Inaugural dissertation

for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of the
Ruprecht – Karls – University
Heidelberg

Presented by
M.Sc Wolfram Gregor Alexander Höps
born in Erlangen, Germany
Oral examination: March 22nd, 2023

Genomic diversity associated with polymorphic inversions in humans and their close relatives

Referees:

Prof. Dr. Benedikt Brors
Prof. Dr. Oliver Stegle
Dr. Judith Zaugg
Prof. Dr. Christian Schaaf

Summary

Individuals of one species share the bulk of their genetic material, yet no two genomes are the same. Aside from displaying classical variation such as deletions, insertions, or substitutions of base pairs, two DNA segments can also differ in their orientation relative to the rest of their chromosomes. Such *inversions* are known for a range of biological implications and contribute critically to genome evolution and disease. However, inversions are notoriously challenging to detect, a fact which still impedes comprehensive analysis of their specific properties. This thesis describes several highly inter-connected projects aimed at identifying and functionally characterizing inversions present in the human population and related great ape species.

First, inversions between human and four great ape species were assessed for their potential to disrupt topologically associating domains (TADs), potentially prompting gene misregulation. TAD boundaries co-located with breakpoints of long inversions, and while disrupted TADs displayed elevated rates of differentially expressed genes, this effect could be attributed the vicinity to inversion breakpoints, suggesting overall robustness of gene expression in response to TAD disruption.

The second part of this thesis describes contributions to a collaborative project aimed at characterizing the full spectrum of inversions in 43 humans. In this study, I co-developed a novel inversion genotyping algorithm based on Strand-specific DNA sequencing and contributed to the description of 398 inversion polymorphisms. Inversions exhibited various underlying formation mechanisms, promotion of gene dysregulation, widespread recurrence, and association with genomic disease. These results suggest that long inversions are much more prominent in humans than previously thought, with at least 0.6% of the genome subject to inversion recurrence and, sometimes, the associated risk of subsequent deleterious mutation.

With a focus on the link between inversions and disease-causing copy number variations, the last project describes a novel algorithm to identify loci hit sequentially by several overlapping mutation events. This algorithm enabled the description of detailed mutation sequences in 20 highly dynamic regions in the human genome, and additional complex variants on chromosome Y. Six complex loci associate directly with a genomic disease, thereby highlighting in detail the intrinsic link between inversions and CNVs.

In summary, these projects provide novel insights into the landscape of inversions in humans and primates, which are much more frequent, and often more complex than previously thought. These findings provide a basis for future inversion studies and highlight the crucial contribution of this class of mutation to genome variation.

Zusammenfassung

Individuen einer Art teilen den Großteil ihres genetischen Materials, jedoch gleichen sich zwei Genome nie gänzlich. Neben klassischen Mutations wie gelöschten, eingefügten oder ausgetauschten Basenpaaren können sich zwei DNA-Abschnitte auch in ihrer Orientierung relativ zum Rest ihres Chromosoms unterscheiden. Solche *Inversionen* sind für eine Vielzahl biologischer Implikationen bekannt und tragen entscheidend zur Genomentwicklung und zum Entstehen von Erbkrankheiten bei. Eine umfassende Analyse ihrer spezifischen Eigenschaften war bisher technisch kaum umzusetzen, da Inversionen notorisch schwer zu detektieren sind. Sie zählen daher zu den am wenigsten untersuchten Klassen genetischer Variation überhaupt. Die vorliegende Dissertation beschreibt mehrere miteinander verwobene Projekte, die darauf abzielen, Inversionen zu identifizieren und funktionell zu charakterisieren, die in der menschlichen Population und in verwandten Menschenaffen vorkommen.

Zunächst untersucht wurden Inversionen zwischen Menschen und vier Menschenaffenarten auf ihre potenzielle Störung von topologisch assoziierenden Domänen (TADs). Frühere Studien legen nahe, dass solche Störungen eine Fehlregulation von Genen auslösen kann. Die Auswertung der Analysen zeigte, dass TAD-Grenzen mit Bruchpunkten langer Inversionen zusammenfallen, und während gestörte TADs erhöhte Raten differentiell exprimierter Gene aufweisen, kann dieser Effekt der Nähe zu Inversionsbruchpunkten zugeschrieben werden, was auf eine allgemeine Robustheit der Genexpression bezüglich TAD-disruptionen hindeutet.

Als nächstes werden Beiträge zu einem Gemeinschaftsprojekt beschrieben, das darauf abzielte, das gesamte Spektrum von Inversionen von 43 menschlichen Individuen zu charakterisieren. Im Verlauf dieser Studie habe ich einen neuartigen Algorithmus zur Genotypisierung von Inversionen, basierend auf Strangspezifischer DNA-Sequenzierung mitentwickelt und zur Beschreibung von 398 Inversionspolymorphismen beigetragen, die verschiedene zugrunde liegende Bildungsmechanismen, Förderung von Gen-Disregulation, weit verbreitete Rekurrenz und Assoziation mit genomischen Erkrankungen aufweisen. Diese Ergebnisse deuten darauf hin, dass lange Inversionen beim Menschen viel häufiger auftreten als bisher angenommen, wobei mindestens 0,6% des Genoms einem Risiko für rekurrente Inversionen und gelegentlich dem damit verbundenen Risiko einer nachfolgenden schädlichen Mutation ausgesetzt sind.

Das letzte Projekt konzentriert sich auf die Verknüpfung zwischen Inversionen und krankheitsverursachenden Deletionen und Duplikationen von genetischem Material und beschreibt einen neuartigen Algorithmus zur Identifizierung von Loci, die nacheinander von mehreren überlappenden Mutationsereignissen getroffen werden. Unter Verwendung dieses neuartigen Algorithmus werden detaillierte Mutationssequenzen in 20 hochdynamischen Regionen im menschlichen Genom beschrieben, von denen 6 direkt mit genomischen Erkrankungen assoziiert sind und dadurch die intrinsische Verbindung zwischen Inversionen und anderen Strukturvarianten hervorheben.

Zusammengenommen bieten diese Projekte neue Einblicke in die Landschaft genomischer Inversionen bei Menschen und Primaten, die sich als viel zahlreicher und oft komplexer erwiesen als bisher angenommen. Diese Ergebnisse bilden zusammen mit den dabei erzielten technologischen Fortschritten eine Grundlage für zukünftige Inversionsstudien und unterstreichen den entscheidenden Beitrag dieser Mutationsklasse zur Variabilität von Genomen.

Acknowledgments

First and foremost, I would like to thank my supervisor, Jan Korbel, for his guidance and support throughout my PhD journey. His expertise and insights have been invaluable to me, and I am deeply grateful for the time and effort he has dedicated to elevating me as a researcher. Many thanks also to Ashley Sanders, whose skill and expertise has impressed me deeply, and to who I am grateful for introducing me to the world of inversions – thank you for everything! I further thank my TAC members Oliver Stegle, Benedikt Brors and Isidro Cortés-Ciriano for ample support and many great suggestions.

My PhD experience was a very collaborative one, and I am grateful for many inspiring people who I got to work with and learn from over the years. This includes David Proubsky, who has been a steady source of creative inspration, and whose enthusiasm for science is unparalleled. Big thanks also to Hufsah Ashraf, my partner-in-crime during some of the most turbulent times of our PhDs – I'm looking forward to designing new tools with you in the future! I thank Pille Hallast for introducing me to the strange biology of the Y chromsome. I feel grateful towards the whole HGSVC consortium, whose members have provided a framework for most of the research represented here, and which is skillfully co-led by Jan Korbel, Charles Lee, Tobias Marschall and Evan Eichler. Furthermore, I want to thank Fritz Sedlazeck for his dedication in supporting me professionally and personally.

I also want to express my gratitude to all (past and present) members of the Korbel lab. Their friendship and collaboration have made my time in the lab enjoyable and productive. I have learned a lot from them, and I am grateful for the experience of working with such a talented and dedicated team. Special thanks to Nina Habermann and Martina Peskoller-Fuchs, two good souls of EMBL who I could always rely on in times of need.

I salute the valued members of the *therapy couch*™: Fergus, Anna, and Gilberto, whose paintings I get to admire daily and whose Mario Kart skills seem to have miraculously stagnated despite playing for three years. Thanks for sticking around with me! Finally, I want to thank Jana and my family for their unwavering support and encouragement over the past years.

My time at EMBL has been an incredibly challenging but rewarding experience, and I am grateful to all people who have supported me along the way.

Abbreviations

BIR break-induced replication

bp base pair

BrdU 5-Bromo-2'-deoxyuridine

CGR complex genomic rearrangement

CNV copy number variant

DSB double-strand break

FACS Fluorescence-activated cell sorting

FISH fluorescence in situ hybridization

FoSTeS fork stalling and template switching

HGSVC Human Genome Structural Variation Consortium

LCR low-copy repeat

LD linkage disequilibrium

MMBIR microhomology-mediated break-induced replication

MNase micrococcal nuclease enzyme

NAHR non-allelic homologous recombination

NHEJ non-allelic homologous end joining

NHP non-human primate

ONT Oxford Nanopore technologies

PacBio Pacific Biosciences

PAR pseudo-autosomal region

PCR polymerase chain reaction

SD segmental duplication

SNP single nucleotide polymorphism

SRS serial replication slippage

Strand-Seq Single-cell DNA template strand sequencing

SV structural variant

TAD topologically associating domain

WGS whole genome sequencing

Contents

1

General Introduction

In April 2022, a team of more than ninety international scientists from the 'Telomere-to-Telomere' Consortium published the first *complete* sequence of a human genome, in which they identified and allocated each of the > 3 billion DNA base pairs (bps) which constitute the blueprint for our species [Nurk et al., 2022]. This announcement was widely regarded as the *definitive* endpoint of the *Human Genome Project*, which started in 1989 with the very same goal of deciphering the human genome. Surprisingly, in 2001, the initial draft of the human genome encompassed already 94% of the human genetic code (Fig. 1.1) [International Human Genome Sequencing Consortium, 2001]. Many commentaries of the time assumed that resolving the few remaining unresolved regions would be a matter of few years – including the authors of the original draft, who confidently stated that "All chromosomes should be essentially completed by 2003, if not sooner" [International Human Genome Sequencing Consortium, 2001]. In hindsight, such early optimism reflects how vastly the complexity of 'complex' DNA regions has been underestimated initially. Likewise, the delayed completion of the human genome – 2 vs. 20 years – is a testimony to the amount of technological and conceptual advances that were necessary to complete the task.

However, no two genomes are the same. While the completion of *one* human genome represents a milestone for the field, it is the *variability* of genome content that produces the diversity of traits observed across individuals. Indeed, advances in DNA sequencing technology continue to provide an ever more comprehensive picture of how genomes differ from each other – both between individuals of the same species (e.g., humans [Ebert et al., 2021]) and across species (e.g., humans vs. primates [Suntsova and Buzdin, 2020]).

Mutations that affect many base pairs at once – so-called structural variants (SVs) – are the predominant source of genomic variation in a typical human

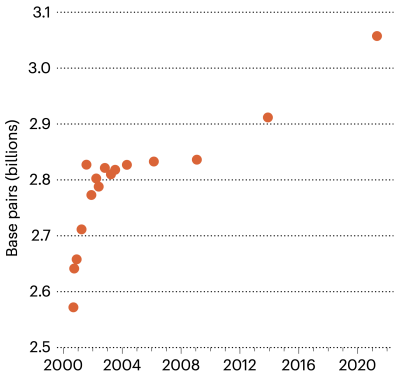

Figure 1.1: Number of resolved basepairs in genome assemblies since the release of the first human reference genome. Chromosome Y and mitochondrial DNA are excluded. Figure adapted from [Reardon, 2021].

genome, affecting more base pairs than single-base substitutions [Chaisson et al., 2019]. Consequently, structural variants have emerged as critical factors in human genetic variation [Ebert et al., 2021], adaptive evolution, and speciation [Perry et al., 2008, Zhang et al., 2021, Wellenreuther et al., 2019], while similarly acting as the predominant class of driver mutation in most cancers [Cosenza et al., 2022] and genetic disorders [Weischenfeldt et al., 2013, Collins et al., 2020, Liu et al., 2022].

Technological advances today allow the detection of SVs which have been impossible to study in such detail mere years ago. Inversions, a class of SVs that reverse the orientation of a genomic region, are particularly affected by this trend. Recent studies have been able to increase the number of known (human) inversions gradually [Chaisson et al., 2019, Giner-Delgado et al., 2019], and it has become clear that inversions affect evolutionary processes [Faria et al., 2019, Hsieh et al., 2021], gene regulation [Loveland et al., 2021, Giner-Delgado et al., 2019] and the formation of genetic diseases [Kozel et al., 2021, Koolen et al., 2016, Yuan et al., 2015] much more prominently than previously thought.

Despite these advances, our knowledge of certain classes of SVs is still incomplete – a notion especially true for inversions, which are among the hardest-to-detect classes of SV [Alkan et al., 2011]. For example, inversions have played a role in the human-primate evolution [Catacchio et al., 2018], but the extent and molecular consequences of this process remain unclear. Similarly, while examples of inversions found in human genomes are abundant [Giner-Delgado et al., 2019, Chaisson et al., 2019], detecting the full spectrum of human inversions has proven to be a significant challenge. Accordingly, the contribution of inversions to phenotypic variation in the human population remains to be determined. Lastly, the co-clustering of human inversions with disease-associated genomic regions

suggests a molecular relationship between the two. However, the molecular underpinning of this association remains vague and requires further clarification.

The work presented here encompasses several projects investigating different aspects of genomic inversions in a cross-species (humans vs. primates) and human-specific context. The remainder of this introductory chapter will first review the role of SVs in genomic variation and explain how SVs correlate with complex DNA regions. Subsequently, the key features known about the biology of inversions will be reviewed, highlighting various sub-classes and their associations with genome function. The next section will review current technologies for identifying SVs and inversions in particular. Eventually, an outlook will be given toward the context of the work presented in this thesis.

1.1 Structural Variation: how a forest was missed for the trees

The last decades have seen significant progress in unraveling the spectrum of variation in human genomes, facilitated by ever-growing pools of sequencing data provided by efforts like the 1000 genomes project [1000 Genomes Project Consortium et al., 2012] or UK Biobank [Szustakowski et al., 2021]. At the same time, comprehensive catalogs of variation in human genomes have quickly become an invaluable asset in advancing genetic and medical research [Kidd and Kidd, 2007]. In late 2022, the *dbSNP* database has counted approximately 324 Million single nucleotide polymorphisms (SNPs) in human genomes [Sherry et al., 2001], demonstrating the capacity to discover small-scale variations of one or few base pairs (bp) at high accuracy. Biased by this technological accessibility, such mutations have long been regarded as the primary source of genetic variation. Consequently, genome-wide associated studies (GWAS) concentrated on the effect of SNPs, while SVs received little attention at best [Krueger, 2012]. However, only 10% (3-4 million bp) of variable nucleotides between two typical human genomes originate from SNPs. In contrast, roughly ten times this number of nucleotides is affected by indels and genomic rearrangements, so-called SV [Trost et al., 2021]. SVs are typically defined as deletions, insertions, translocations, and inversions spanning at least 50 bp. Although SVs are frequently described as single events, combinations of SVs forming more complex events can also be observed [Carvalho and Lupski, 2016]. Furthermore, SVs are associated with gene function, regulation, and phenotypic outcomes far more frequently than SNPs [Sudmant et al., 2015, Ebert et al., 2021]. However, the number of SVs reported is rarely consistent across studies - at times differing by orders of magnitudes depending on technologies used for the survey - which highlights the challenge that structural variation detection still poses to screening technologies used to date [Ho et al., 2020].

From a naive perspective, it may seem surprising that single-base substitutions are more readily detectable than genomic rearrangements, which sometimes span millions of nucleotides. Indeed, early cytogenetic studies reported almost exclusively on megabase-sized chromosomal aberrations identified with microscopic or indirect inference-based methods [Kannan and Zilfalil, 2009, Sturtevant, 1917]. However, modern whole genome sequencing (WGS) is preceded by fragmentation of DNA molecules and thus produces fragments of 'mere' tens, hundreds, or thousands of base pairs in length. On the one hand, the information contained within an individual read (such as SNPs or indels) can typically be extracted directly with high confidence. On the other hand, inference of read-spanning events such as long SVs require more sophisticated analysis and are prone to

false calls [Mahmoud et al., 2019]. For this reason, modern geneticists find themselves in a paradoxical situation: The smallest genomic variants are routinely determined with basepair precision, while rearrangements involving entire chromosomal compartments can often remain entirely undetected. In other words, we are missing the forest of SVs for the trees of SNPs.

