature Biotechnology* 29.8, pp. 723–730. ISSN: 1087-0156. DOI: 10.1038/nbt.1904.

Lieber, Michael R. (2008). "The Mechanism of Human Nonhomologous DNA End Joining". In: *Journal of Biological Chemistry* 283.1, pp. 1–5. ISSN: 0021-9258. DOI: 10.1074/jbc.R700039200.

Lieberman-Aiden, E., N. L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, E. S. Lander, and J. Dekker (2009). "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome". In: *Science* 326.5950, pp. 289–293. ISSN: 0036-8075. DOI: 10.1126/science.1181369.

Liu, Zhi, Tuantuan Gui, Zhen Wang, Hong Li, Yunhe Fu, Xiao Dong, and Yixue Li (2016). "cisASE: A likelihood-based method for detecting putative cis-regulated allele-specific expression in RNA sequencing data". In: *Bioinformatics*, btw416. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw416.

Lodato, M. A., M. B. Woodworth, S. Lee, G. D. Evrony, B. K. Mehta, A. Karger, S. Lee, T. W. Chittenden, A. M. D'Gama, X. Cai, L. J. Luquette, E. Lee, P. J. Park, and C. A. Walsh (2015). "Somatic mutation in single human neurons tracks developmental and transcriptional history". In: *Science* 350.6256, pp. 94–98. ISSN: 0036-8075. DOI: 10.1126/science.aab1785.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel, and Stefan Mundlos (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions". In: *Cell* 161.5, pp. 1012–1025. ISSN: 10974172. DOI: 10.1016/j.cell.2015.04.004.

Lupiáñez, Darío G., Malte Spielmann, and Stefan Mundlos (2016). "Breaking TADs: How Alterations of Chromatin Domains Result in Disease". In: *Trends in Genetics* 32.4, pp. 225–237. ISSN: 13624555. DOI: 10.1016/j.tig.2016.01.003.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson (2017). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic Acids Research* 45.D1, pp. D896–D901. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1133.

Mackay, Trudy F. C., Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F. Ayroles, Dianhui Zhu, Sònia Casillas, Yi Han, Michael M. Magwire, Julie M. Cridland, Mark F. Richardson, Robert R. H. Anholt, Maite Barrón, Crystal Bess, Kerstin Petra Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura Duncan, Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N. Jhangiani, Katherine W. Jordan, Fremiet Lara, Faye Lawrence, Sandra L. Lee, Pablo Librado, Raquel S. Linheiro, Richard F. Lyman, Aaron J. Mackey, Mala Munidasa, Donna Marie Muzny, Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel Ràmia, Jeffrey G. Reid, Stephanie M. Rollmann, Julio Rozas, Nehad Saada, Lavanya Turlapati, Kim C. Worley, Yuan-Qing Wu, Akihiko Yamamoto, Yiming Zhu, Casey M. Bergman, Kevin R. Thornton, David Mittelman, and Richard A. Gibbs (2012). "The Drosophila melanogaster Genetic Reference Panel". In: *Nature* 482.7384, pp. 173–178. ISSN: 0028-0836. DOI: 10.1038/nature10811.

Marouli, Eirini et al. (2017). "Rare and low-frequency coding variants alter human adult height". In: *Nature* 542.7640, pp. 186–190. ISSN: 0028-0836. DOI: 10.1038/nature21039.

Marschall, Tobias and The Computational Pan-Genomics Consortium (2016). "Computational pan-genomics: status, promises and challenges". In: *Briefings in Bioinformatics*, bbw089. ISSN: 1467-5463. DOI: 10.1093/bib/bbw089.

Marschall, Tobias, Ivan G Costa, Stefan Canzar, Markus Bauer, Gunnar W Klau, Alexander Schliep, and Alexander Schönhuth (2012). "CLEVER: clique-enumerating variant finder." In: *Bioinformatics (Oxford, England)* 28.22, pp. 2875–82. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts566.

Marschall, Tobias, Iman Hajirasouliha, and Alexander Schönhuth (2013). "MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels". In: *Bioinformatics* 29.24, pp. 3143–3150. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btt556.

Massouras, Andreas, Sebastian M. Waszak, Monica Albarca-Aguilera, Korneel Hens, Wiebke Holcombe, Julien F. Ayroles, Emmanouil T. Dermitzakis, Eric A. Stone, Jeffrey D. Jensen, Trudy F. C. Mackay, and Bart Deplancke (2012). "Genomic Variation and Its Impact on Gene Expression in Drosophila melanogaster". In: *PLoS Genetics* 8.11. Ed. by Brian Oliver, e1003055. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003055.

Mayba, Oleg, Houston N Gilbert, Jinfeng Liu, Peter M Haverty, Suchit Jhunjhunwala, Zhaoshi Jiang, Colin Watanabe, and Zemin Zhang (2014). "MBASED: allele-specific expression detection in cancer tissues and cell lines". In: *Genome Biology* 15.8, p. 405. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0405-3.

Maydan, Jason S, Adam Lorch, Mark L Edgley, Stephane Flibotte, and Donald G Moerman (2010). "Copy number variation in the genomes of twelve natural isolates of Caenorhabditis elegans". In: *BMC Genomics* 11.1, p. 62. ISSN: 1471-2164. DOI: 10.1186/1471-2164-11-62.

McCarroll, Steven A, Alan Huett, Petric Kuballa, Shannon D Chilewski, Aimee Landry, Philippe Goyette, Michael C Zody, Jennifer L Hall, Steven R Brant, Judy H Cho, Richard H Duerr, Mark S Silverberg, Kent D Taylor, John D Rioux, David Altshuler, Mark J Daly, and Ramnik J Xavier (2008). "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease." In: *Nature genetics* 40.9, pp. 1107–12. ISSN: 1061-4036. DOI: 10.1038/ng.215.

McMaster Pathophysiology Review (website) (2018). *Aneuploidy and non-disjunction*. Website, accessed 12 Frebruary 2018. Authors: Parashar, Richa and Johnson-Keeping, Cheryl. URL: http://www.pathophys.org/aneuploidy/.