1.1.1 Rearrangements mediated and obscured by low-copy repeats

Identifying variation is more difficult in some genomic regions than in others. One reason for this is the presence of stretches of highly similar (>95%) duplicated sequences between a few kbp and several Mb in length, so-called segmental duplications (SDs) or low-copy repeats (LCRs). Several such duplications can also partially overlap, creating complex webs of interspersed segmental duplications. The repetitive nature of such regions makes read mapping extremely challenging and impedes confident SV calling [Mahmoud et al., 2019]. From a biological perspective, the expansion of SDs is especially prominent in the recent hominid evolution, as human SDs contain multiple gene families which have likely played essential roles in the evolution of the human brain [Cantsilieris et al., 2020]. Approximately 7.0% of the human genome corresponds to segmental duplications [Vollger et al., 2022], with the amount decreasing with phylogenetic distance to humans (chimp: ±0%; orangutan: -40%, macaque: -50% SD content compared to humans) [Marques-Bonet et al., 2009]. SD content varies widely in the animal kingdoms, and other orders show similar variances between related species, e.g., a recent study in butterflies and moths reports a species-dependent span of 1.2% and 15% SD content, most of which emerged late in speciation [Zhao et al., 2017]. Apart from their evolutionary role in duplicating and diversifying genomic content, SDs are roughly 10-fold enriched for standard copy number variation [Vollger et al., 2022]. Due to their long and highly similar nature, SDs also promote a class of structural variation via non-allelic homologous recombination (NAHR), a form of ectopic homologous recombination that results in a gain, loss, or inversion of genetic content. Many of the largest SVs are very hard to study since the same SD-rich sequence architecture that promotes their formation also negates accurate mapping of reads. Figuratively speaking, one may be tempted to conclude that scientists and cellular enzymes are equally overwhelmed by the repetitive nature of these sequences.

Section 1.3 will highlight how technological advances of the last 3–5 years have

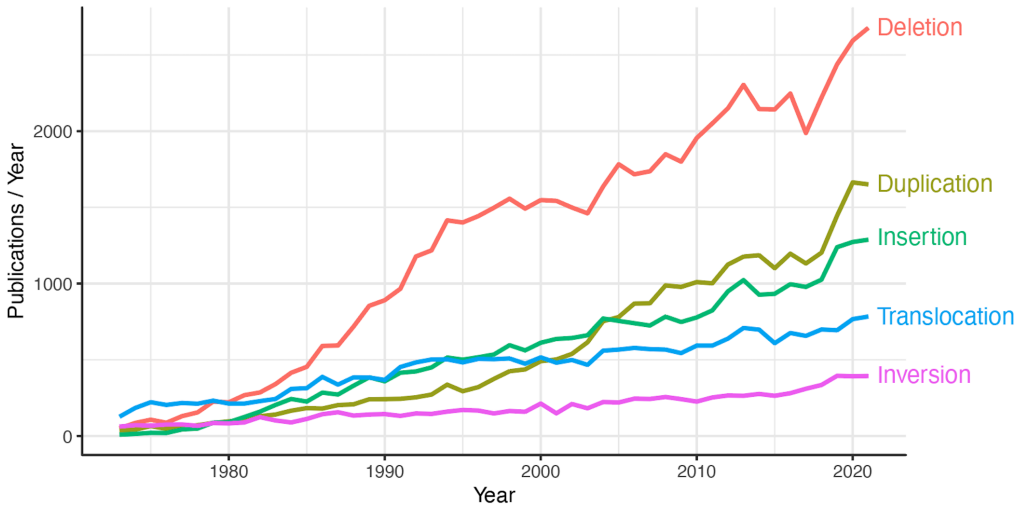

Figure 1.2: Number of peer-reviewed publications from the field of genetics mentioning different SVs in the title or abstract. Data were obtained through the dimensions.ai platform; publications were filtered by 'Fields of Research group: 3105 genetics' according to the ANZSRC 2020 scientific classification standard.

brought the study of SD-rich regions into a feasible range, and the full spectrum of SVs mediated by and between them has now become partially accessible. Due to the implied gain or loss of sequence, NAHR-mediated insertions and deletions can be confidently identified with classical WGS technologies. NAHR-mediated Inversions, however, have widely escaped previous genomic screenings (see Fig. 1.2), and it is for most recent technological advances that we now have the opportunity to explore this widely under-explored class of inversions with unprecedented resolution.

1.2 The biology of inversions

The first evidence of a chromosomal inversion was published in 1921 by Alfred Sturtevant, a pioneer in the field of genetic mapping [Sturtevant, 1921]. As he correctly identified then, inversions are central players in the evolution of species by their ability to suppress homologous recombination in heterozygous settings [Stevison et al., 2011]. Inversions remained a central topic to population geneticists for the next half a century until research foci eventually shifted more towards molecular genetics, partly driven by technological advances in this direction [Kirkpatrick, 2010]. Finally, 100 years after Sturtevant's discovery, modern genetics has fully recovered its interest and ability to survey large

genomic events. As a result, inversions are also getting back into the focus of geneticists [Kirkpatrick, 2010]. This section will explore the biological foundations of inversions, focusing on inversions in humans and occasional digressions to examples from other animals. Before the biological significance of this class of SV is described, however, the different sub-events summarized under the term 'inversion' will be clarified in the next section first.

1.2.1 Inversions: An umbrella term for a plethora of DNA rearrangement events

Generally, inversions appear whenever a genomic segment re-integrates itself in its original location in reverse orientation [Kirkpatrick, 2010]. However, a diverse group of chromosomal rearrangements is hiding behind this broad definition, and it is helpful to illuminate the different classes of inversions first.

Non homology mediated inversions

The simplest form of inversions occurs if a paired DNA double-strand break (DSB) isolates a DNA segment, and the segment is re-integrated in reverse orientation by DSB-repair via non-allelic homologous end joining (NHEJ) [Carvalho and Lupski, 2016]. Breakpoints of such inversions are most frequently simple, blunt ends or display microhomology (1-3 bp), although occasionally small deletions or insertions can be present, too. [Pannunzio et al., 2014]. The majority of NHEJ-mediated inversions are smaller than 20 kbp [Porubsky et al., 2022b]

An estimated 45% - 66% of non-homology-mediated inversions show evidence of additional insertions or deletions of >50 bp length [Giner-Delgado et al., 2019, Porubsky et al., 2022b]. Such secondary SVs are indicative of replication-based mechanisms (RBM) such as break-induced replication break-induced replication (BIR), microhomology-mediated break-induced replication (MMBIR), serial replication slippage (SRS) and fork stalling and template switching (FoSTeS), which are reviewed in more detail, e.g., in [Hastings et al., 2009]. These mutational events tend to create complex genomic rearrangements (CGRs) with more than two breakpoint junctions and, aside from inversions, can also include duplications, insertions, deletions, and translocations (Fig. 1.3) [Collins et al., 2017]. In other cases, inversions dominate CRGs, and the distinction between CGRs and simple inversions can sometimes become blurry. The length of the

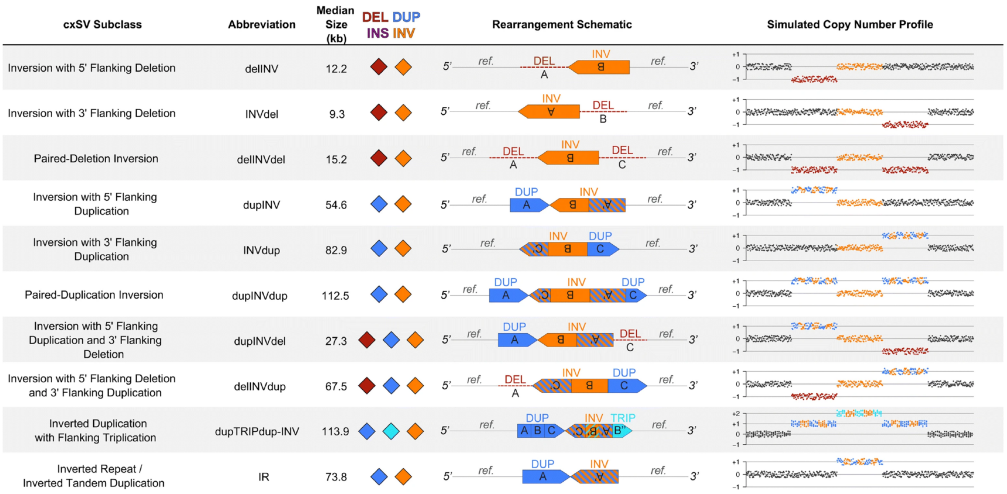

Figure 1.3: Selected inversion-containing complex genomic rearrangements found across the genomes of 689 individuals with developmental abnormalities. Likely formed primarily by replication based mechanisms, complex genomic rearrangements display a large variability of individual events in which inversions can be nested. Figure adapted from [Collins et al., 2017].

inverted fraction of CGRs is typically in a similar range as NHEJ-mediated inversions [Giner-Delgado et al., 2019, Porubsky et al., 2022b]. However, events >100 kbp have also been observed in patients with developmental abnormalities [Collins et al., 2017].

Homology mediated inversions

Pairs of highly identical segmental duplications mediate the second class of inversions through ectopic crossing over between homologous sequences via non-allelic homologous recombination (NAHR). Such events can appear during mitosis and meiosis, although only the latter can get fixated in the germline and thus contributes to genomic variation. For topological reasons, NAHR along inversely oriented repeats promotes inversions of the intermediate sequence, while the same process along directly oriented repeats creates deletions or duplications (see Fig. 1.4).

NAHR-mediated inversions are challenging to detect and have thus yet to be studied extensively. However, the opposite is true for NAHR-mediated copy number variants (CNVs), whose association with human disease is long-standing and well-documented [McKusick, 1970, Stankiewicz and Lupski, 2002]. Thus, despite lacking significant numbers of reported inversions, some of their specific characteristics can be extrapolated from our knowledge of the NAHR mechanism. A classic assumption is that SDs of at least ten kbp length and 97% sequence

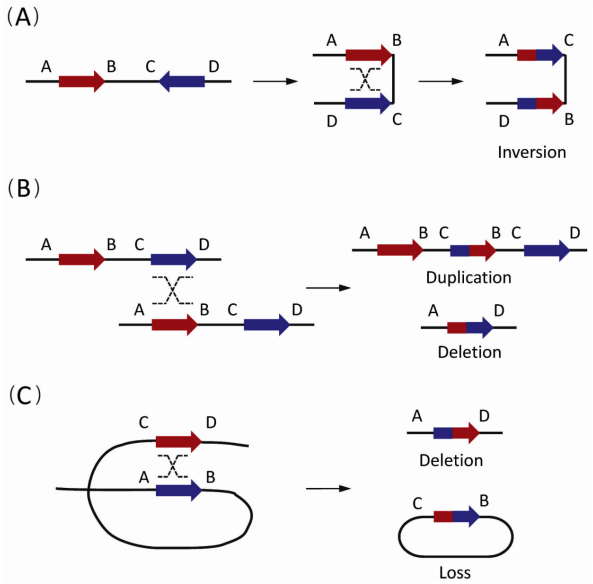

Figure 1.4: Scenarios of non-allelic homologous recombination (NAHR) leading to various SV outcomes. A Crossing over between inversely oriented SDs (blue/red arrows) leads to inversion of the intermediated segment. **B** Inter-chromosomal NAHR between directly oriented repeats lead to sequence transfer from one homologous partner to the other, creating a duplication/deletion pair. **C** Extrusion of a sequence through intrachromsomal NAHR between directly oriented repeats leading to a deletion. Figure adapted from [Chen et al., 2014].

identity can act as substrates for NAHR, especially in the presence of sequence motifs for the *PRDM9* gene, which is involved in the determination of recombination hotspots [Stankiewicz and Lupski, 2002, Paigen and Petkov, 2018]. However, crossovers have also been observed for much smaller segments (34 – 114 bp), or sequences with as low as 94% sequence similarity [Steinmann et al., 2007, Lam and Jeffreys, 2006]. The size of NAHR-events does typically not exceed 5 Mbp [Stankiewicz and Lupski, 2002], and the frequency of NAHR correlates positively with SD length, SD similarity, and the presence of *PRDM9* motifs, but negatively with the distance between repeats [Liu et al., 2011]. Consequently, the positions and frequencies of NAHR-mediated inversions are pre-determined by existing SD pairs, analysis of which suggests that about 12% of the human genome may potentially be susceptible to inversions [Zhang et al., 2010]. Indeed, while the number of NAHR-mediated inversions has been unclear until recently, several reports suggest that the vast majority of human inversions longer than ten kbp are mediated by this mechanism [Kidd and Kidd, 2007, Giner-Delgado et al., 2019].

The presence of SDs at the flanks of NAHR-mediated inversions paves the way for recurrent mutations, a phenomenon that has been termed “inversion

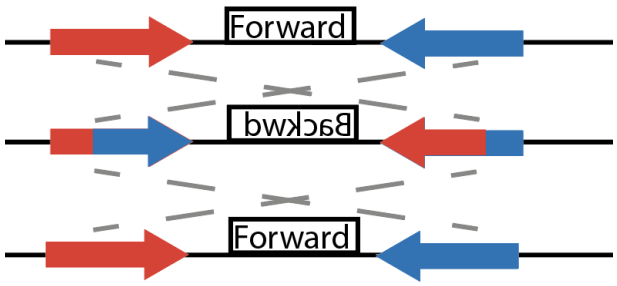

Figure 1.5: Schematic view of inversion recurrence between a pair of inversely oriented segmental duplications. Balanced NAHR-mediated inversions are mediated by long inversely oriented repeats (red/blue arrows). The homologous nature of these repeats remains intact even after an inversion event, allowing for repeated inversion in a process called *Inversion toggling* or *recurrence*.

toggling" [Zody et al., 2008] (Fig. 1.5). Several studies have concluded that disease-associated genomic regions changed their orientation multiple times between a direct and inverted state during primate evolution. This includes, for example, chromosomal regions 15q25, 15q13.3, 16p12.2, 16p11.2, 8p23.1, Xq22, 17q21.31, as well as at the hemophilia A locus (Xq28) [Lozier et al., 2002, Zody et al., 2008, Antonacci et al., 2014, Catacchio et al., 2018, Maggiolini et al., 2020b, Maggiolini et al., 2020a, Porubsky et al., 2020, Puig et al., 2020]. In human cohorts, targeted assays applied to a limited number of loci have revealed shared SNPs between directly oriented and inverted haplotypes, hinting at ongoing inversion toggling in humans [Zody et al., 2008, Aguado et al., 2014, Puig et al., 2020]. In the most comprehensive study by 2019, Giner-Delgado and colleagues reported signs of inversion recurrence in 20 out of 45 inversions analyzed [Giner-Delgado et al., 2019], suggesting that toggling may be a common feature among inversions. However, just like the full spectrum of inversion loci has only been examined comprehensively in the study presented in chapter 4, the extent of inversion toggling within human genomes could not be estimated until recently and will be discussed further in Chapters 2 and 4. NAHR-mediated inversions constitute the focus of the work presented in chapter 4, where we will revisit the characteristics discussed here.

1.2.2 Three associations between inversions and genome function

Most inversions are characterized by a conserved pool of overall sequence content. This notion, at first glance, argues against a positive fitness effect and thus against the spreading of inversions through populations. Moreover, centromere-spanning inversions are present between human and chimpanzee genomes [Feuk et al., 2005] despite selective disadvantages: Such inversions are at risk of producing unviable gametes during crossover and should thus have negative fitness effects when they first emerge. Similarly, balanced inversions can also cause or predispose to disease in humans. This notion raises the question of how inverted segments can carry (dis)beneficial properties that their uninverted counterparts lack. At least three distinct mechanisms exist through which inversions can shape the functional genomic landscape (Figure 1.6), which will be highlighted in this section.

First, inversions can disrupt genes directly, create fusion genes or, more indirectly, affect gene expression by exchanging regulatory neighborhoods, bringing sequences into proximity that would otherwise be distant, and vice versa [Puig et al., 2015, Giner-Delgado et al., 2019, Lupiáñez et al., 2015]. Such gene misregulation carries the potential of deleterious but occasionally also adaptive effects. For example, two inversions are known to disrupt the *FVIII* gene, causing almost 50% of severe cases of *hemophilia A* [Park et al., 2014, Lakich et al., 1993]. Other examples include the inversion of a single exon in the *RHOH* gene [Giner-Delgado et al., 2019] and various inversion-mediated fusion transcripts such as *ZNF257* [Puig et al., 2015], the *IFITM2/IFITM3* gene pair [Giner-Delgado et al., 2019] or the *CTRB1/CTRB2* genes. Notably, fusion transcripts of the latter are known as a risk factor for chronic pancreatitis [Rosendahl et al., 2018]. On the level of gene regulation, individual loci are susceptible to disruption of topological domains and enhancer hijacking mediated by inversions and other SVs. This encompasses inversion-mediated misexpression of *Pitx1* in forelimbs leading to partial arm-to-leg transformation [Kragesteen et al., 2018], and SVs altering the TAD-spanning *WNT6/IHH/EPHA4/PAX3* locus [Lupiáñez et al., 2015].

Second, several studies have indicated a close link between inversions, microdeletions, and microduplications. One effect contributing to this association is the phenomenon of inversion-associated CNVs. As described earlier, roughly half of non-homology-mediated inversions are associated with insertions or deletions up to 200 kbp in length. In this way, inversions can be associated with functional effects, even though they may not always confer them directly [Giner-Delgado et al., 2019]. However, even some balanced inversions can be associated with modified microdeletion and -duplication formation rates. Such CNVs often span >1 Mb in size and are frequently causative of disease [Koolen et al., 2016, Osborne

et al., 2001]. Statistical relationships to inversions have been documented on a wide range of disease-associated CNVs, e.g., at the 3q29, 8p23, 15q13.3, 15q24, 17q12 and 17q21.31 loci [Antonacci et al., 2009a, Mostovoy et al., 2021, Zody et al., 2008]. Anecdotal evidence suggests that both the CNVs and inversions in these loci may be predominantly driven by NAHR occurring within the complex webs of SDs in certain regions. Well-documented cases include Williams-Beuren Syndrome [Kozel et al., 2021], Koolen de Vries syndrome [Koolen et al., 2016, Itsara et al., 2012] and the *NPHP1*-containing locus on 2q13 [Yuan et al., 2015]). These examples demonstrate that inversions can reorganize the local landscape of segmental duplications, facilitating (or protecting against) the formation of subsequent SD-driven copy-number variants through NAHR [Carvalho and Lupski, 2016, Hsieh et al., 2021]. However, highly complex SD structures in these loci have so far impeded accurate inversion genotyping across large cohorts, and the detailed mechanistic underpinnings of inversion-CNV relationships are still missing in most cases.

Lastly, it has long been known that balanced inversions can suppress recombination as heterozygotes [Sturtevant, 1921], thus promoting genetic isolation between individuals, which can be a path to speciation [Kirkpatrick and Barton, 2006]. For example, in *drosophila*, the recombination rate between heterozygous inversions falls magnitudes below that of non-rearranged regions [Andolfatto et al., 2001]. In his review article, M. Kirkpatrick suggests an analogy between the populations of (a) inverted and uninverted chromosomes and (b) a pair of coexisting biological species: While each species individually evolves under Mendelian rules, there is little to no genetic exchange between the species (or chromosomes). As a result of ecological competition, this will lead either to the coexistence of two species (i.e., stable polymorphism) or replacement of one by the other (i.e., fixation of one haplotype) [Kirkpatrick, 2010].

The genetic separation of inverted and uninverted haplotypes allows for two or more alleles to co-segregate and emerge as alternative versions of genetic content specialized in different environments. This concept is the basis for inversion-based local adaptation, most famously observed in the inversion 3RP in *Drosophila meanogaster*. In this example, two inversion haplotypes provide fitness benefits in different climatic environments. Consequently, the geographical cline between individuals carrying one haplotype over the other has shifted away from the equator over the last 100 years, reflecting the continued increase in global temperatures [Anderson et al., 2005]. A well-known example in humans is the ca 900 kbp inversion region on 17q21.31, whose direct and inverted haplotypes harbor two distinct lineages. The lineages show little evidence of recombination

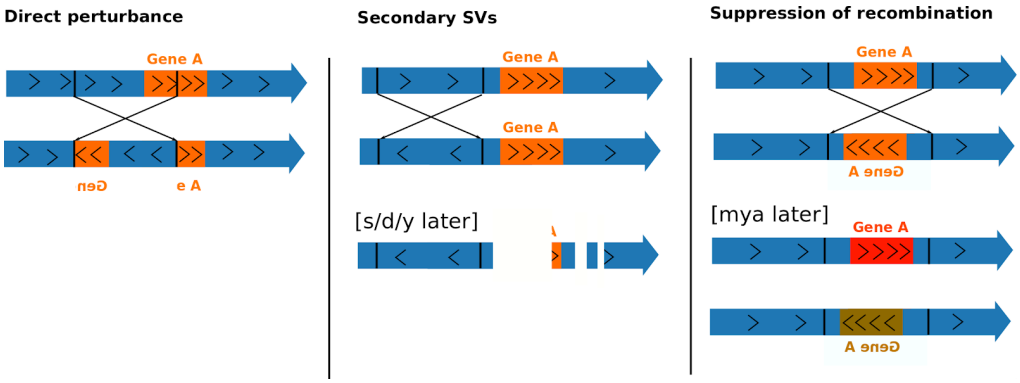

Figure 1.6: Schematic illustration of the three classes of interactions between inversions and genome function. Left: Inversions can directly perturb gene bodies or, more generally, disrupt regulatory neighborhoods, including topologically associating domains (*TADs*) or gene-enhancer pairs. **Middle:** Inversions can also be part of copy number variations, which can either occur as part of the initial inversion event or be facilitated by an inversion-mediated rearrangement of segmental duplications. **Right:** Large inversions additionally have the potential to suppress recombination, facilitating the emergence of independent haplotypes.

over the last 3 million years, and the inverted haplotype has been associated with increased fertility in the Icelandic population [Stefansson et al., 2005]. A recent study similarly reported 21 large (>1.5 Mbp) inversion polymorphisms in deer mice across the north American continent which cause near-complete suppression of recombination and likely shape local adaptation [Harringmeyer and Hoekstra, 2022].

Another prominent example of fitness benefits provided by blocking recombination through inversion is the Y chromosome which determines sex in mammals and other groups like flies [Landeen and Presgraves, 2013]. The mammalian Y chromosome is derived from the X chromosome – originally a regular autosome – and the two chromosomes still recombine in selected regions, so-called pseudoautosomal regions (PARs) [Landeen and Presgraves, 2013]. In humans, a series of inversions on chrY have established suppression of recombination with the X chromosome [Lahn and Page, 1999]. Such sex-specific separation of alleles is beneficial in the light of sex-antagonistic selection, i.e., alleles that provide fitness benefits to males or females.

Generally, the evolutionary significance of inversions has gained interest in recent years. While not immediately relevant for the further flow of the work presented here, we refer to [Wellenreuther and Bernatchez, 2018] for more examples of inversions in evolution and ecology from all branches of life.

1.3 Traditional and emerging approaches for detecting inversions

As discussed in the previous sections, the identification of inversions is technologically challenging, and no routine method has been established yet. Inversions and other SVs share most of their technological options for detecting, and thus inversions suffer from similar difficulties as other SVs, though to a larger extent given some of its unique features:

1. Balanced inversions exhibit **copy-number neutrality**. Sequencing depth-based inference, as performed, e.g., routinely on large deletions/duplications via short reads, is therefore not viable.

2. **Large segmental duplications** at the flanks of NAHR-mediated inversions pose a challenge to direct breakpoint mapping.

3. Inversions are often **recurrent**, resulting in low linkage disequilibrium (LD) with other variants. Inference-based methods like PanGenie [Ebler et al., 2022] rely on such metrics and are thus unsuited for detecting recurrent events.

Identifying mutations, including SVs, generally follows one of two objectives: de-novo *calling* describes identifying mutations without prior knowledge about their location. In contrast, *genotyping* refers to the determination of genotypes of pre-defined candidate loci. Out of these two related problems, genotyping can be considered the conceptually 'easier' one due to a reduced risk of false positive calls, which is gained by sacrificing the ability to detect previously unseen variation [Mahmoud et al., 2019].

According to a classification suggested in a review by M. Mahmoud and colleagues, technologies for calling or genotyping SVs fall into different categories, which include short-read mapping, long-read mapping, de novo assembly, Strand-Seq technology, and other methods (Hi-C, Optical mapping, 10x Genomics, Multimethods SV caller) [Mahmoud et al., 2019]. With a general focus on inversions, this section will review the most critical technologies used in SV calling and genotyping and elaborate on their strengths and weaknesses.

Short-read mapping

Short paired-end sequencing is still the most commonly applied sequencing technique due to its cost efficiency and matured protocols. SV calling from short reads is still the standard approach and has been applied to large cohorts such as the 1000 genomes project [Sudmant et al., 2015, Mahmoud et al., 2019]. Three

different kinds of SV-relevant information can be extracted from short read pairs aligned to a reference: First, the depth of read coverage can be informative of copy-number variations in a genomic location. Second, SVs can lead to *paired-end read discordance*, a phenomenon in which the relative orientation or distance between two reads of a pair is altered, indicative of inversion, loss, or gain of sequence in between. Lastly, reads overlapping SV breakpoints show distinct mapping patterns, such as mapping to two different genomic locations (presenting so-called *split-reads*). While more than 100 short-read-based SV calling algorithms have been released to date, the limitations inherent to the technology prohibit reliable identification of all classes of SVs across all size ranges [Mahmoud et al., 2019]. This notion is especially true for multi-kbp, copy-number-neutral inversions, which are essentially invisible to short-read-based SV calling (see also section 4.1, where existing inversion callsets are highlighted).

It was recently demonstrated that SV genotyping (but not SV calling) from short reads can be drastically improved by comparing reads to a pan-genome graph in a process called *genomic inference* [Liao et al., 2022]. Imagine, as an analogy, being asked to reconstruct a city's metro system based on a few blurry holiday photos – impossible without further information. However, if one is given access to subway maps of every city on earth, and one of the photos depicts a piece of the Eiffel tower, reconstruction of the subway system will be straightforward[1]. The same principle is used in genomic inference: high-quality pan-genomes provide statistical links between hard-to-call SVs (i.e. the underground) and easy-to-call SNPs (i.e. the skyline), and the correlation between the two can be used to infer one from the other [Ebler et al., 2022]. While the adoption of this technique can double to amount of SVs genotyped from short read data sets [Liao et al., 2022], recurrent events are exempt from this small revolution due to a lack of linkage equilibrium between SNPs and recurrent SVs.

(Ultra-) Long-read mapping

Sequencing methods by Pacific Biosciences (PacBio) and Oxford Nanopore technologies (ONT) can produce reads spanning dozens, sometimes hundreds, or even thousands of kilobases [Loose et al., 2018]. Longer reads can more easily be anchored in complex genomic regions than shorter ones. Consequently, long reads display improved performance in calling SVs in repetitive or SD-rich regions [Ebert et al., 2021]. Additionally, long reads grant the ability to resolve long, complex SVs which contain a combination of simple events. However, long reads often display

[1]This analogy has been conceived after a weekend trip to Paris.

a higher sequencing error rate than short reads. Depending on the system used, long reads additionally exhibit specific biases, such as a tendency of ONT-reads to collapse homopolymers [Delahaye and Nicolas, 2021]. As a resposne, specialized mapping pipelines such as minimap2 have been tailored to take such biases into account [Li, 2018]. Again, multiple SV calling algorithms have been developed over the years, with *PBSV* (https://github.com/PacificBiosciences/pbsv, unpublished) and *Sniffles* [Sedlazeck et al., 2018] serving as prominent examples. Long-read SV calling outperforms short-read-based methods, calling between 2 – 4 fold more SVs per sample [Mahmoud et al., 2019].

De novo Genome assembly

Traditionally, de novo genome assembly has been regarded as a resource-intense process and was thus reserved for generating reference genomes [Nagarajan and Pop, 2013]. However, the process of assembling individual genomes has become more feasible over the last years, and > 100 human genomes have been assembled to date, with contiguity comparable to – or exceeding – that of the hg38 reference (as measured by the *N50* metric) [Ebert et al., 2021, Liao et al., 2022]. Assemblies are typically generated using (ultra) long reads, sometimes with additional phasing information provided by Strand-Seq, Hi-C, or Mother-Father-Child trios [Ebert et al., 2021]. SV detection can be performed by comparing assembled genomes to a reference and subsequently identifying discontinuities. While this method can be compelling for identifying all kinds of SVs, including inversions, the most significant benefit of using de novo genome assemblies lies in detecting multi-kbp long insertions [Tian et al., 2018]. However, two bottlenecks arise: First, genome assemblies have a tendency to be disrupted or collapsed in SD-rich regions, resulting in many inversion-associated regions which are not fully resolved ([Porubsky et al., 2022a] and Fig. 1.7). Furthermore, even in the case of a high-quality contig spanning an inversion, local alignment to a reference poses a challenge. Assembly-based SV calling algorithms are still sparse, with prominent examples being *paftools.js* [Li, 2018], *SyRI* [Goel et al., 2019] and *PAV* [Ebert et al., 2021].

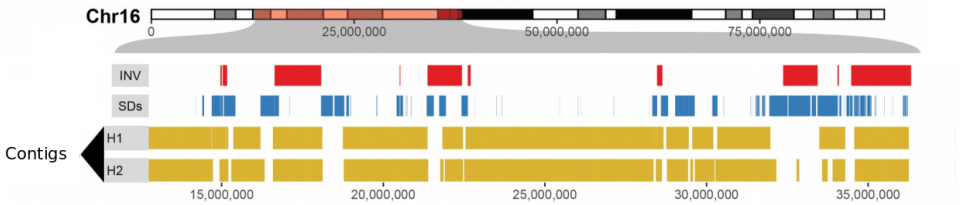

Figure 1.7: View of de-novo assembled contigs, Inversions, and SDs on chr16 in one sample presented in [Ebert et al., 2021]. Characteristically, long inversions display SDs at their breakpoints, prohibiting uninterrupted sequencing of their flanks. It is typical to observe contig breaks in these locations, a fact that currently limits the power of de-novo sequenced genomes to represent inversions accurately.

Strand-Sequencing

Single-cell DNA template strand sequencing (Strand-Seq) is a sequencing technique which can selectively sequence individual *strands* of DNA, preserving their directionality [Falconer et al., 2012, Sanders et al., 2017].

To achieve this, Strand-Seq follows a specialized protocol for preparing and selectively digesting strands of DNA (Fig. 1.8A). The DNA double helix consists of two oppositely directed strands, the plus ('Watson') and minus ('Crick') strands. During replication, the helix is unwinded, and each plus and minus strand serves as a template to which a nascent strand gets added. In the Strand-Seq protocol, cells undergo one round of cell division in the presence of 5-Bromo-2'-deoxyuridine (BrdU), a thymidine analog that gets integrated into the nascent DNA strands. Following cell division, several daughter cells (typically n=96) from a cell pool are selected via Fluorescence-activated cell sorting (FACS) and processed further in ixndividual wells. The DNA of these cells gets digested using a micrococcal nuclease enzyme (MNase), ligated to sequencing adapters, and fragments containing BrdU are degraded by photolytic cleavage, leaving only fragments from the template strands. After adding cell-specific barcodes, the fragments are processed by short-read whole genome sequencing [Sanders et al., 2017].