Mercy, Guillaume, Julien Mozziconacci, Vittore F. Scolari, Kun Yang, Guanghou Zhao, Agnès Thierry, Yisha Luo, Leslie A. Mitchell, Michael Shen, Yue Shen, Roy Walker, Weimin Zhang, Yi Wu, Ze-xiong Xie, Zhouqing Luo, Yizhi Cai, Junbiao Dai, Huanming Yang, Ying-Jin Yuan, Jef D. Boeke, Joel S. Bader, Héloïse Muller, and Romain Koszul (2017). "3D organization of synthetic and scrambled chromosomes". In: *Science* 355.6329, eaaf4597. ISSN: 0036-8075. DOI: 10.1126/science.aaf4597.

Mertens, Fredrik, Bertil Johansson, Thoas Fioretos, and Felix Mitelman (2015). "The emerging complexity of gene fusions in cancer". In: *Nature Reviews Cancer* 15.6, pp. 371–381. ISSN: 1474-175X. DOI: 10.1038/nrc3947.

Miller, Danny E, Kevin R Cook, Alexandra V Arvanitakis, and R Scott Hawley (2016). "Third Chromosome Balancer Inversions Disrupt Protein-Coding Genes and Influence Distal Recombination Events in Drosophila melanogaster." In: *G3 (Bethesda, Md.)* Pp. 1–18. ISSN: 2160-1836. DOI: 10.1534/g3.116.029330.

Miller, Danny E, Kevin R Cook, Elizabeth A Hemenway, Vivienne Fang, Angela L Miller, Karen G Hales, and R Scott Hawley (2018). "The Molecular and Genetic Characterization of Second Chromosome Balancers in Drosophila melanogaster". In: *G3 (Bethesda, Md.)* ISSN: 2160-1836. DOI: 10.1534/g3.118.200021.

Mills, Ryan E, Klaudia Walter, Chip Stewart, Robert E Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R Keira Cheetham, Asif Chinwalla, Donald F Conrad, Yutao Fu, Fabian Grubert, Iman Hajirasouliha, Fereydoun Hormozdiari, Lilia M Iakoucheva, Zamin Iqbal, Shuli Kang, Jeffrey M Kidd, Miriam K Konkel,

Joshua Korn, Ekta Khurana, Deniz Kural, Hugo Y K Lam, Jing Leng, Ruiqiang Li, Yingrui Li, Chang-Yun Lin, Ruibang Luo, Xinmeng Jasmine Mu, James Nemesh, Heather E Peckham, Tobias Rausch, Aylwyn Scally, Xinghua Shi, Michael P Stromberg, Adrian M Stütz, Alexander Eckehart Urban, Jerilyn a Walker, Jiantao Wu, Yujun Zhang, Zhengdong D Zhang, Mark a Batzer, Li Ding, Gabor T Marth, Gil McVean, Jonathan Sebat, Michael Snyder, Jun Wang, Kenny Ye, Evan E Eichler, Mark B Gerstein, Matthew E Hurles, Charles Lee, Steven a McCarroll, and Jan O Korbel (2011). "Mapping copy number variation by population-scale genome sequencing." In: *Nature* 470, pp. 59–65. ISSN: 0028-0836. DOI: 10.1038/nature09708.

Morin, Ryan, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor Pugh, Helen McDonald, Richard Varhol, Steven Jones, and Marco Marra (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." In: *BioTechniques* 45.1, pp. 81–94. ISSN: 0736-6205. DOI: 10.2144/00011290.

Muller, H J (1928). "The Production of Mutations by X-Rays." In: *Proceedings of the National Academy of Sciences of the United States of America* 14.9, pp. 714–26. ISSN: 0027-8424.

Mullis, K B (1990). "The unusual origin of the polymerase chain reaction." In: *Scientific American* 262.4, pp. 56–61, 64–5. ISSN: 0036-8733.

Myers, E. W. (2000). "A Whole-Genome Assembly of Drosophila". In: *Science* 287.5461, pp. 2196–2204. ISSN: 00368075. DOI: 10.1126/science.287.5461.2196.

Narendra, V., P. P. Rocha, D. An, R. Raviram, J. A. Skok, E. O. Mazzoni, and D. Reinberg (2015). "CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation". In: *Science* 347.6225, pp. 1017–1021. ISSN: 0036-8075. DOI: 10.1126/science.1262088.

Nathans, J, T P Piantanida, R L Eddy, T B Shows, and D S Hogness (1986). "Molecular genetics of inherited variation in human color vision." In: *Science (New York, N.Y.)* 232.4747, pp. 203–10. ISSN: 0036-8075.

Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks, and Michael Wigler (2011). "Tumour evolution inferred by single-cell sequencing". In: *Nature* 472.7341, pp. 90–94. ISSN: 0028-0836. DOI: 10.1038/nature09807.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L. van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard (2012). "Spatial partitioning of the regulatory landscape of the X-inactivation centre". In: *Nature* 485.7398, pp. 381–385. ISSN: 0028-0836. DOI: 10.1038/nature11049. arXiv: NIHMS150003.

Northcott, Paul a., Catherine Lee, Thomas Zichner, Adrian M. Stütz, Serap Erkek, Daisuke Kawauchi, David J. H. Shih, Volker Hovestadt, Marc Zapatka, Dominik Sturm, David T. W. Jones, Marcel Kool, Marc Remke, Florence M. G. Cavalli, Scott Zuyderduyn, Gary D. Bader, Scott VandenBerg, Lourdes Adriana Esparza, Marina Ryzhova, Wei Wang, Andrea Wittmann, Sebastian Stark, Laura Sieber, Huriye Seker-Cin, Linda Linke, Fa-

bian Kratochwil, Natalie Jäger, Ivo Buchhalter, Charles D. Imbusch, Gideon Zipprich, Benjamin Raeder, Sabine Schmidt, Nicolle Diessl, Stephan Wolf, Stefan Wiemann, Benedikt Brors, Chris Lawerenz, Jürgen Eils, Hans-Jörg Warnatz, Thomas Risch, Marie-Laure Yaspo, Ursula D. Weber, Cynthia C. Bartholomae, Christof von Kalle, Eszter Turányi, Peter Hauser, Emma Sanden, Anna Darabi, Peter Siesjö, Jaroslav Sterba, Karel Zitterbart, David Sumerauer, Peter van Sluis, Rogier Versteeg, Richard Volckmann, Jan Koster, Martin U. Schuhmann, Martin Ebinger, H. Leighton Grimes, Giles W. Robinson, Amar Gajjar, Martin Mynarek, Katja von Hoff, Stefan Rutkowski, Torsten Pietsch, Wolfram Scheurlen, Jörg Felsberg, Guido Reifenberger, Andreas E. Kulozik, Andreas von Deimling, Olaf Witt, Roland Eils, Richard J. Gilbertson, Andrey Korshunov, Michael D. Taylor, Peter Lichter, Jan O. Korbel, Robert J. Wechsler-Reya, and Stefan M. Pfister (2014). "Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma". In: *Nature*. ISSN: 0028-0836. DOI: 10.1038/nature13379.