The strand from which a read originates (plus/W or minus/C) eventually determines the orientation in which the read maps to a reference genome. All reads from the same DNA molecule are expected to map in the same direction. However, given the diploid nature of human cells, each cell contains two template strands per chromosome –one per homolog–, and the signals of the two are overlaid. This process results in a random fraction of libraries that display reads mapping only in the 'W' direction (both template strands were plus strands), only in 'C' or reads which are split between both directions ('WC'/'CW' libraries)

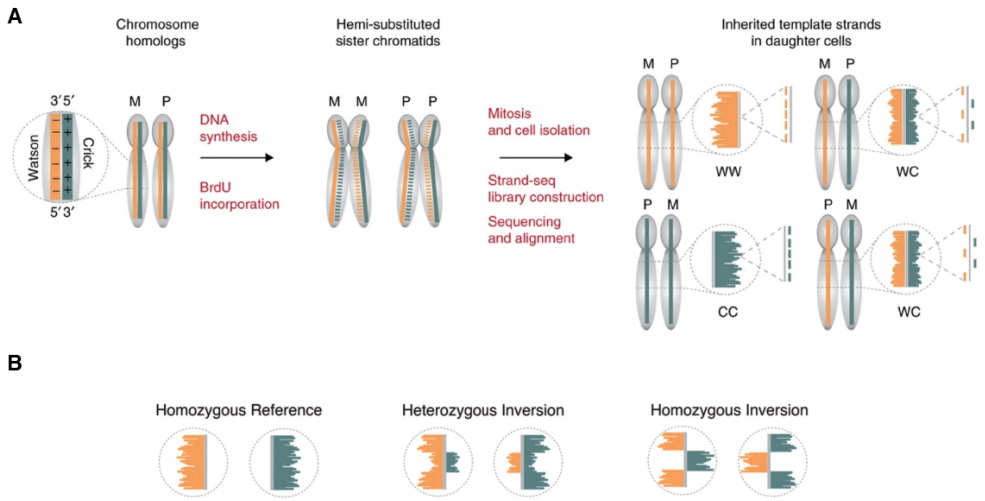

Figure 1.8: Basic principles and workflow of the Strand-Seq technology. A Cells initially undergo one round of replication in the presence of BrdU, which acts as a label for the nascent strands of replicated chromosomes. Per chromosome, each daughter cell inherits two Watson ('WW'), two Crick ('CC') or one Watson and one Crick ('WC' / 'CW') template strands. As part of library preparation, the BrdU-labeled nascent strands are removed, and only template strands are passed on for sequencing. **B** The Strand-Seq specific information is contained in the directionality of mapping reads. Heterozygous and homozygous inversions can be identified from different characteristic patterns of strand switching.

(Fig. 1.8A). In the latter case, all reads in the same direction belong to the same haplotype, a property that can be exploited for long read phasing [Ghareghani et al., 2018].

Certain SVs, such as inversions and sister chromatid exchanges, break the unidirectionality of reads from one haplotype. Such breaks manifest themselves in a switch of read orientation in the affected loci. For example, a library that has inherited 'WW' reads on one chromosome is expected to switch to 'WC' in a region containing a heterozygous inversion and to 'CC' in the case of a homozygous inversion. Likewise, a 'WC' cell will switch to 'WW' or 'CC' for a heterozygous inversion but to 'CW' for a homozygous inversion – which is indistinguishable from 'WC.' By identifying analogous patterns, also other SVs can be detected, such as deletions, duplications, and inverted duplications [Sanders et al., 2017]. Strand-Seq can thus identify SVs based on read directionality without the need to sequence breakpoints – ideal for detecting long inversions with impenetrable flanking regions. However, as a single-cell sequencing technique with extensive pre-processing, the read coverage is exceptionally shallow (typically around 0.03× coverage), which means that only larger SVs (> 1 - 10 kbp) can be detected. Additionally, due to the low sequencing resolution, complex or nested

events can typically not be resolved by Strand-Seq alone. Existing computational frameworks for processing Strand-Seq data, as well as the development of a novel Strand-Seq-based genotyping algorithm, will be discussed in chapter 3.

Other Methods

While not central to the work presented in subsequent chapters, other techniques have been used in recent works to identify inversions and other SVs:

1. **Optical mapping** (e.g., produced by BioNano Genomics) uses a fluorescent dye to label specific nucleotide sequences, creating barcode-like fluorescence patterns which can be aligned to one another to identify SVs. This technique is highly cost-efficient, while breakpoint accuracy is limited [Lam et al., 2012].

2. **Hi-C** is used to identify DNA regions that fall in close proximity in 3D space. Alterations of 3D interaction patterns can be informative of long-range SVs (e.g. demonstrated recently by Hi-C based exploration of complex inter- and intrachromosomal rearrangements (so-called 'chromoplexy' and 'chromothripsis') in the germline of 11 patients [Schöpflin et al., 2022]).

3. **Linked reads** provided by 10x genomics resemble classical Illumina-based paired-end short reads but increase the insert size up to 150 kbp, enhancing the ability to detect large variants [Marks et al., 2019].

4. While lacking de-novo calling power, **PCR**-based techniques provide a simple, cost-efficient way to genotype known simple inversions in individuals and large populations [Giner-Delgado et al., 2019].

1.4 Motivation and thesis overview

The previous sections have highlighted the limitations that have contributed to placing inversions among the least well-studied classes of structural variation. Consequently, the field is likely to underestimate the clinical and ecological relevance of this class of SVs, as their role in genome evolution, adaptation, and variation still needs to be thoroughly studied.

More than 100 years after their initial discovery, the identification of inversions of all size ranges has finally approached the edge of technological feasibility. These advances present an opportunity to investigate inversions comprehensively and

in unprecedented detail. The studies presented in this thesis aim to identify and analyze near-complete spectra of inversion polymorphisms in humans and great apes using extensive data sets obtained with recent technologies, including Strand-Sequencing, long-read sequencing, and de-novo genome assembly. Given the limited knowledge about inversions, even fundamental questions about their abundance and length distribution are of great interest. Consequently, the construction of workflows for creating accurate call sets is a strong focus of these studies. Additionally, the projects examine additional aspects of inversions, such as their relationship with gene-regulatory neighborhoods in great apes and their recurrent nature and association with disease-causing copy-number variations in human populations. Finally, motivated by an unexpected level of structural variety associated with inversions, the last part of the thesis is dedicated to the highly diverse genomic variation in the most repeat-rich regions of the genome, where inversions act merely as one of several contributing factors. The specific contents of each chapter are briefly summarized below.

Chapter 2 will initially focus on my contribution to a study of inversion polymorphisms across four great ape species led by A. Sanders and D. Porubsky. This project describes an inversion callset derived from Strand-Seq, which serves as a basis to explore the relationship between inversions and the evolutionary conservation of gene regulatory environments. Throughout this project, inversions are viewed in correlation with topologically associating domains (TADs), revealing that the breakpoints of long inversions preferentially locate near TAD boundaries. I furthermore examine the level to which long inversions contribute to evolutionary conserved gene expression changes and estimate the role that TADs likely play in this process.

Next, **Chapter 3** describes the development of a new Strand-Seq-based inversion genotyper which I co-developed with my colleague H. Ashraf. This novel tool, ArbiGent, is described and tested for performance before being utilized as an essential building block for genotyping human inversions in the following chapter.

The subsequent **Chapter 4** describes a collaborative study co-led by me and my colleagues D. Porubsky and H. Ashraf. This study explores the full spectrum of inversions in the human genome and constitutes the most comprehensive study of human inversions to date. In this process, I attempted to classify inversions in terms of their sequence properties, and my colleagues and I estimated the overall abundance of inversions in human genomes. Specific focuses of this study include (1) the analysis of inversion loci that have switched their orientation multiple

times during evolution, so-called recurrent inversions, and (2) novel insights into copy-number variations for which statistical relationships with inversions have been described previously.

Finally, **Chapter 5** sheds light on inversions in context with other SVs, especially long duplications and deletions. These SVs can all be mediated through similar mechanisms and are sometimes found nested with each other, creating highly dynamic hotspots of genomic diversity. To facilitate their study, I describe a new computational tool, NAHRWhals, designed to identify and untangle such complex events. In this study, chromosome Y is highlighted in more detail, as several complex inversions can be found on this chromosome. Eventually, this chapter investigates more cases of copy-number variations, which are often inherently connected to NAHR-based diversity hotspots.

2

Inversions in humans and great apes disrupt TADs and promote gene dysregulation.

This chapter describes a contribution to a project on inversion recurrence in great ape genomes published in Nature Genetics in 2020 [Porubsky et al., 2020]. At the time when I joined the laboratory of J. Korbel, the project had already been initiated as a collaborative effort with the laboratory of E. Eichler [University of Washington, US], and was led by A. Sanders [EMBL Heidelberg, Germany] and D. Porubsky [University of Washington, US]. This chapter focuses on the parts of the manuscript that I contributed, with work from collaborators clearly marked as such in the text. Specifically, the data presented in section 2.2 formed the basis of the project and were created by A. Sanders and D. Porubsky. I thank A. Sanders and J. Korbel for patiently supervising my work and providing ample feedback and discussion. I further thank all co- and senior authors for helping me to contribute to this project, especially D. Porubsky, P. Hsieh, A. Sulovari [University of Washington, US] T. Marschall [Heinrich-Heine University Düsseldorf, Germany], A. Sanders, J. Korbel and E. Eichler.

Contents

2.1 Introduction: Inversions may disrupt evolutionarily conserved TADs

Section 1.2 has highlighted molecular mechanisms by which inversions likely contribute to evolutionary processes. This chapter describes a project to assess the effect of non-human primate inversions on the expression of adjacent genes. Specifically, it is examined if a gene regulatory effect might be transferred through the disruption of gene regulatory neighborhoods, so-called topologically associating domains (TADs). The introductory section will first review previous attempts of discovering inversions in non-human primates (NHPs), before introducing TADs, which likely play an essential role in gene regulation and are potentially interlinked with inversions.

2.1.1 Known inversions in nonhuman primate genomes

An early mention of inversions found between the chromosomes of humans and non-human primes was made in 1982, where seven large inversion polymorphisms were observed via karyotyping [Yunis and Prakash, 1982]. Over time this finding was extended, and it became clear, that human and chimpanzee chromosomes display a consistent set of at least nine large, pericentric (centromere-spanning) inversions. Later studies have described a few dozen human regions inverted in other ape species, although detection power was still limited to large events typically exceeding 100 kbp [Ventura et al., 2001, Carbone et al., 2002, Kehrer-Sawatzki et al., 2005, Capozzi et al., 2012]. More recently, steady improvements in the quality of great ape genomes have enabled direct genomic sequence comparisons between species. An early study reported 1,526 inversions between the human and first chimpanzee assemblies (Fig. 2.1) [Feuk et al., 2005]. However, only 1.7% of these inversions were experimentally validated, and a significant portion of these calls likely represents false positives, reflecting the comparatively low quality of the initial genome assemblies. Drawing from more recent genome assemblies (released between 2011 and 2017), Cataccio and colleagues, too, compared the genomes of human, chimpanzee, gorilla and orangutan to map inversion polymorphisms [Catacchio et al., 2018]. Out of 156 inversions initially described, 120 were subjected to validation through fluorescence in situ hybridization (FISH) and sequencing-based methods, revealing that 37 (31%) corresponded to false calls tracing back to misoriented regions in one of the assemblies. The remaining 83 inversions mapped to 67 human loci between 103 kbp and

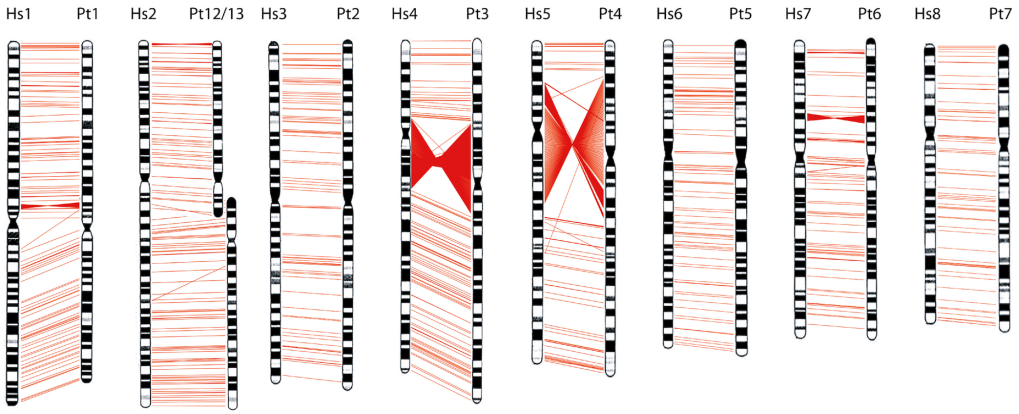

Figure 2.1: Visualization of inversions found between the human and chimpanzee genome assemblies in [Feuk et al., 2005]. Only eight of the 23 human chromosomes are displayed here. Two large pericentric inversions on human chromosomes 4 and 5 are clearly visible. A fraction of the remaining inversion calls may be artifacts representing assembly errors. Figure adapted from [Feuk et al., 2005].

5 Mbp. The validated inversions in this dataset overlapped with 81 human genes, mapped to ten fragile human sites frequently associated with genomic disease, and included 39 large known inversions described previously [Feuk et al., 2005]. In the same year, analysis of de-novo long-read based assemblies of chimpanzee and orangutan genomes exhibited 29 inversions between the three species between 100 kbp and 5 Mbp, roughly half of which had not been described previously [Kronenberg et al., 2018]. Furthermore, 93% of those inversions displayed SDs at their breakpoints, and 28% co-occurred in hotspots of human genomic disease. Moreover, expression data from brain organoid models revealed 18 differentially expressed genes associated with human-chimpanzee inversions (three of which were upregulated in human cells).

Reports state unanimously that the vast majority of human – ape inversions are nested in SDs [Feuk et al., 2005, Catacchio et al., 2018, Kronenberg et al., 2018], similar to long inversions in the human population [Sanders et al., 2016]. This notion explains the widely varying numbers of reported inversions in great ape species given that such inversions are very hard to detect for most sequencing methods (see section 1.3).

On the functional perspective, several reports have underlined the pronounced effect of reduced recombination in inverted regions. For example, protein divergence in human vs chimpanzee was reported to occur 2.2 times faster in rearranged compared to collinear chromsomes [Navarro and Barton, 2003]. Accordingly, several studies have noted the implications of this phenomenon for chromosomal speciation (see, e.g., [Farré et al., 2013] and section 1.2.2). Beyond suppression

of recombination, functional effects of such inversions are less clear. The human evolution was accompanied by a rapid expansion of segmental duplications often containing gene families associated to brain development [Cantsilieris et al., 2020]. Work from my colleagues has shown that these same SD-rich regions are frequently found inverted between humans and apes, suggesting an interconnection between SD expansion and inversions, although the details of this association are not yet fully understood [Porubsky et al., 2020]. Lastly, relatively little is known about another mode of functional effects of inversions in ape evolution: the change of gene regulation by disruption of gene regulatory environments.

2.1.2 TADs (co-)shape the 3D organization of genomes

The 3D organization of chromatin has been a subject of discussion in the field of genomics for decades [Rowley and Corces, 2018]. The genome is hierarchically folded inside the nucleus: from DNA winding to nucleosome clusters, chromatin loops, topologically associating domains (TADs), and eventually, chromosomal compartments spanning many Megabases in size [Dekker and Heard, 2015]. TADs have been the latest of these organizational units to be recognized widely [Dixon et al., 2012], and no full consensus has been reached in explaining their function, conservation, and importance. TADs are genomic regions which form interactions preferentially within themselves [Pombo and Dillon, 2015]. Such regions typically span length-scales of around 0.5 to 2 Mb in humans, and can be visualized e.g., by chromosome conformation capture techniques such as Hi-C [Belton et al., 2012] (Fig. 2.2). On a molecular level, binding motifs for the insulator protein *CTCF* are typically found at the boundaries of TADs, and the loop extrusion protein *cohesin* likely plays a crucial role in their formation and maintenance [Wutz et al., 2017]. Functionally, TADs have been suspected of playing a critical role in gene regulation, given that most enhancer-promoter interactions occur within them [Jost et al., 2017].