Novak, Adam M, Glenn Hickey, Erik Garrison, Sean Blum, Abram Connelly, Alexander Dilthey, Jordan Eizenga, M. A. Saleh Elmohamed, Sally Guthrie, André Kahles, Stephen Keenan, Jerome Kelleher, Deniz Kural, Heng Li, Michael F Lin, Karen Miga, Nancy Ouyang, Goran Rakocevic, Maciek Smuga-Otto, Alexander Wait Zaranek, Richard Durbin, Gil McVean, David Haussler, and Benedict Paten (2017). "Genome Graphs". In: *bioRxiv*. DOI: 10.1101/101378. eprint: https://www.biorxiv.org/content/early/2017/01/18/101378.full.pdf.

Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler (2004). "Circular binary segmentation for the analysis of array-based DNA copy number data". In: *Biostatistics* 5.4, pp. 557–572. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxh008.

Onishi-Seebacher, Megumi and Jan O Korbel (2011). "Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond." In: *BioEssays : news and reviews in molecular, cellular and developmental biology* 33.11, pp. 840–50. ISSN: 1521-1878. DOI: 10.1002/bies.201100075.

Oster, I I (1956). "A new crossing-over suppressor in chromosome 2 effective in the presence of heterologous inversions". In: *Drosophila Information Service* 30, p. 145.

Ott, Jurg, Jing Wang, and Suzanne M Leal (2015). "Genetic linkage analysis in the age of whole-genome sequencing." In: *Nature reviews. Genetics* 16.5, pp. 275–84. ISSN: 1471-0064. DOI: 10.1038/nrg3908.

Pendleton, Matthew, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi-Yang Fritz, Han Cao, Ariella Cohain, Gintaras Deikus, Russell E Durrett, Scott C Blanchard, Roger Altman, Chen-Shan Chin, Yan Guo, Ellen E Paxinos, Jan O Korbel, Robert B Darnell, W Richard McCombie, Pui-Yan Kwok, Christopher E Mason, Eric E Schadt, and Ali Bashir (2015). "Assembly and diploid architecture of an individual human genome via single-molecule technologies". In: *Nature Methods* 12.8, pp. 780–786. ISSN: 1548-7091. DOI: 10.1038/nmeth.3454.

Perry, George H, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, Fernando a Villanea, Joanna L Mountain, Rajeev Misra, Nigel P Carter, Charles Lee, and Anne C Stone (2007). "Diet and the evolution of hu-

man amylase gene copy number variation." In: *Nature genetics* 39.10, pp. 1256–1260. ISSN: 1061-4036. DOI: 10.1038/ng2123.

Pirinen, Matti, Tuuli Lappalainen, Noah A. Zaitlen, Emmanouil T. Dermitzakis, Peter Donnelly, Mark I. McCarthy, and Manuel A. Rivas (2015). "Assessing allele-specific expression across multiple tissues from RNA-seq read data". In: *Bioinformatics* 31.15, pp. 2497–2504. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv074.

Pope, Benjamin D., Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L. Vera, Yanli Wang, R. Scott Hansen, Theresa K. Canfield, Robert E. Thurman, Yong Cheng, Günhan Gülsoy, Jonathan H. Dennis, Michael P. Snyder, John A. Stamatoyannopoulos, James Taylor, Ross C. Hardison, Tamer Kahveci, Bing Ren, and David M. Gilbert (2014). "Topologically associating domains are stable units of replication-timing regulation". In: *Nature* 515.7527, pp. 402–405. ISSN: 14764687. DOI: 10.1038/nature13986. arXiv: 15334406.

Porubsky, David, Ashley D Sanders, Niek Van Wietmarschen, Ester Falconer, Mark Hills, Diana C J Spierings, Marianna R Bevova, Victor Guryev, and Peter M Lansdorp (2016). "Direct chromosome-length haplotyping by single cell sequencing". In: *Genome Research*, pp. 1565–1574. ISSN: 1088-9051. DOI: 10.1101/gr.209841.116.26.

Putnam, Nicholas H, Brendan L O'Connell, Jonathan C Stites, Brandon J Rice, Marco Blanchette, Robert Calef, Christopher J Troll, Andrew Fields, Paul D Hartley, Charles W Sugnet, David Haussler, Daniel S Rokhsar, and Richard E Green (2016). "Chromosome-scale shotgun assembly using an in vitro method for long-range linkage." In: *Genome research* 26.3, pp. 342–50. ISSN: 1549-5469. DOI: 10.1101/gr.193474.115.

Ramírez, Fidel, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke (2018). "High-resolution TADs reveal DNA sequences underlying genome organization in flies". In: *Nature Communications* 9.1, p. 189. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02525-w.

Ramírez, Fidel, Thomas Lingg, Sarah Toscano, Kin Chung Lam, Plamen Georgiev, Ho-Ryun Chung, Bryan R Lajoie, Elzo de Wit, Ye Zhan, Wouter de Laat, Job Dekker, Thomas Manke, and Asifa Akhtar (2015). "High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila." In: *Molecular cell* 60.1, pp. 146–62. ISSN: 1097-4164. DOI: 10.1016/j.molcel.2015.08.024.

Rao, Suhas S.P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden (2014). "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". In: *Cell* 159.7, pp. 1665–1680. ISSN: 00928674. DOI: 10.1016/j.cell.2014.11.021.

Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel (2012). "DELLY: structural variant discovery by integrated paired-end and split-read analysis." In: *Bioinformatics (Oxford, England)* 28.18, pp. i333–i339. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bts378.

Razzaghian, Hamid Reza, Mehdi Hayat Shahi, Lars A Forsberg, Teresita Diaz de Ståhl, Devin Absher, Niklas Dahl, Maxwell P Westerman, and Jan P Dumanski (2010). "So-

matic mosaicism for chromosome X and Y aneuploidies in monozygotic twins heterozygous for sickle cell disease mutation." In: *American journal of medical genetics. Part A* 152A.10, pp. 2595–8. ISSN: 1552-4833. DOI: 10.1002/ajmg.a.33604.

Redon, Richard, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L Freeman, Juan R González, Mònica Gratacòs, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R MacDonald, Christian R Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P Carter, Hiroyuki Aburatani, Charles Lee, Keith W Jones, Stephen W Scherer, and Matthew E Hurles (2006). "Global variation in copy number in the human genome." In: *Nature* 444.7118, pp. 444–54. ISSN: 1476-4687. DOI: 10.1038/nature05329.

Regev, Aviv, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants (2017). "The Human Cell Atlas." In: *eLife* 6. ISSN: 2050-084X. DOI: 10.7554/eLife.27041.

Rhoads, Anthony and Kin Fai Au (2015). "PacBio Sequencing and Its Applications". In: *Genomics, Proteomics & Bioinformatics* 13.5, pp. 278–289. ISSN: 16720229. DOI: 10.1016/j.gpb.2015.08.002.

Rickman, D. S., T. D. Soong, B. Moss, J. M. Mosquera, J. Dlabal, S. Terry, T. Y. MacDonald, J. Tripodi, K. Bunting, V. Najfeld, F. Demichelis, A. M. Melnick, O. Elemento, and M. A. Rubin (2012). "Oncogene-mediated alterations in chromatin conformation". In: *Proceedings of the National Academy of Sciences* 109.23, pp. 9083–9088. ISSN: 0027-8424. DOI: 10.1073/pnas.1112570109.

Rizk, G., A. Gouin, R. Chikhi, and C. Lemaitre (2014). "MindTheGap: integrated detection and assembly of short and long insertions". In: *Bioinformatics* 30.24, pp. 3451–3457. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu545.

Roller, Eric, Sergii Ivakhno, Steve Lee, Thomas Royce, and Stephen Tanner (2016). "Canvas: versatile and scalable detection of copy number variants". In: *Bioinformatics* 32.15, pp. 2375–2377. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw163.

Romanel, Alessandro, Sara Lago, Davide Prandi, Andrea Sboner, and Francesca Demichelis (2015). "ASEQ: fast allele-specific studies from next-generation sequencing data". In: *BMC Medical Genomics* 8.1, p. 9. ISSN: 1755-8794. DOI: 10.1186/s12920-015-0084-2.

Rozowsky, Joel, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, Nitin Bhardwaj, Mark Rubin, Michael Snyder, and Mark Gerstein (2014). "AlleleSeq: analysis of allele-specific expression and binding in a network framework". In: *Molecular Systems Biology* 7.1, pp. 522–522. ISSN: 1744-4292. DOI: 10.1038/msb.2011.54.

Rudewicz, Justine, Hayssam Soueidan, Raluca Uricaru, Hervé Bonnefoi, Richard Iggo, Jonas Bergh, and Macha Nikolski (2016). "MICADo - Looking for Mutations in Targeted PacBio Cancer Data: An Alignment-Free Method." In: *Frontiers in genetics* 7, p. 214. ISSN: 1664-8021. DOI: 10.3389/fgene.2016.00214.

Ruiz-Velasco, Mariana, Manjeet Kumar, Mang Ching Lai, Pooja Bhat, Ana Belen Solis-Pinson, Alejandro Reyes, Stefan Kleinsorg, Kyung-Min Noh, Toby J. Gibson, and Judith B. Zaugg (2017). "CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals". In: *Cell Systems* 5.6, 628–637.e6. ISSN: 24054712. DOI: 10.1016/j.cels.2017.10.018.

Ruiz-Velasco, Mariana and Judith B. Zaugg (2017). "Structure meets function: How chromatin organisation conveys functionality". In: *Current Opinion in Systems Biology* 1.i, pp. 129–136. ISSN: 24523100. DOI: 10.1016/j.coisb.2017.01.003.

Saini, Natalie, Steven A. Roberts, Leszek J. Klimczak, Kin Chan, Sara A. Grimm, Shuangshuang Dai, David C. Fargo, Jayne C. Boyer, William K. Kaufmann, Jack A. Taylor, Eunjung Lee, Isidro Cortes-Ciriano, Peter J. Park, Shepherd H. Schurman, Ewa P. Malc, Piotr A. Mieczkowski, and Dmitry A. Gordenin (2016). "The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts". In: *PLOS Genetics* 12.10. Ed. by Martin Taylor, e1006385. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006385.

Sanders, Ashley D, Ester Falconer, Mark Hills, Diana C J Spierings, and Peter M Lansdorp (2017). "Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs". In: *Nature Protocols* 12.6, pp. 1151–1176. ISSN: 1754-2189. DOI: 10.1038/nprot.2017.029.

Sanders, Ashley D, Mark Hills, David Porubský, Victor Guryev, Ester Falconer, and Peter M. Lansdorp (2016). "Characterizing polymorphic inversions in human genomes by single-cell sequencing". In: *Genome Research* 26.11, pp. 1575–1587. ISSN: 1088-9051. DOI: 10.1101/gr.201160.115.

Sanger, F, S Nicklen, and A R Coulson (1977). "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12, pp. 5463–7. ISSN: 0027-8424.

Saxena, R. K., D. Edwards, and R. K. Varshney (2014). "Structural variations in plant genomes". In: *Briefings in Functional Genomics* 13.4, pp. 296–307. ISSN: 2041-2649. DOI: 10.1093/bfgp/elu016.

Schmitt, Anthony D., Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L. Barr, and Bing Ren (2016). "A Compendium

of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome". In: *Cell Reports* 17.8, pp. 2042–2059. ISSN: 22111247. DOI: 10.1016/j.celrep.2016.10.061.

Schwartz, D C, X Li, L I Hernandez, S P Ramnarain, E J Huff, and Y K Wang (1993). "Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping." In: *Science (New York, N.Y.)* 262.5130, pp. 110–4. ISSN: 0036-8075.

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.-H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.-C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler (2007). "Strong Association of De Novo Copy Number Mutations with Autism". In: *Science* 316.5823, pp. 445–449. ISSN: 0036-8075. DOI: 10.1126/science.1138659.

Sebat, Jonathan, B Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy, James Hicks, Kenny Ye, Andrew Reiner, T Conrad Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler (2004). "Large-scale copy number polymorphism in the human genome." In: *Science (New York, N.Y.)* 305.5683, pp. 525–8. ISSN: 1095-9203. DOI: 10.1126/science.1098918.

Selvaraj, Siddarth, Jesse R Dixon, Vikas Bansal, and Bing Ren (2013). "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing". In: *Nature Biotechnology* 31.12, pp. 1111–1118. ISSN: 1087-0156. DOI: 10.1038/nbt.2728.

*Seq, by Lior Pachter (website) (2018). *A list of *Seq assays*. Website, accessed 12 Frebruary 2018. URL: https://liorpachter.wordpress.com/seq/.

Sexton, Tom and Giacomo Cavalli (2015). "The Role of Chromosome Domains in Shaping the Functional Genome". In: *Cell* 160.6, pp. 1049–1059. ISSN: 00928674. DOI: 10.1016/j.cell.2015.02.040.

Sexton, Tom, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli (2012). "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome". In: *Cell* 148.3, pp. 458–472. ISSN: 00928674. DOI: 10.1016/j.cell.2012.01.010.

Sharp, Andrew J, Devin P Locke, Sean D McGrath, Ze Cheng, Jeffrey A Bailey, Rhea U Vallente, Lisa M Pertz, Royden A Clark, Stuart Schwartz, Rick Segraves, Vanessa V Oseroff, Donna G Albertson, Daniel Pinkel, and Evan E Eichler (2005). "Segmental duplications and copy-number variation in the human genome." In: *American journal of human genetics* 77.1, pp. 78–88. ISSN: 0002-9297. DOI: 10.1086/431652.

Shen, Yin, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V. Lobanenkov, and Bing Ren (2012). "A map of the cis-regulatory sequences in the mouse genome". In: *Nature* 488.7409, pp. 116–120. ISSN: 0028-0836. DOI: 10.1038/nature11243.

Shendure, Jay and Hanlee Ji (2008). "Next-generation DNA sequencing". In: *Nature Biotechnology* 26.10, pp. 1135–1145. ISSN: 1087-0156. DOI: 10.1038/nbt1486.

Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat (2006). "Nuclear organization of active and in-

active chromatin domains uncovered by chromosome conformation capture–on-chip (4C)". In: *Nature Genetics* 38.11, pp. 1348–1354. ISSN: 1061-4036. DOI: 10.1038/ng1896.

Skelly, Daniel a., Marnie Johansson, Jennifer Madeoy, Jon Wakefield, and Joshua M. Akey (2011). "A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data". In: *Genome Research* 21.10, pp. 1728–1737. ISSN: 1088-9051. DOI: 10.1101/gr.119784.110.

Smallwood, Andrea and Bing Ren (2013). "Genome organization and long-range regulation of gene expression by enhancers". In: *Current Opinion in Cell Biology* 25.3, pp. 387–394. ISSN: 09550674. DOI: 10.1016/j.ceb.2013.02.005.

Snijders, Antoine M., Norma Nowak, Richard Segraves, Stephanie Blackwood, Nils Brown, Jeffrey Conroy, Greg Hamilton, Anna Katherine Hindle, Bing Huey, Karen Kimura, Sindy Law, Ken Myambo, Joel Palmer, Bauke Ylstra, Jingzhu Pearl Yue, Joe W. Gray, Ajay N. Jain, Daniel Pinkel, and Donna G. Albertson (2001). "Assembly of microarrays for genome-wide measurement of DNA copy number". In: *Nature Genetics* 29.3, pp. 263–264. ISSN: 1061-4036. DOI: 10.1038/ng754.

Solares, Edwin A, Mahul Chakraborty, Danny E Miller, Shannon Kalsow, Kate E Hall, Anoja G Perera, J J Emerson, and R Scott Hawley (2018). "Rapid low-cost assembly of the Drosophila melanogaster reference genome using low-coverage, long-read sequencing". In: *bioRxiv*. URL: http://biorxiv.org/content/early/2018/02/18/267401.abstract.

Speicher, Michael R. and Nigel P. Carter (2005). "The new cytogenetics: blurring the boundaries with molecular biology". In: *Nature Reviews Genetics* 6.10, pp. 782–792. ISSN: 1471-0056. DOI: 10.1038/nrg1692.

Spencer, David H, Manoj Tyagi, Francesco Vallania, Andrew J Bredemeyer, John D Pfeifer, Rob D Mitra, and Eric J Duncavage (2014). "Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data." In: *The Journal of molecular diagnostics : JMD* 16.1, pp. 75–88. ISSN: 1943-7811. DOI: 10.1016/j.jmoldx.2013.09.003.

Stankiewicz, Paweł and James R Lupski (2010). "Structural variation in the human genome and its role in disease." In: *Annual review of medicine* 61, pp. 437–55. ISSN: 1545-326X. DOI: 10.1146/annurev-med-100708-204735.

Stegle, Oliver, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin (2012). "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses". In: *Nature Protocols* 7.3, pp. 500–507. ISSN: 1754-2189. DOI: 10.1038/nprot.2011.457.

Steinberg, Karyn Meltz, Valerie A. Schneider, Tina A. Graves-Lindsay, Robert S. Fulton, Richa Agarwala, John Huddleston, Sergey A. Shiryev, Aleksandr Morgulis, Urvashi Surti, Wesley C. Warren, Deanna M. Church, Evan E. Eichler, and Richard K. Wilson (2014). "Single haplotype assembly of the human genome from a hydatidiform mole". In: *Genome Research* 24.12, pp. 2066–2076. ISSN: 1088-9051. DOI: 10.1101/gr.180893.114.