Indeed, naturally occurring and artificially induced alterations of TAD boundaries have been associated with changes in gene expression in several cases and across multiple species [Valton and Dekker, 2016, Krefting et al., 2018, Despang et al., 2019, Lupiáñez et al., 2016] and were shown to act as drivers in cancer and other genetic diseases [Lupiáñez et al., 2015, Hnisz et al., 2016]. However, in contrast to these results, other studies have noted a somewhat limited significance of TADs for gene expression. For example, disruption of TADs in artificially

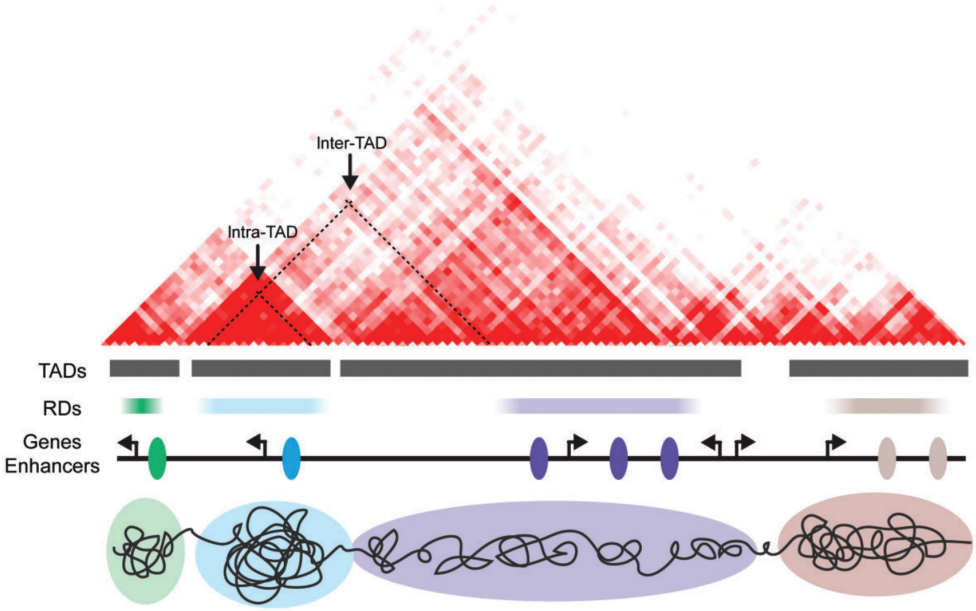

Figure 2.2: Visualization of example HI-C data to illustrate topologically associating domains (TADs). DNA interactions within TADs ('Intra-TAD') are more common than those across ('Inter-TAD'). Genes and enhancers encapsulated in the same TAD can form regulatory domains (RDs). Figure adapted from [Remeseiro et al., 2016].

shattered 'balancer' chromosomes in drosophila resulted in only slight changes to gene expression [Ghavi-Helm et al., 2019]. Similarly, the genome-wide fusion of TADs via depletion of the TAD-forming DNA-binding factors CTCF or cohesin did not produce significant transcription changes [Despang et al., 2019, Rao et al., 2017]. In the most extensive study of TAD breaks to date – published later than the work presented in this chapter –, Akdemir and colleagues report that roughly 14% of somatic TAD boundary deletions found across 2,658 cancers resulted in a more than twofold change in expression of nearby genes [Akdemir et al., 2020]. In summary, while there is evidence for various response patterns to TAD disruptions, substantial genetic dysregulation appears to be more an exception than the rule. The reasons for this variability in response remain largely unclear to date.

2.1.3 Aims of this study

The work discussed in this chapter is part of an encompassing study on inversion polymorphisms in NHPs, in which my colleagues have used the Strand-Seq

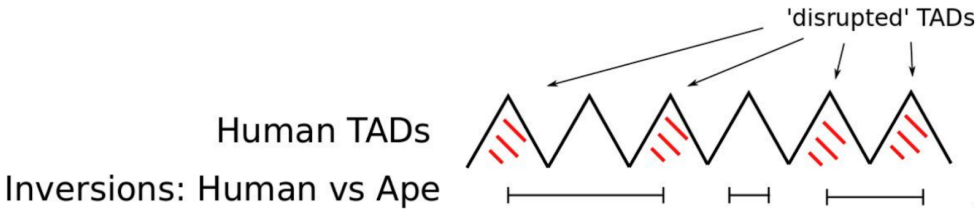

Figure 2.3: Schematic drawing illustrating the concept of 'disrupting' TADs through inversions. Inversions in the genome of great apes can be mapped onto the human genome, where their overlap with existing TADs can be identified. Such 'broken' TADs (red) are candidate regions for genetic re-wiring during the hominid evolution.

technology to identify and analyze the landscape of inversion polymorphisms between the human reference and that of four NHP species to date. Focusing on the functional impact of inversions on genes and gene expression, I conducted analyses to test whether the re-arrangement of TAD structures through inversions may have played an essential role in the hominid evolution. This could, in principle, be achieved by re-wiring selected regulatory neighborhoods and thus enabling re-regulation or diversification of the transcriptome. Being specifically interested in the differences in 3D genome organization between great apes and humans, I focussed on the intersections of these newly identified inversions between human and ape genomes with TADs (Fig. 2.3) and characterized 'broken' TADs in terms of gene expression.

2.2 Identification of 682 inversions between human and ape genomes

My colleagues A. Sanders and D. Porubsky have used the Strand-Seq technique to identify inversions in samples from chimpanzee, bonobo, gorilla and orangutan compared to the human reference genome, leading to a set of 682 simple inversions and 387 inverted duplications [Porubsky et al., 2020]. Owing to the sparseness of Strand-Seq data, the breakpoints of these inversions are denoted with an uncertainty window of ca. 50 kbp, which poses a limitation for in-depth analyses of the breakpoint regions. Nevertheless, the callset comprises a large number of long events (86 variants larger than 1 Mbp), suggesting that many of these inversions have the potential to disrupt TADs and mediate novel chromatin contacts across long chromosomal distances (Fig. 2.4).

Apart from serving as a basis for the analyses presented subsequently, the same

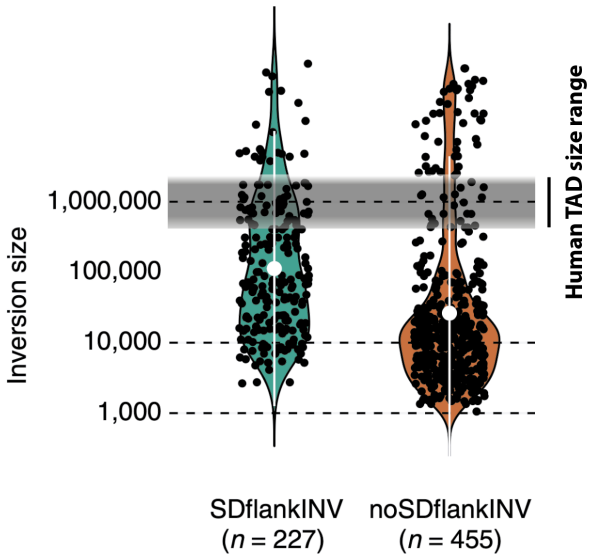

Figure 2.4: The length of discovered inversions overlaid with the typical size range of human TADs. Inversions were further subgrouped based on the presence of flanking segmental duplications at the breakpoints. The callset includes 86 Inversions longer than 1 Mbp, suggesting a large potential for disruptions of TADs. Figure adapted from David Porubsky's and Ashley Sanders' figure in [Porubsky et al., 2020].

callset has also been used for studying inversion recurrence and the formation of breakpoint clusters. However, these phenomena were not the focus of my work and are therefore not further highlighted here. The complete in-depth analysis, including analyses performed by my colleagues, can be found in the published manuscript [Porubsky et al., 2020].

Species	#Strand-seq libs	#Simple Inversions	#Inverted Duplications
Chimpanzee	62	159	71
Bonobo	51	153	63
Gorilla	81	160	122
Orangutan	60	210	131

Table 2.1: Inversion callset stratified by species. In order to exclude effects on differential expression caused by copy number changes, only simple inversions were retained for the study of broken TADs.

2.3 Breakpoints of long inversions co-cluster with TAD boundaries

I initially attempted to identify human TADs which overlap and might thus be disrupted by NHP inversions. As a reference for TAD boundaries in the human genome, I utilized a callset derived from Hi-C data on human embryonic stem cells from a widely recognized paper that described TADs for the first time [Dixon et al., 2012]. These data, defined on the hg19 reference genome, were translated to the GRCh38 reference assembly using the liftOver tool from the UCSC Genome Browser, successfully mapping all but one TAD to the new reference. While early studies have suggested that TADs are widely cell-type independent [Dixon et al., 2012, Rao et al., 2014], newer evidence suggests that their presence may be variable across cell types and -states [Akdemir et al., 2020]. In light of this, the choice of a fixed TAD reference for this study poses a potential weakness of this study (discussed further in section 2.5).

Initial visualizations of the TAD- and inversion locations suggested a non-random co-localization pattern of the two, motivating a formal analysis of the spatial co-distribution. To this end, I measured the distance of the breakpoints of these inversion loci to the closest TAD boundaries separately for short, medium, and very long inversions (<100 kb, <10 Mb, >10 Mb). As a baseline control, inversion-TAD distances were also calculated after n=1000 random permutations of the inversion coordinates using the *RegioneR* package [Gel et al., 2016]. I further specified for each of the four ape species individually a list of human TADs disrupted by inversions, which was used in subsequent analyses. TADs were marked as 'disrupted' when only one breakpoint of a given inversion was positioned within the TAD or as 'intact' otherwise.

The analysis revealed that the breakpoints of long inversions (>100 kbp) tend to co-localize with human TAD boundaries, while shorter inversions showed no such tendency (Fig. 2.5). In contrast, short inversions cause 67.1% fewer TAD disruptions than expected by random (determined through 100-fold randomization of inversion-locations), suggesting that those inversions are strongly depleted from spanning TAD boundaries. These results agree with a prior study conducted in the Gibbon genome, where a similar effect was observed for long inversions [Lazar et al., 2018]. These findings will be further discussed in section 2.5.

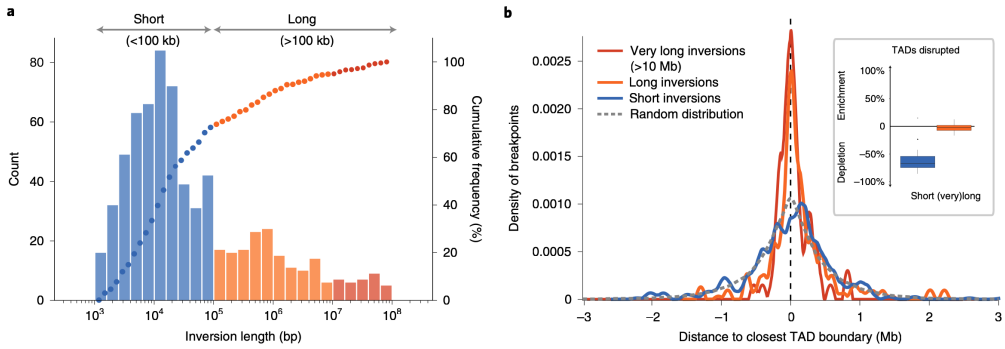

Figure 2.5: Length distribution and spatial co-distribution of NHP inversions and TADs. a Length distribution of all nonredundant simple inversions ('short' <100 kb; blue) and 'long' >100 kb, orange). **b** Distance of each inversion breakpoint (centered at 0) to the closest TAD boundary, stratified by inversion length (color coding according to a). The expected distance distribution for randomly placed breakpoints is indicated by the gray dashed line. The distribution of distances to the closest TAD boundaries for each inversion size category was drawn as a kernel density estimation-fitted curve.

Tissue	Human	Chimpanzee	Bonobo	Gorilla	Orangutan
Brain	6 individuals	6	3	2	2
Cerebellum	2	2	2	2	1
Heart	3	2	2	2	2
Kidney	3	2	2	2	2
Liver	2	2	2	2	2
Testis	2	1	1	1	0

Table 2.2: Bulk RNA-sequencing data used for this study. The underlying data set, stratified by tissue and species, was obtained from [Brawand et al., 2011].

2.4 Inversion-disrupted TADs are enriched for differentially expressed genes

To quantify gene expression in great apes, I utilized existing 75bp paired-end bulk RNA-sequencing data for humans and four NHPs (chimpanzee, bonobo, gorilla, orangutan) across six tissues (brain, cerebellum, heart, kidney, liver, testis) with zero to six individuals each (Median: 2, see Table 2.2) obtained from [Brawand et al., 2011].

The comparison of gene expression across species poses several practical complications, which will be briefly discussed here. First and foremost, the alignment of reads from other species to the human genome assembly is problematic, as divergent gene bodies typically lead to reduced mapping accuracy. This effect generally makes genes from more distant species appear less expressed [Liu et al.,

2014]. In cases where reference assemblies exist, this effect can be mitigated by mapping reads to their respective genomes. However, some problems remain: First, the quality of reference genomes is variable, and errors or minor alleles can impact the number of mapped reads, thus again making genes appear less expressed. Second, copy-number variations can introduce signals that are hard to distinguish from differential expression signals. Lastly, even if copy-number variations are excluded, gene bodies may differ in length or, more prominently, in the use of alternative transcript isoforms, which can hinder direct comparability of expression.

In response to these considerations, RNA-seq reads were first mapped to their respective reference genomes using the STAR aligner [Dobin et al., 2013] and the Ensembl database (version 91) for reference assemblies and annotations [Aken et al., 2017] (reference assemblies; humans: hg38, chimp: Pan_tro_3.0, bonobo: panpan1.1, gorilla: gorGor4, orangutan: PPYG2). Next, read counts were determined using Ensembl v91 gene annotations and the featureCounts tool [Liao et al., 2014]. Subsequent analysis was restricted to a set of 15,117 ortholog genes sampled as follows: Only genes denoted by Ensembl v91 as '1:1:1:1:1' orthologs across human and the four NHP species were considered initially. From this set, 91 X inactivation-escape genes were removed (obtained from [Tukiainen et al., 2017]) due to expected sex-specific expression bias. Lastly, genes were included if they did not display signs of gene expression (>1 fragment per kilobase million) in at least one sample and tissue.

2.4.1 Tissue-specific RNA-seq analysis reveals DE genes

Differential expression levels per gene were calculated using DESeq2 [Love et al., 2014] v.1.24.0, with information about sex included as a cofactor. All NHPs were tested separately against human. Additionally, between-species differential expression analyses were performed for matched tissues (for example, human brain versus chimpanzee brain, human brain versus bonobo brain, human kidney versus orangutan kidney). Using this strategy, 23 differential expression comparisons were conducted (4 species $\times$ 6 tissues, excluding orangutan testis (no data)). Genes with an absolute shrunken fold change >2 and an adjusted Shannon information value (also known as 'surprisal (s) value') below 0.005 were considered as differentially expressed. Fig 2.6 depicts differential expression in the brain as an example. Overall differential expression levels per ape genome were consistent

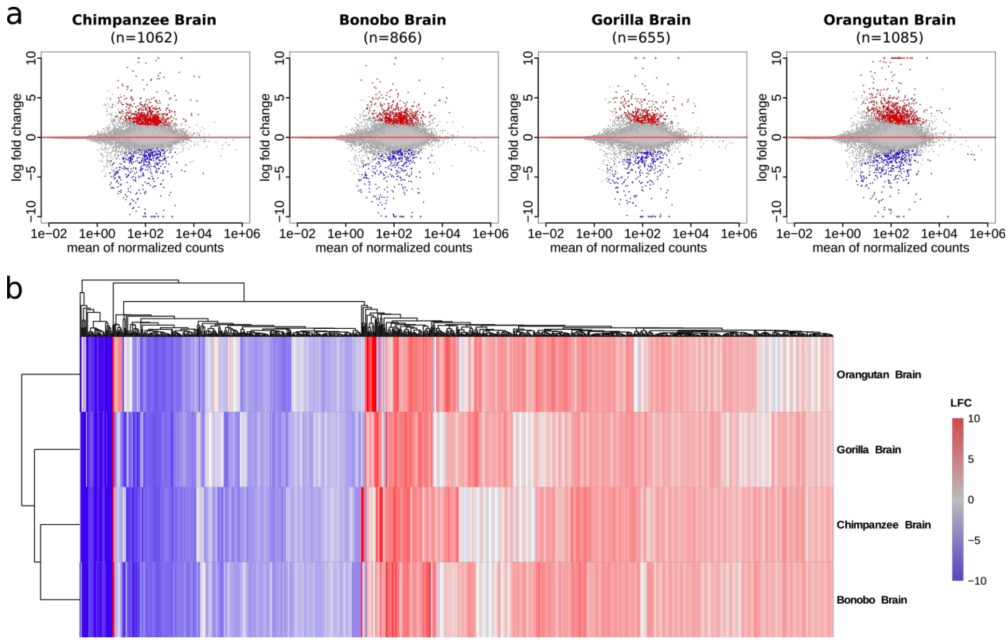

Figure 2.6: DE genes between humans and non-human primate samples. **a** Scatterplots of RNA-seq data show the log_2 fold change versus mean read counts for orthologous genes, comparing the ape lineage to human lineage for brain tissue. In each comparison, differentially expressed (DE) genes that display an absolute fold change >2 and a Shannon information s-value < 0.005 are highlighted, with upregulated genes shown in red, downregulated genes in blue, and the total number of DE genes (N) listed above. **b** Clustered heatmap showing all DE genes (columns) found in brain tissue of at least one ape lineage (rows) LFC: log_2 fold change.

with NHP phylogeny and species divergence (Fig. 2.6b).