Strimmer, K. (2008). "fdrtool: a versatile R package for estimating local and tail area-based false discovery rates". In: *Bioinformatics* 24.12, pp. 1461–1462. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn209.

Sudmant, Peter H, S Mallick, B J Nelson, F Hormozdiari, N Krumm, J Huddleston, B P Coe, C Baker, S Nordenfelt, M Bamshad, L B Jorde, O L Posukh, H Sahakyan, W S Watkins, L Yepiskoposyan, M S Abdullah, C M Bravi, C Capelli, T Hervig, J T S Wee, C Tyler-Smith, G van Driem, I G Romero, A R Jha, S Karachanak-Yankova, D Toncheva, D Comas, B Henn, T Kivisild, A Ruiz-Linares, A Sajantila, E Metspalu, J Parik, R Villems, E B Starikovskaya, G Ayodo, C M Beall, A Di Rienzo, M Hammer, R Khusainova, E Khusnutdinova, W Klitz, C Winkler, D Labuda, M Metspalu, S A Tishkoff, S Dryomov, R Sukernik, N Patterson, D Reich, and E E Eichler (2015). "Global diversity, population stratification, and selection of human copy number variation". In: *Science*, science.aab3761–. ISSN: 0036-8075. DOI: 10.1126/science.aab3761.

Sudmant, Peter H, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, Miriam K Konkel, Ankit Malhotra, Adrian M Stütz, Xinghua Shi, Francesco Paolo Casale, Jieming Chen, Fereydoun Hormozdiari, Gargi Dayama, Ken Chen, Maika Malig, Mark J P Chaisson, Klaudia Walter, Sascha Meiers, Seva Kashin, Erik Garrison, Adam Auton, Hugo Y K Lam, Xinmeng Jasmine Mu, Can Alkan, Danny Antaki, Taejeong Bae, Eliza Cerveira, Peter Chines, Zechen Chong, Laura Clarke, Elif Dal, Li Ding, Sarah Emery, Xian Fan, Madhusudan Gujral, Fatma Kahveci, Jeffrey M. Kidd, Yu Kong, Eric-Wubbo Lameijer, Shane McCarthy, Paul Flicek, Richard A Gibbs, Gabor Marth, Christopher E Mason, Androniki Menelaou, Donna M Muzny, Bradley J Nelson, Amina Noor, Nicholas F Parrish, Matthew Pendleton, Andrew Quitadamo, Benjamin Raeder, Eric E. Schadt, Mallory Romanovitch, Andreas Schlattl, Robert Sebra, Andrey A Shabalin, Andreas Untergasser, Jerilyn A Walker, Min Wang, Fuli Yu, Chengsheng Zhang, Jing Zhang, Xiangqun Zheng-Bradley, Wanding Zhou, Thomas Zichner, Jonathan Sebat, Mark A Batzer, Steven A McCarroll, Ryan E Mills, Mark B Gerstein, Ali Bashir, Oliver Stegle, Scott E Devine, Charles Lee, Evan E Eichler, and Jan O Korbel (2015). "An integrated map of structural variation in 2,504 human genomes". In: *Nature* 526.7571, pp. 75–81. ISSN: 0028-0836. DOI: 10.1038/nature15394.

Sulem, Patrick, Hannes Helgason, Asmundur Oddson, Hreinn Stefansson, Sigurjon A Gudjonsson, Florian Zink, Eirikur Hjartarson, Gunnar Th Sigurdsson, Adalbjorg Jonasdottir, Aslaug Jonasdottir, Asgeir Sigurdsson, Olafur Th Magnusson, Augustine Kong, Agnar Helgason, Hilma Holm, Unnur Thorsteinsdottir, Gisli Masson, Daniel F Gudbjartsson, and Kari Stefansson (2015). "Identification of a large set of rare complete human knockouts". In: *Nature Genetics* 47.5, pp. 448–452. ISSN: 1061-4036. DOI: 10.1038/ng.3243.

Svartman, Marta, Gary Stone, and Roscoe Stanyon (2005). "Molecular cytogenetics discards polyploidy in mammals". In: *Genomics* 85.4, pp. 425–430. ISSN: 08887543. DOI: 10.1016/j.ygeno.2004.12.004.

Swaminathan, G. J., E. Bragin, E. A. Chatzimichali, M. Corpas, A. P. Bevan, C. F. Wright, N. P. Carter, M. E. Hurles, and H. V. Firth (2012). "DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders". In: *Human Molecular Genetics* 21.R1, R37–R44. ISSN: 0964-6906. DOI: 10.1093/hmg/dds362.

Symmons, Orsolya, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz (2014). "Functional and topological characteristics of mammalian regulatory domains". In: *Genome Research* 24.3, pp. 390–400. ISSN: 15495469. DOI: 10.1101/gr.163519.113.

Tadros, W. and H. D. Lipshitz (2009). "The maternal-to-zygotic transition: a play in two acts". In: *Development* 136.18, pp. 3033–3042. ISSN: 0950-1991. DOI: 10.1242/dev.033183.

Tan, Adrian, Gonçalo R. Abecasis, and Hyun Min Kang (2015). "Unified representation of genetic variants". In: *Bioinformatics* 31.13, pp. 2202–2204. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btv112.

Tattini, Lorenzo, Romina D'Aurizio, and Alberto Magi (2015). "Detection of Genomic Structural Variants from Next-Generation Sequencing Data". In: *Frontiers in Bioengineering and Biotechnology* 3.6, pp. 1–8. ISSN: 2296-4185. DOI: 10.3389/fbioe.2015.00092.

Teague, B., M. S. Waterman, S. Goldstein, K. Potamousis, S. Zhou, S. Reslewic, D. Sarkar, A. Valouev, C. Churas, J. M. Kidd, S. Kohn, R. Runnheim, C. Lamers, D. Forrest, M. A. Newton, E. E. Eichler, M. Kent-First, U. Surti, M. Livny, and D. C. Schwartz (2010). "High-resolution human genome structure by single-molecule analysis". In: *Proceedings of the National Academy of Sciences* 107.24, pp. 10848–10853. ISSN: 0027-8424. DOI: 10.1073/pnas.0914638107.

Telenti, Amalio, Levi C. T. Pierce, William H. Biggs, Julia di Iulio, Emily H. M. Wong, Martin M. Fabani, Ewen F. Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C. Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A. Perkins, Franz J. Och, Yaron Turpaz, and J. Craig Venter (2016). "Deep sequencing of 10,000 human genomes". In: *Proceedings of the National Academy of Sciences* 113.42, pp. 11901–11906. ISSN: 0027-8424. DOI: 10.1073/pnas.1613365113.

Templado, C., L. Uroz, and A. Estop (2013). "New insights on the origin and relevance of aneuploidy in human spermatozoa". In: *MHR: Basic science of reproductive medicine* 19.10, pp. 634–643. ISSN: 1460-2407. DOI: 10.1093/molehr/gat039.

Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov (2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". In: *Briefings in Bioinformatics* 14.2, pp. 178–192. ISSN: 1467-5463. DOI: 10.1093/bib/bbs017. URL: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs017.

Tijo, Joe Hin and Albert Levan (1956). "THE CHROMOSOME NUMBER OF MAN". In: *Hereditas* 42.1-2, pp. 1–6. ISSN: 00180661. DOI: 10.1111/j.1601-5223.1956.tb03010.x. URL: http://doi.wiley.com/10.1111/j.1601-5223.1956.tb03010.x%20http://www.ncbi.nlm.nih.gov/pubmed/345813.

Tinderholt, I I (1960). "New Mutants (TM3, Ser: Third Mulitple 3, Serrate)". In: *Drosophila Information Service* 34, pp. 53–54.

Travers, K. J., C.-S. Chin, D. R. Rank, J. S. Eid, and S. W. Turner (2010). "A flexible and efficient template format for circular consensus sequencing and SNP detection". In: *Nucleic Acids Research* 38.15, e159–e159. ISSN: 0305-1048. DOI: 10.1093/nar/gkq543.

Turcatti, Gerardo, Anthony Romieu, Milan Fedurco, and Ana-Paula Tairi (2008). "A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible ter-

minators for DNA sequencing by synthesis." In: *Nucleic acids research* 36.4, e25. ISSN: 1362-4962. DOI: 10.1093/nar/gkn021.

Tuzun, Eray, Andrew J Sharp, Jeffrey A Bailey, Rajinder Kaul, V Anne Morrison, Lisa M Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, Maynard V Olson, and Evan E Eichler (2005). "Fine-scale structural variation of the human genome". In: *Nature Genetics* 37.7, pp. 727–732. ISSN: 1061-4036. DOI: 10.1038/ng1562.

Tyson, John R, Nigel J O'Neill, Miten Jain, Hugh E Olsen, Philip Hieter, and Terrance P Snutch (2017). "Whole genome sequencing and assembly of a Caenorhabditis elegans genome with complex genomic rearrangements using the MinION sequencing device". In: *bioRxiv*. URL: http://biorxiv.org/content/early/2017/01/08/099143.abstract.

Uemura, Sotaro, Colin Echeverría Aitken, Jonas Korlach, Benjamin A. Flusberg, Stephen W. Turner, and Joseph D. Puglisi (2010). "Real-time tRNA transit on single translating ribosomes at codon resolution". In: *Nature* 464.7291, pp. 1012–1017. ISSN: 0028-0836. DOI: 10.1038/nature08925.

UK10K Consortium, The (2015). "The UK10K project identifies rare variants in health and disease". In: *Nature* 526.7571, pp. 82–90. ISSN: 0028-0836. DOI: 10.1038/nature14962.

Untergasser, Andreas, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen (2012). "Primer3—new capabilities and interfaces". In: *Nucleic Acids Research* 40.15, e115–e115. ISSN: 1362-4962. DOI: 10.1093/nar/gks596.

Van de Geijn, Bryce, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard (2015). "WASP: allele-specific software for robust molecular quantitative trait locus discovery". In: *Nature Methods* September. ISSN: 1548-7091. DOI: 10.1038/nmeth.3582.

Van Loo, P., S. H. Nordgard, O. C. Lingjaerde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Borresen-Dale, and V. N. Kristensen (2010). "Allele-specific copy number analysis of tumors". In: *Proceedings of the National Academy of Sciences* 107.39, pp. 16910–16915. ISSN: 0027-8424. DOI: 10.1073/pnas.1009843107.

Veeramah, Krishna R. and Michael F. Hammer (2014). "The impact of whole-genome sequencing on the reconstruction of human population history". In: *Nature Reviews Genetics* 15.3, pp. 149–162. ISSN: 1471-0056. DOI: 10.1038/nrg3625.

Venkatraman, E. S. and A. B. Olshen (2007). "A faster circular binary segmentation algorithm for the analysis of array CGH data". In: *Bioinformatics* 23.6, pp. 657–663. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl646.

Venter, J. C. (2001). "The Sequence of the Human Genome". In: *Science* 291.5507, pp. 1304–1351. ISSN: 00368075. DOI: 10.1126/science.1058040. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1058040.

Volik, Stanislav, Shaying Zhao, Koei Chin, John H Brebner, David R Herndon, Quanzhou Tao, David Kowbel, Guiqing Huang, Anna Lapuk, Wen-Lin Kuo, Gregg Magrane, Pieter De Jong, Joe W Gray, and Colin Collins (2003). "End-sequence profiling: sequence-based analysis of aberrant genomes." In: *Proceedings of the National Academy of Sciences of the United States of America* 100.13, pp. 7696–701. ISSN: 0027-8424. DOI: 10.1073/pnas.1232418100.

Wang, Min, Christine R Beck, Adam C English, Qingchang Meng, Christian Buhay, Yi Han, Harsha V Doddapaneni, Fuli Yu, Eric Boerwinkle, James R Lupski, Donna M Muzny, and Richard A Gibbs (2015). "PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations." In: *BMC genomics* 16, p. 214. ISSN: 1471-2164. DOI: 10.1186/s12864-015-1370-2.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0056. DOI: 10.1038/nrg2484.

Ward, L (1923). "The Genetics of Curly Wing in Drosophila. Another Case of Balanced Lethal Factors." In: *Genetics* 8.3, pp. 276–300. ISSN: 0016-6731. URL: http://www.ncbi.nlm.nih.gov/pubmed/17246014%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1200750.