Applying the DE analysis described above, a median of 1,499 differentially expressed genes in each NHP (compared to the corresponding human tissue) was observed. Differentially expressed genes were located more frequently (approximately 1.15-fold increase, p = 0.0048, one-sided permutation test) in TADs disrupted by an inversion compared to intact TADs that did not contain an inversion breakpoint (Fig. 2.7a). The possibility arose that this effect may have been predominantly driven by unrecognized copy-number variations of genes contained inside inversion-flanking segmental duplications (SDs), which are known as hotspots of both rapid sequence evolution and errors in sequence assemblies [Sharp et al., 2005]. However, masking SD regions only marginally reduces the observed effect to a 1.13-fold enrichment of DE genes in broken TADs (p = 0.0145, one-sided permutation test). When testing differential expression with respect to inversion breakpoints, I observed more differentially expressed genes near the breakpoints of large inversions (>100 kb) compared to small inversions (<100 kb) (Fig. 2.7b).

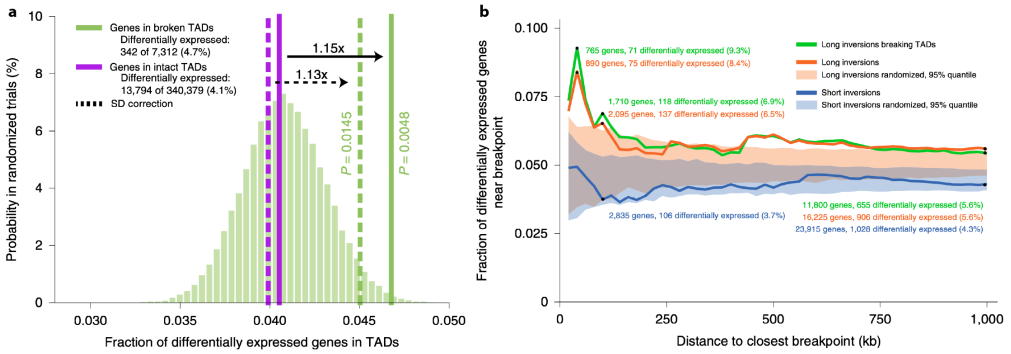

Figure 2.7: Differential gene expression in broken and intact TADs. a Proportion of differentially expressed genes in TADs classified as either 'broken' (solid green line) or 'intact' (solid purple line). The underlying histogram depicts the expected differentially expressed frequency after randomizing TAD labels. The dotted lines represent the differentially expressed proportion after excluding genes in SDs. One-sided permutation testing was used to derive the P values **b** Proportion of differentially expressed genes relative to inversion breakpoints and stratified by inversion length or whether the inversion disrupted a TAD. The shaded areas show the expected differentially expressed proportion measured in matched randomized breakpoints.

2.5 Discussion

Embedded into a more extensive study of inversion polymorphisms in NHPs, the results presented in this section contribute to further understanding the impact of inversions on altering gene expression in the context of the great ape evolution.

In general, the enrichment of co-locating inversion- and TAD breakpoints results in long inversions typically inverting entire TADs, rather than disrupting such structures. It should be noted, however, that the chosen model of TADs may oversimplify the biological structure of genomic loci, as TADs are likely variable across cell types and developmental stages [Akdemir et al., 2020]. While such simplification may, in principle, introduce artifacts, a similar observation to ours has been made on the Gibbon genome before [Lazar et al., 2018], suggesting that the effect is not specific to the choice of species or experimental methods of this study. The co-localization itself could result from several processes: (1) While inversions may occur randomly throughout the genome initially, those that do disrupt TADs may have a higher propensity for negative fitness effects and, thus, a lower likelihood for long-term propagation. Meanwhile, inversions that affect whole TADs may be more frequently fitness-neutral, allowing them to propagate as silent passengers in a population. (2) An alternative explanation could be to assume an unequal prior distribution of inversions, with TAD-breakpoints located preferentially near these inversion breakpoint hotspots. The reverse effect, short

inversions being depleted from TAD boundaries, can be explained analogously. Regarding unequal prior distributions, other co-authors of the enveloping project have identified inversion hotspots characterized by segmental duplications that promote the formation of inversions through non-allelic homologous recombination (NAHR). The implied enrichment of TAD breakpoints in SDs is consistent with documented cases of SD-embedded TAD breakpoints (e.g., in [Cheng et al., 2022]), but the connection still needs to be systematically studied in more detail.

The analysis presented has also highlighted a mild yet statistically significant (1.13 – 1.15-fold) enrichment of differential gene expression in the vicinity of breakpoints of large inversions, which typically break TADs. However, subsequent analysis showed that this enrichment is mainly driven by genes adjacent to the breakpoint site (0 – 150 kbp), suggesting that inversion-associated gene expression changes are more likely to be caused by a direct mechanism associated with the inversion rather than by the disturbance of TADs. A possible objection to this analysis is the lack of sequence resolution of Strand-Seq-based inversion calls (resolution: ca 50 kbp). Indeed, the 'coarse' definition of inversion loci may have obscured further complexity associated with these inversions. This notion becomes clear when contrasting the present study with [Kronenberg et al., 2018], in which the sequence-resolved analysis of human-chimpanzee inversions revealed evidence of secondary CNVs near the breakpoints of inversion in 38% of inversions. A more detailed understanding of these events could lead to a finer definition of 'disrupted' TADs and might help distinguish expression changes caused by the disturbance of whole regulatory neighborhoods from those driven by a more direct mechanism associated directly with a long inversion. A second limitation is owed to the fact that Strand-Seq and RNA-seq data were measured in different individuals from the same species, which may have introduced false signals in the case of inversinos which are polymorphic within an ape species, where occasionally wrong genotypes may have been assumed for subsequent analysis. Despite these limitations, the insights gained from this study agree with a general notion in the field, which suggests that only a subset of TAD disturbance events displays acute effects on gene expression (e.g., [Akdemir et al., 2020, Ghavi-Helm et al., 2019]). Instead, more complex gene regulatory relationships appear to be at play [Ghavi-Helm et al., 2019]. Two years past the publication of the study presented here, Schöpflin and colleagues also drew a similar conclusion concerning differential expression patterns around the breakpoints of chromothripsis and chromoplexy-associated translocations in human patients. In this study, too, differential expression peaked in the initial 150 kbp around breakpoints but did not expand across the affected TADs [Schöpflin et al., 2022].

In the future, more detailed analyses of SVs in the primate evolution will likely provide further molecular insights into their formation and evolutionary role. As discussed previously, the technological revolution in genome assembly has also affected non-human species, including non-human primates, for which the quality of reference genomes has been steeply improving over the last decade [Vollger et al., 2022]. Future studies can draw from these richer resources and apply more fine-grained techniques to resolve inversions in all detail. Such studies will likely provide further insights into the nature of rearrangements contributing to speciation and help distinguish 'passengers' from 'driver' mutations. This notion is especially true for regions containing recurrrent inversions, which are most difficult to study (see chapter 1) but are also likely a crucial component in the human-primate evolution [Vollger et al., 2022, Porubsky et al., 2020].

3

ArbiGent: A general-purpose Strand-Seq based genotyper

This chapter describes the development of a new computational tool, ArbiGent, which has been utilized as the primary source of inversion genotypes in two published projects [Ebert et al., 2021, Porubsky et al., 2022b]. I wish to acknowledge foremost H. Ashraf [Heinrich-Heine University Düsseldorf], who was involved in all stages of the development and helped especially with the mappability correction in section 3.2.1. Furthermore, I thank A. Sanders, J. Korbel, and T. Marschall, who provided ample feedback and advice throughout the whole development phase. My colleague H. Jeong helped set up the Mosaicatcher pipeline and explained the concepts behind the various steps. ArbiGent has been built on top of a conceptual basis developed by S. Meiers [EMBL Heidelberg, Germany] and M. Ghareghani [Heinrich-Heine University Düsseldorf, Germany] in a previous project, who I also thank for their support.

Contents

3.1 Introduction: SV genotyping capabilities inherent to Strand-Seq

The introductory section 1.3 has highlighted specific challenges associated with genotyping inversions. In section 1.3, we have furthermore identified advantages that the Strand-Seq technology offers in identifying inversions compared to other methods, specifically the absence of a need for reads spanning SV breakpoints. Consequently, Strand-Seq had been chosen as the technological basis for inversion calling in the preceding chapter.

From a computational perspective, Strand-Seq requires specialized data processing approaches. As a result, our lab and others have developed several tools to facilitate the analysis of this kind of data over time. Early during the preparation for a project on inversions in the human population (which will be the focus of the subsequent chapter 4), the necessity emerged for a computational method that can genotype inversions across dozens or hundreds of human genomes sequenced with Strand-Seq. In this chapter, I will describe the development of such a new structural variation genotyping method, which we termed *ArbiGent*, and which was realized in collaboration with H. Ashraf, a Ph.D. student from T. Marschall's laboratory at Heinrich-Heine Universität Düsseldorf.

One of the existing tools that can assist Strand-Seq-based SV detection is *breakpointR* (https://github.com/daewoooo/breakpointR; unpublished). The algorithm uses a step-wise binning procedure to dynamically estimate coordinate ranges for *template strand switches*, which are the traces of SV breakpoints in Strand-Seq. *breakpointR* has contributed to inversion calling in several projects [Porubsky et al., 2020, Ebert et al., 2021, Porubsky et al., 2022b] (including the ones chapter 2 4). However, while *breakpointR* calls SV breakpoints confidently, the subsequent task of reviewing these breakpoints and calling structural variation between them is left to the manual curator, somewhat limiting the scalability of this approach. Furthermore, human judgment is never devoid of biases, and especially complex or SD-associated genomic regions can rarely be interpreted unambiguously.

An alternative with a different focus is the *scTRIP* analysis implemented as a workflow in the *MosaiCatcher* tool [Sanders et al., 2020], which was developed as a somatic SV caller. *MosaiCatcher*, which is currently undergoing re-structuring through my colleague T. Weber [EMBL Heidelberg], encompasses several tools and steps which work together to identify SVs above 50-100 kbp in individual cells using a probabilistic read-count-based model. While a detailed description of the method exceeds the scope of this thesis, we will briefly highlight the key steps

here: Initially, aligned Strand-Seq reads are binned into windows of 100 kb. A workflow based on Hidden Markov Models then determines a 'joint segmentation' to effectively identify likely SV breakpoints. Finally, after further steps associated with read phasing and haplotype deconstruction, the probability of SVs in each segment is determined using a negative binomial model of expected vs. observed read counts. The mathematical principle behind this last step also builds the core of the tool developed in this chapter. It will be further elucidated in a later section of this chapter (see 3.2.2). Overall, the workflow is designed to identify SVs – clonal and subclonal – in individual samples and is restricted to a relatively coarse window size of typically 100 kbp, prohibiting the discovery of smaller events.

Another use-case of Strand-Seq data is to genotype known SV locations across individuals – a task that the tools presented here have not been designed for. *breakpointR* specializes in de-novo breakpoint discovery but not in assigning genotypes, and *MosaiCatcher* provides insufficient resolution for small events. Finally, both cannot incorporate information from other sources, especially long-read alignments. The project described in chapter 4, though, requires a genotyping software that (1) could be used to unify calls across samples, (2) verify calls made with other platforms, and (3) integrate information about inversion loci across samples. This chapter describes the development of such a computational tool: ***Arbi****trary segment* ***Genot****yper (Arbigent)*, a Strand-Seq based algorithm to genotype pre-defined inversion segments across large populations.

3.2 Key modules of a new Strand-Seq genotyper

ArbiGent is built as an extension to *MosaiCatcher* and thus shares its mathematical foundations while enhancing the concept to accept pre-defined coordinate ranges – acting as an SV **genotyper** – and to integrate information from individuals cells of the same sample. Furthermore, additional features were included, such as a read-count normalization to enhance sensitivity in difficult-to-map regions, post-hoc inversion phase correction, and a population-based filtering procedure, which will be discussed in this section. The conceptual use case for *ArbiGent* as an inversion genotype is depicted in (Fig. 3.1).

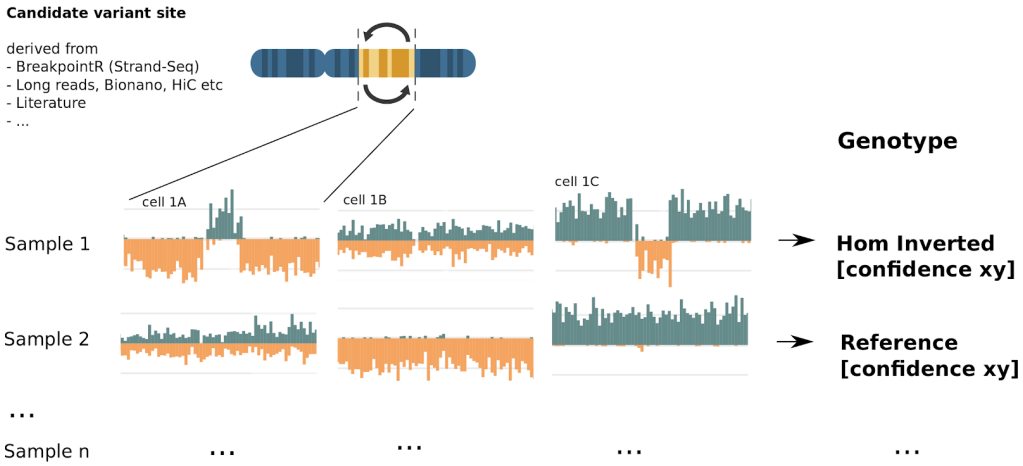

Figure 3.1: Concept of the ArbiGent method. Given arbitrary loci of >500 bp unique sequence, ArbiGent computes inversion genotype likelihoods for inversions and copy-number changes. SV genotype likelihoods derived from individual cells from the same sample are then concatenated by summing up log-likelihoods across cells to result in a combined genotype likelihood estimate per sample and genomic locus of interest.

3.2.1 Read mappability estimation

Based on *MosaiCatcher*, *ArbiGent* also utilizes Strand-Seq read count as a central metric for determining genotype likelihoods. Inversions are often found near the most complex regions of the genome, where a significant portion of short reads can not be mapped unambiguously due to segmental duplications or gaps in the reference [Eslami Rasekh et al., 2017]. This effect can lead to a reduced low read count in these regions, falsely skewing genotype predictions towards lower copy numbers. To quantify the extent of 'lost' reads, H. Ashraf designed an experiment in which the GRCh38 reference was split into 75-mers starting from every reference base, resulting in 3 billion artificial reads. Using the same mapping procedure typically used for Strand-Seq, all 75-mers were mapped back to the genome, and the reads were assigned 'back' to their correct location were counted. Using this method, my colleague was able to create a 'mappability' track that notes the percentage of 'uniquely mappable' basepairs per 100 bp bin. Given an arbitrary segment, this track can be used to express the 'mappability' of this region as a factor between 0 (no reads mapping) and 1 (all reads mapping).