Waszak, Sebastian M., Yehudit Hasin, Thomas Zichner, Tsviya Olender, Ifat Keydar, Miriam Khen, Adrian M. Stütz, Andreas Schlattl, Doron Lancet, and Jan O. Korbel (2010). "Systematic Inference of Copy-Number Genotypes from Personal Genome Sequencing Data Reveals Extensive Olfactory Receptor Gene Content Diversity". In: *PLoS Computational Biology* 6.11. Ed. by Wyeth W. Wasserman, e1000988. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000988.

Weber, James L. and Eugene W. Myers (1997). "Human Whole-Genome Shotgun Sequencing". In: *Genome Research* 7.5, pp. 401–409. ISSN: 1088-9051. DOI: 10.1101/gr.7.5.401.

Weese, D, A.-K. Emde, T Rausch, A. Doring, and K Reinert (2009). "RazerS–fast read mapping with sensitivity control". In: *Genome Research* 19.9, pp. 1646–1654. ISSN: 1088-9051. DOI: 10.1101/gr.088823.108.

Weischenfeldt, Joachim, Taronish Dubash, Alexandros P Drainas, Balca R Mardin, Yuanyuan Chen, Adrian M Stütz, Sebastian M Waszak, Graziella Bosco, Ann Rita Halvorsen, Benjamin Raeder, Theocharis Efthymiopoulos, Serap Erkek, Christine Siegl, Hermann Brenner, Odd Terje Brustugun, Sebastian M Dieter, Paul A Northcott, Iver Petersen, Stefan M Pfister, Martin Schneider, Steinar K Solberg, Erik Thunissen, Wilko Weichert, Thomas Zichner, Roman Thomas, Martin Peifer, Aslaug Helland, Claudia R Ball, Martin Jechlinger, Rocio Sotillo, Hanno Glimm, and Jan O Korbel (2016). "Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking". In: *Nature Genetics* 49.1, pp. 65–74. ISSN: 1061-4036. DOI: 10.1038/ng.3722.

Weischenfeldt, Joachim, Orsolya Symmons, Francois Spitz, and Jan O Korbel (2013). "Phenotypic impact of genomic structural variation: insights from and for human disease". In: *Nat Rev Genet* 14.2, pp. 125–138. ISSN: 1471-0056.

Wood, Andrew R et al. (2014). "Defining the role of common variation in the genomic and biological architecture of adult human height". In: *Nature Genetics* 46.11, pp. 1173–1186. ISSN: 1061-4036. DOI: 10.1038/ng.3097.

Xie, Chao and Martti T Tammi (2009). "CNV-seq, a new method to detect copy number variation using high-throughput sequencing". In: *BMC Bioinformatics* 10.1, p. 80. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-80.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning (2009). "Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from

paired-end short reads". In: *Bioinformatics* 25.21, pp. 2865–2871. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp394.

Yin, Tengfei, Dianne Cook, and Michael Lawrence (2012). "ggbio: an R package for extending the grammar of graphics for genomic data". In: *Genome Biology* 13.8, R77. ISSN: 1465-6906. DOI: 10.1186/gb-2012-13-8-r77.

Youssoufian, Hagop and Reed E. Pyeritz (2002). "Mechanisms and consequences of somatic mosaicism in humans". In: *Nature Reviews Genetics* 3.10, pp. 748–758. ISSN: 1471-0056. DOI: 10.1038/nrg906.

Zepeda-Mendoza, Cinthya J., Jonas Ibn-Salem, Tammy Kammin, David J. Harris, Debra Rita, Karen W. Gripp, Jennifer J. MacKenzie, Andrea Gropman, Brett Graham, Ranad Shaheen, Fowzan S. Alkuraya, Campbell K. Brasington, Edward J. Spence, Diane Masser-Frye, Lynne M. Bird, Erica Spiegel, Rebecca L. Sparkes, Zehra Ordulu, Michael E. Talkowski, Miguel A. Andrade-Navarro, Peter N. Robinson, and Cynthia C. Morton (2017). "Computational Prediction of Position Effects of Apparently Balanced Human Chromosomal Rearrangements". In: *The American Journal of Human Genetics* 101.2, pp. 206–217. ISSN: 00029297. DOI: 10.1016/j.ajhg.2017.06.011.

Zhang, Feng, Wenli Gu, Matthew E. Hurles, and James R. Lupski (2009). "Copy Number Variation in Human Health, Disease, and Evolution". In: *Annual Review of Genomics and Human Genetics* 10.1, pp. 451–481. ISSN: 1527-8204. DOI: 10.1146/annurev.genom.9.081307.164217.

Zhang, Feng, Mehrdad Khajavi, Anne M Connolly, Charles F Towne, Sat Dev Batish, and James R Lupski (2009). "The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans." In: *Nature genetics* 41.7, pp. 849–53. ISSN: 1546-1718. DOI: 10.1038/ng.399.

Zhang, Zhengdong D, Jiang Du, Hugo Lam, Alex Abyzov, Alexander E Urban, Michael Snyder, and Mark Gerstein (2011). "Identification of genomic indels and structural variations using split reads". In: *BMC Genomics* 12.1, p. 375. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-375.

Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang (2014). "CrossMap: a versatile tool for coordinate conversion between genome assemblies." In: *Bioinformatics (Oxford, England)* 30.7, pp. 1006–7. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt730.

Zhao, Zhihu, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti, and Rolf Ohlsson (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions." In: *Nature genetics* 38.11, pp. 1341–7. ISSN: 1061-4036. DOI: 10.1038/ng1891.

Zichner, Thomas, David A Garfield, Tobias Rausch, Adrian M Stütz, Enrico Cannavó, Martina Braun, Eileen E M Furlong, and Jan O Korbel (2013). "Impact of genomic structural variation in Drosophila melanogaster based on population-scale sequencing." In: *Genome research* 23.3, pp. 568–79. ISSN: 1549-5469. DOI: 10.1101/gr.142646.112.

Zong, C., S. Lu, A. R. Chapman, and X. S. Xie (2012). "Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell". In: *Science* 338.6114, pp. 1622–1626. ISSN: 0036-8075. DOI: 10.1126/science.1229164.