3.2.2 Mappability correction and integration of single cells

I subsequently explored ways to extend existing code by *MosaiCatcher* to add a correction based on read mappability factors obtained by the approach described in the previous section. The original implementation of *MosaiCatcher* determines likelihoods for a set of 72 possible haplotype configurations (such as ref/ref, ref/inv, ref/inv-dup, ref/del, ref/dup, ref/trip, inv/ref, ...) using a negative binomial model:

$$L_{SV} = NB(c_W, s_W, p) \cdot NB(c_C, s_C, p) \tag{3.1}$$

where LLH_{SV} is the likelihood of an SV given read counts c_W and c_C in a segment, and p is the 'scale' parameter determined heuristically by mosaicatcher per sample. The 'shape' parameter s (given by equation 3.2) is, in turn, a function of p, the average number of reads per 100 kbp bin across all chromosomes (measured once per library) and the number of bins that a segment encompasses.

$$s_x = \frac{p}{1-p} \cdot n_{expected\_reads\_per\_bin} \cdot n_{bins\_in\_segment} \tag{3.2}$$

An intuitive way to implement read mappability normalization would be to re-scale measured read counts. If, e.g., 100 reads were counted in a segment of average mappability 50%, the corrected measurement would yield 200 reads. However, such a correction would not preserve the original mean/variance relationship of the data and lead to data distortions, especially for low mappability regions. To understand this notion, consider an example of a segment with 10% read mappability (not unrealistic for SD-rich regions), in a cell with an NB model with $mean_{expected\ reads} = 500$ (Fig. 3.2A). (Fig. 3.2C) simulates the naive approach: ten random draws with 10% of reads are obtained and scaled up by a factor of ten. The normalized draws are over-dispersed compared to the underlying NB distribution and will thus yield distorted results.

A favorable approach is to scale down the expected value of the model to match the reduced mappability (Fig. 3.2D). This strategy can be imagined as slightly 'cheating' the model: if, e.g., 45 reads fall in a segment of 10 kbp length and 10% mappability, we instead pretend that these reads come from a 1 kbp sequence with perfect mappability – and all other calculations remain unchanged. Practically, this is implemented by multiplying the mappability factor with the shape parameter s (Equation 3.3).

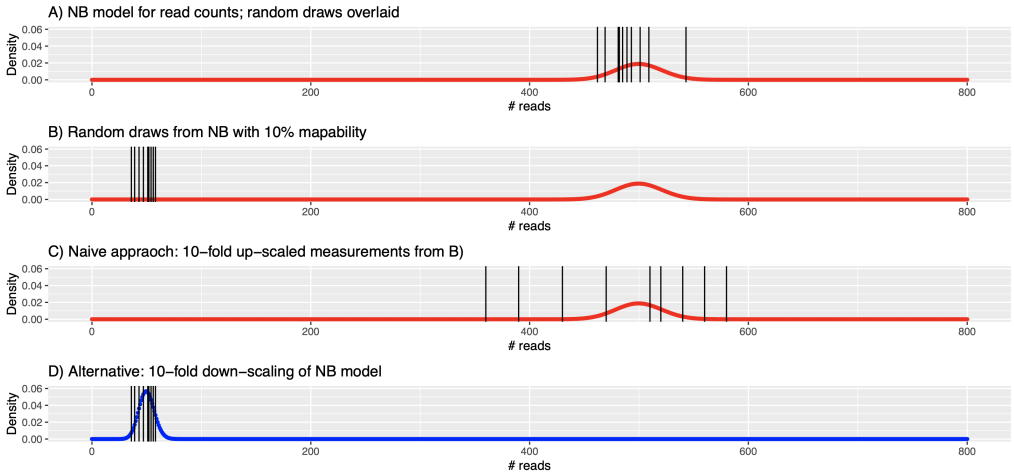

Figure 3.2: Visualization of different read-count normalization strategies. **A** expected read counts for an example segment modeled by a negative binomial model. Ten random draws are overlaid. **B** Random draws simulated with mappability 10%. **C** A naive up-scaling of the reads observed in B does not yield NB-distributed counts. **D** Alternatively, the expected mean of the model can be down-scaled, preserving a meaningful data model.

$$s'_x = \frac{p}{1-p} \cdot n_{expected\_reads\_per\_bin} \cdot n_{bins\_in\_segment} \cdot mappability\_factor \tag{3.3}$$

As the last step, ArbiGent concatenates the likelihoods of individual cells to a bulk likelihood by summing up the log-likelihoods across cells from the same sample. This is done to transfer the predictions from a single-cell basis to sample-wide genotypes. The resulting number can be interpreted as the likelihood of a particular SV state underlying the measurements in each cell. Like in *MosaiCatcher*, the basis for SV calls is finally the likelihood ratio of SV states vs. the reference state: SV genotypes with at least 500 bp of uniquely mappable sequence and a likelihood ratio over reference state $> 10^3$ are considered confident.

3.2.3 Inversion phasing

Inversion calls made by *ArbiGent* are by design phased according to a Strand-Seq-based de-novo phasing procedure implemented by the *StrandPhaseR* tool [Porubsky et al., 2020], in which the assignment of the two haplotypes, h1 and h2,

is random. There emerged, however, a need to combine phased *ArbiGent*-based genotypes with separately phased genotypes made by other technologies. I thus implemented a function, *phase_anchor*, which accepts two vcf files of phased SNPs as input (e.g. one from Strand-Seq phasing, one based on another source like long reads), identifies the intersection of the two and determines the pairwise identities of the haplotypes (i.e. $h1_{vcf1} = h1_{vcf2}$ or $h1_{vcf1} = h2_{vcf2}$). Moreover, since Strand-Seq-based phasing is highly accurate [Porubsky et al., 2020], outputs from this tool can also help to identify incorrectly phased chromosome regions in other vfcs (see Section 3.3.3).

3.2.4 Population-based filtering tools

Calling inversion breakpoints is an error-prone process, and inversion loci given as an input to *ArbiGent* may not always be valid. This notion is especially true in regions of high SD content, where tools like *breakpointR* or long-read-based SV callers are prone to making false calls. Being conceptualized as a population-based genotyper, *ArbiGent* provides options for identifying potentially spurious inversion loci based on population-wide genotypes. Markers assigned to potentially problematic regions are:

1. **False Positive**: segments genotyped as 'reference' in all samples.

2. **Always Complex**: segments with complex genotypes in every sample, where simple inversions or the reference state are never observed.

3. **Mendel Fail**: segments in which genotypes of at least one father-mother-child trio violate mendelian inheritance (given that trios are specified).

4. **Misorient**: segments with a reported homozygous inversion in every sample (indicative of a misoriented reference region).

5. **Inverted Duplication**: for segments with an inverted duplication in at least one sample.

6. **Low Confidence**: segments with less than 500 bp of uniquely mappable sequence (75PE reads).

3.3 Testing & Benchmarking

Being based on the thoroughly tested *MosaiCatcher* tool [Chaisson et al., 2019], *ArbiGent* could similarly be expected to show a high performance in SV calling, as the core functionalities are shared between the two. To test the new features of our tool, we designed several benchmarks and use cases which we describe in this section. To adjust to the flow of this thesis, results from genotyping a large inversion callset across >40 samples with *ArbiGent* are retained for the next chapter, where they will be embedded into context with accompanying efforts around building a comprehensive inversion callset.

3.3.1 Recapitulation and refinement of inversion genotypes

ArbiGent was subjected to benchmark experiments to validate the basic functionality as a genotyper. As a truth set, we initially considered the sample HG00512 (a sample from the 1000 genomes sample pool), for which inversion calls had been determined in a previous paper [Chaisson et al., 2019] using a multi-technology approach and which we considered to be of high quality. Re-genotyping of these inversion loci led to genotype congruence in 113 of 134 loci (84%) (Fig. 3.3A). Out of 21 disagreeing genotypes, 19 accounted for regions that had been classified as heterozygous inversions in the truth set, but *ArbiGent* considers non-inverted or homozygous. We speculate that a proportion of these inversions are, in fact, wrong assignments in the truth set. This notion is supported by the observation that semi-manual inversion calling with *breakpointR* tends to falsely call SD-rich regions as heterozygous inversions due to their similar appearance in the merged 'composite files' [Hanlon et al., 2021].

P. Audano [The Jackson Laboratory, US] also created a separate callset using the PAV tool, a structural variant caller based on de-novo assemblies, which were available for most of the samples from [Ebert et al., 2021]. We re-genotyped 53 inversions with matching genotypes in 46 instances (87%) (Fig. 3.3B) and found no apparent biases, suggesting overall a high genotyping performance.

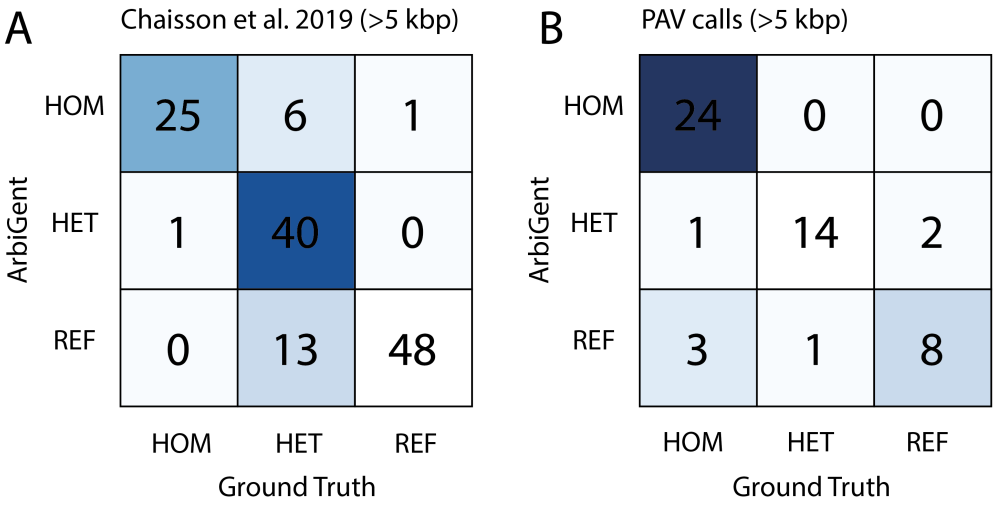

Figure 3.3: Confusion matrices of *ArbiGent* versus two truth sets. A Result of re-genotyping of all simple inversions above 5 kbp in the sample HG00512 reported in [Chaisson et al., 2019]. **B** Comparison of ArbiGent-derived genotypes and 53 inversion calls above 5 kbp made by the PAV caller using phased CLR PacBio reads for 35 samples.

3.3.2 Subsampling experiments and estimated cell number thresholds

Strand-Seq data is typically processed on 96-well-plates, yielding up to 96 individual libraries derived from single cells. However, for technical reasons, not all cells produce viable libraries, leading to typically ~40-70 viable cells per sample. Furthermore, Strand-Seq can also be applied on cell 'pools,' a procedure that typically produces less than ten cells per sample. With help from H. Ashraf, a downsampling experiment was set up to estimate how much the number of cells influences the predictive performance of *ArbiGent*.

As this experiment was conducted in a later stage of the *ArbiGent* development, we were able to utilize inversion loci discovered in the project described in chapter 4. Focussing on sample HG00733, we subsampled random sets of cells in multiple rounds and used ArbiGent in each step to re-genotype inversions, comparing them to the 'truth' set obtained from running *ArbiGent* on the complete set of 115 cells (Fig. 3.4). As expected, inversions above 10 kbp mappable sequence reach near-maximum genotype concordance even with relatively few cells. At the same time, small inversions with hundreds or thousands of bp of mappable sequence profit significantly from a higher number of cells, highlighting the benefit of accumulating information across single cells.

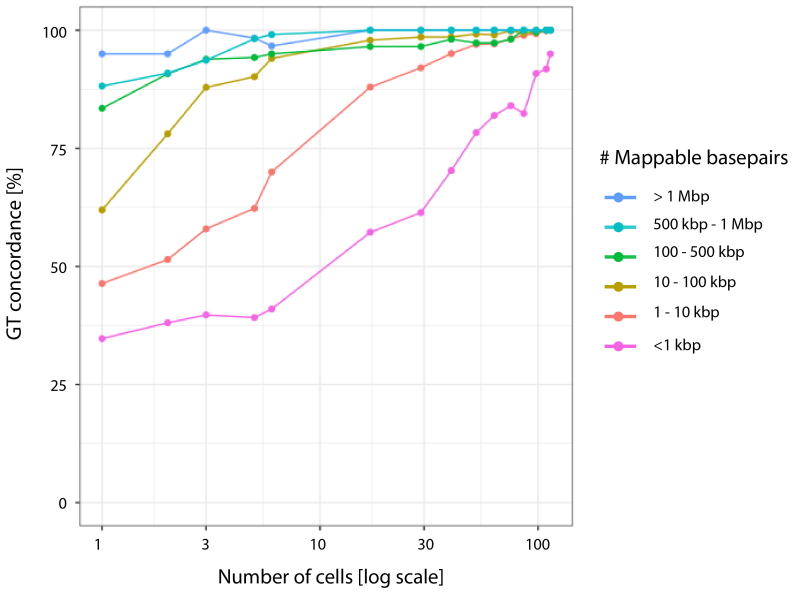

Figure 3.4: Concordance of ArbiGent-based SV calls for different numbers of sub-sampled cells. In each round of the subsampling experiment, SV genotypes of the same inversion set were determined with ArbiGent based on a subset of cells.

3.3.3 Experimental verification of phase correction

To test the performance of the phase correction implemented in *ArbiGent*, we compared Strand-seq-based phased vcfs of 805 chromosomes (35 independent samples * 23 chromosome sets) to the phase assignments created by PAV based on phased genome assemblies of the same samples/chromosomes [Ebert et al., 2021]. We leave the systematic description of this experiment to chapter 4, while highlighting here proof-of-principle phase comparisons of one successful (top left, Fig. 3.5) and three problematic chromosome-wide phase assignments (remaining panels, Fig. 3.5), all of which were further followed up and confirmed by my colleagues P. Audano and P. Ebert [Heinrich-Heine Universität Düsseldorf, Germany], who created the alternative phasing. Section 4.2 will discuss this experiment in context with the underlying data in more detail.

3.4 Discussion

In the preceding chapter, a new Strand-Seq-based genotyping algorithm was presented in response to the challenges associated with inversion detection. The

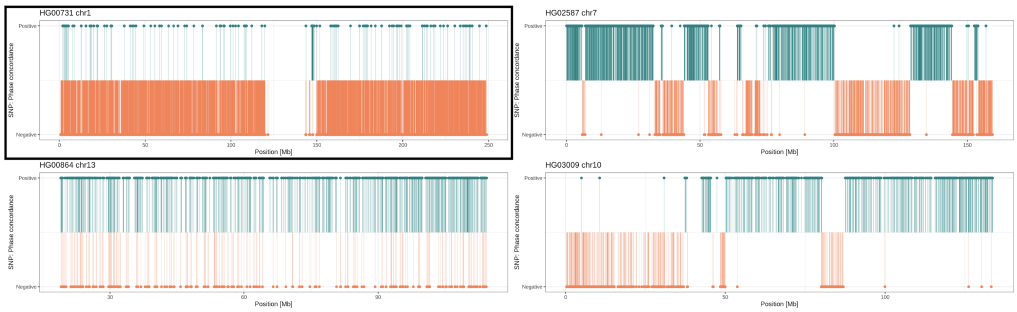

Figure 3.5: Comparison of phased heterozygous (het) SNP identity between calls made by Strand-seq (StrandPhaseR) and PAV. Visualization of phasing from one successful (top left) and three outlier chromosomes. With the two phas ing approaches conducted independently, we expect het SNPs per chromosome to be either close to 100% (h1StrandSeq = h1PAV) or 0% (h1StrandSeq = h2PAV) identical, which suggests concordant or discordant phasing. Inversion genotypes derived from Strand-Seq data on discordant chromosomes were flipped in phase to match the haplotype assignment used for PAV. Each lollipop represents one SNP; color and direction of the lollipop indicate phase agreement (yes/no).

algorithm accurately genotypes SV loci across an arbitrary number of samples by utilizing the full information content found in Strand-Seq data. ArbiGent outperforms previous Strand-Seq-based genotyping approaches, most notably by improving read count normalization in SD-rich regions with low read mappability. Furthermore, while previous approaches have relied on so-called composite files (effectively losing 50% of information content by rejecting libraries with 'WC' and 'CW' strand configuration), ArbiGent utilizes the information inherent in all Strand-Seq libraries and integrates them in a final SV genotyping step. These improvements have reduced a bias in previous Strand-Seq-based approaches to over-call heterozygous inversions (confirmed in [Hanlon et al., 2021]). As additional features, ArbiGent introduces utilities to phase inversions and correctly filter calls based on population-wide metrics.

A significant limitation of the ArbiGent approach is its relatively low sequence resolution, which results from the sparseness of Strand-Seq data which limits the genotyping power for SVs shorter than 1-10 kbp. Also, as a genotyper, ArbiGent does not define likely inversion regions by itself. While this behavior is desired for integration with other methods with higher resolution (e.g., long reads), sample-specific differences, such as alternative breakpoints, can be obscured by this approach. Specifically, more complex events, like nested inversions or inversions associated with deletions, can only be classified correctly if the correct individual segments are passed as input.

It shall be noted here that ArbiGent was paralleled by another Strand-Seq-based inversion genotyper [Hanlon et al., 2021], which was developed for a related,

yet not synonymous task and shares some of its functionalities. This parallel development thus provides an opportunity for cross-testing and improving both algorithms in future iterations (elucidated further in section 3.4). Apart from serving as the primary source of genotypes in two publications, ArbiGent can be expected to contribute to future Strand-Seq-based studies. Currently, the HGSVC is planning to adopt a 'pooled' Strand-Seq technique, aiming at sequencing many more samples than previously by reducing the average number of cells per sample (e.g., to n=3-5 cells, instead of n=50-100). ArbiGent is expected to provide a basis for genotyping large inversion hotspots in these regions. To this end, T. Weber (EMBL Heidelberg) has kindly ensured the integration of ArbiGent as a module in the newest version of the Mosaicatcher pipeline [Weber et al., Manuscript in preparation].

4

Full-spectrum analysis of human inversions reveals hotspots of recurrence associated with genomic disorders

This chapter covers a multi-year collaborative project with members of the Human Genome Structural Variation Consortium (HGSVC) consortium published as a resource article in Cell [Porubsky et al., 2020]. Contents of the manuscript have been re-written here with a focus on the research conducted primarily by me, while work contributed by co-authors is always marked clearly in the text. In particular, this concerns the following sections: the initial inversion discovery was performed by A. Sanders, D. Porubsky, F. Yilmaz, and P. Audano [the latter two: The Jackson Laboratory, US]. Genotyping and filtering of inversion calls were done in close collaboration with H. Ashraf and with input from all co-authors. PCR experiments were conducted at EMBL Heidelberg by E. Garragorri and P. Hasenfeld under my guidance. Figure 4.7 was created by D. Porubsky. Furthermore, H. Ashraf, P. Hsieh, M. Steinrücken [University of Chicago, US], and T. Marschall identified recurrent inversions, and P. Hsieh also created Figure 4.12. Lastly, D. Porubsky conducted an initial analysis on the co-location of inversions and CNVs (section 4.6.1) and several other experiments which could not be reported in this chapter. I thank all co-authors, especially D. Porubsky, H. Ashraf, B. Rodriguez [EMBL Heidelberg, Germany], B. Hsieh as well as T. Marschall, E. Eichler and J. Korbel for their collaborative attitude and immense support.

Contents

4.1 Introduction: Historical and recent inversion callsets

Many factors contribute to the difficulties that are still associated with studying structural variation today, and especially inversions are notoriously challenging to detect and are thus understudied, for reasons discussed in detail earlier (section 1.3). This chapter describes a large-scale effort involving over 20 scientists from multiple labs and countries, which was aimed at identifying a comprehensive set of inversions in 44 human individuals and performing in-depth analyses to shed much-needed light onto this class of structural variation in humans. The current understanding of inversions and their biological background were discussed extensively in section 1.2. Therefore, this introductory section will instead highlight the study's motivation and historical background, which traces back to past efforts from the 1000 genomes consortium and its successor, the Human Genome Structural Variation Consortium. After introducing these past efforts, key challenges and questions associated with this study will be discussed.

4.1.1 Inversions in the Human Genome SV Consortium

The Human Genome Structural Variation Consortium (*HGSVC*) has formed as one of several quasi-successors of the 1000 Genomes Project, which in turn had the goal of providing a comprehensive catalog of genomic variation based on originally 1,000, but eventually, 3,202 human genomes of diverse origin [1000 Genomes Project Consortium et al., 2015]. Indeed, the effort succeeded at providing catalogs of single-nucleotide polymorphisms and Structural Variants which were unmatched at the time and led to a cascade of technological advances and biological findings (reviewed, e.g., in [Zheng-Bradley and Flicek, 2017]). Structural variation discovery was then primarily performed using 100bp Illumina WGS reads with 7.4-fold genome coverage. While enabling a survey of simpler structural variations, the approach still posed severe limitations for capturing more complex variants or approach regions with high SD content [Sudmant et al., 2015]. While this callset also included 786 inversions (20 of which were breakpoint-resolved – partially with the help of PacBio reads), no inversions above 100 kbp were reported at all (Fig. 4.1).

The 1000 genomes project displayed the difficulties associated with a comprehensive survey of SVs. One of its quasi-successor projects, initiated by the *HGSVC*, next focussed on developing new approaches towards this goal. The most important concept was integrating the most successful SV calling technologies

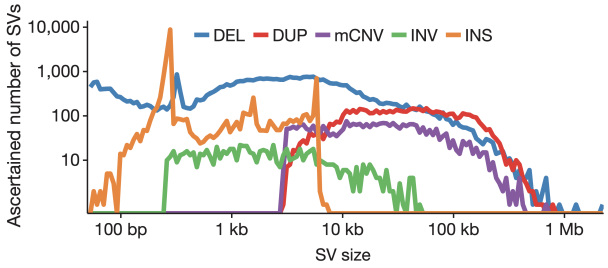

Figure 4.1: SV calls reported in [Sudmant et al., 2015]. Figure taken from the original publication. Among 786 reported inversions, none were larger than 100 kbp, highlighting the technological limitations of short-read sequencing for detecting long inversions.

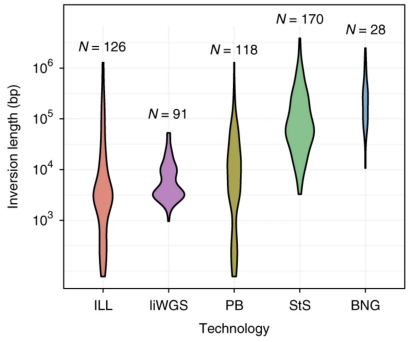

Figure 4.2: Inversion loci reported in [Chaisson et al., 2019]. Figure taken from the original publication. The calls are subdivided by technology, and the number of total inversions contributed by each is listed above each violin plot. The majority of inversions larger than 10^5 bp (100 kbp) was contributed by Strand-Seq. ILL Illumina. liWGS long-insert whole-genome sequencing libraries. PB Pacific Biosciences. StS Strand-seq. BNG Bionano Genomics.

(Illumina short reads, Pacbio long reads, Strand-seq, and 10x Chromium) into a combined SV calling approach, where the technologies would complement each other. On the flip side, the approach was highly intensive in cost and effort, so only nine individuals could be screened initially [Chaisson et al., 2019]. The approach revealed 27,622 SVs above 50bp (compared to 68,645 SVs in $>$ 2,500 samples in [Sudmant et al., 2015]), as well as 308 inversion loci (227 copy-neutral). Strikingly, this callset included around 100 inversion loci above 100 kbp, highlighting the increase in sensitivity for such events significantly contributed by Strand-Seq (Fig. 4.2). Most importantly, the study showcased that combining multiple algorithms and data types was state-of-the-art for maximizing SV discovery. Valid for all SVs, this lesson was especially true for inversions.

In the lastest stage of their project, members of the *HGSVC* built on these insights and expanded the number of humans analyzed to 32, again to set a benchmark for SV discovery and derive biological insights [Ebert et al., 2021].

Owing to rapid developments in long-read sequencing technologies and genome assembly, it had become feasible to creatine phased de-novo genome assemblies for all 32 samples, or 64 haplotypes. This was achieved, again, by combining multiple data sources, especially CLR and HIFI PacBio reads, with read phasing assisted by Strand-Seq [Ghareghani et al., 2018]. The approach yielded a total of 107,590 SV loci and 316 inversions. While identifying these inversions had already required a substantial amount of curation, it became clear that the full spectrum of inversions is likely not yet reflected in this number.

4.1.2 Aims of this study

Chapter 1 has discussed open questions revolving around inversions, and the study presented here offered a chance to answer many of them. The ensuing project around human inversions, which forms the basis of this chapter, fell broadly into two stages, each with its technical challenges and open questions.

The initial goal was to identify, genotype and validate the spectrum of inversions across the genomes of 41 individuals – for the first time, in a truly comprehensive manner and across all length scales. This task, set out at the edge of technical feasibility, was viable only due to the richness of data, recent technological and computational developments (including the work presented in chapter 3), and the expertise of many members of the project, and will be the focus of the first half of this chapter.

After obtaining this novel callset, many assumptions and hypotheses around inversions could be revisited. These hypotheses include relatively simple questions regarding the number, allele frequencies, and size distribution of inversions, as well as their breakpoint architectures, flanking sequences, and formation mechanisms. In addition, it has been proposed that inversions fall into mechanistically distinct classes, but the spectrum of these classes and their specific properties still need to be determined. Furthermore, this study posed an opportunity to identify the propensity of the human genome for inversion recurrence and the consequences that this little-studied phenomenon may have on a molecular and phenotypical basis. Lastly, chapter 1.2 has described the three known modes in which inversions are thought to interact with genome function: disruption of gene- or regulatory neighborhoods, suppression of homologous recombination, and association with secondary duplications and deletions. This chapter will also examine how prominent such inversion-mediated effects are in the germline,

and propose models for the molecular basis of the associations of inversions with copy-number variations.

4.2 Iterative construction of a comprehensive inversion callset

Detecting inversions across the whole size range is challenging. While this project benefits from abundant and diverse data sources, a novel workflow for discovering and genotyping balanced inversions was developed, which integrates these technologies into contributing to a common callset. The description of this new methodology, and the resulting inversion callset, form the core the following section.

4.2.1 Initial inversion discovery and genotyping with ArbiGent

Due to inherent difficulties in detecting balanced inversions (discussed in section 1.3), an initial round of inversion discovery was performed utilizing three orthogonal platforms: Strand-Seq, The assembly-based SV caller *PAV* [Ebert et al., 2021] and Bionano optical mapping [Lam et al., 2012]). This stage was performed by my colleagues, primarily D. Porubsky, A. Sanders (using *breakpointR* and manual inspection of Strand-Seq data), P. Audano (*PAV*) and F. Yilmaz (Bionano). P. Audano then concatenated these individual callsets using a procedure described in a previous publication from the consortium [Ebert et al., 2021], leading to an initial callset of 618 inversion *candidate* loci.

The inversion genotyping and filtering steps were performed with ArbiGent, a custom-developed Strand-Seq-based genotyper described in chapter 3. Utilizing Strand-Seq data available for 41 samples, ArbiGent was used to assign genotypes to all 618 inversions across samples. According to ArbiGent's population-based filtering scheme, calls were filtered out from the callset if labeled as complex ('alwayscomplex', 'alwayscomplex-INVDUP') across all samples or as 'false positive'. This initial filtering reduced the callset to 419 inversion loci, or 68% of the original callset (Fig. 4.3A). The high false positive rate in the original callset reflects technology-specific biases, such as a high FP rate for short StrandSeq or long *PAV*-based calls. Likewise, it highlights the importance of orthogonally validating inversion calls. Loci identified using Bionano displayed the highest rate of rejected calls, with 47 / 82 (57%) of inversions labeled as 'false positive'. Reas-

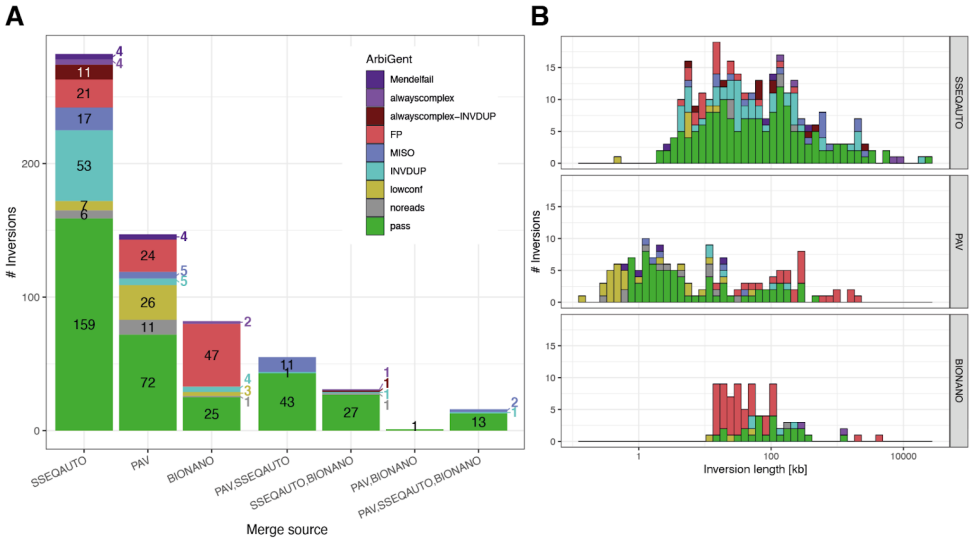

Figure 4.3: ArbiGent-based genotypes for the merged inversion callset. **A)** ArbiGent inversion classification stratified by technologies used for detection. Loci were rejected from the final inversion callset if they showed evidence of absence ('FP', false positive) or were classified as 'complex' in all samples. **B)** ArbiGent classifications, stratified by event length and technique used for initial prediction. The distributions reveal technological biases, such as an enrichment of 'FP' calls in long *PAV*-derived and short Bionano-derived calls.

suringly, only 2 of 103 (2%) inversion loci predicted by more than one orthogonal method were rejected by this approach. To finalize genotyping, predictions with low confidence (likelihood ratio over reference state of $< 10^3$) - most frequently observed in inversions < 5 kbp) were subsequently exchanged with calls based on *PAV*.

Phase correction

The phasing of inversions genotyped by ArbiGent is based on the StrandPhaseR tool. While this phasing is consistent by itself, inversion calls were additionally synchronized with the de-novo assemblies in [Ebert et al., 2021] to facilitate subsequent analysis – essentially ensuring that the assignment of 'h1' and 'h2' refers to the identical haplotypes in both data sets. Using the phase synchronization implemented in ArbiGent (described in section 3.2.3), 709/805 chromosomes (35 independent samples * 23 chromosome sets) could be phase-synchronized. However, the remaining 9 chromosomes showed potential errors in the phasing of the haplotype assemblies (examples depicted earlier in Fig. 3.5), and associated inversion calls were considered as 'unphased'.

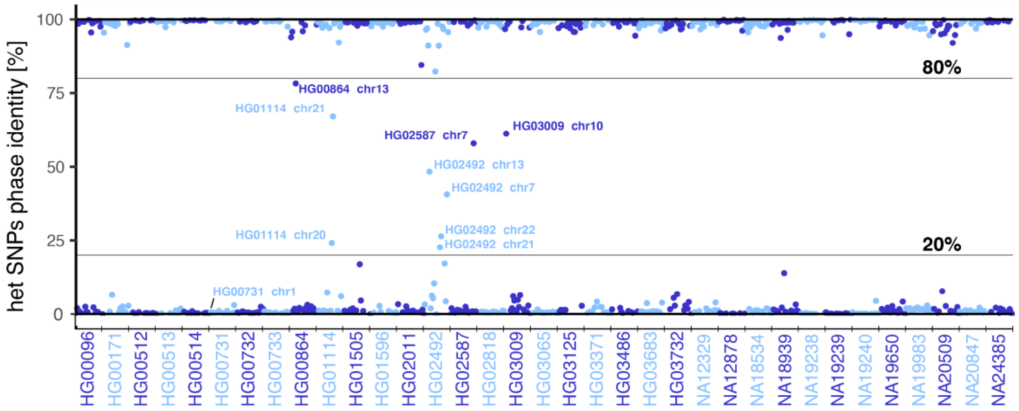

Figure 4.4: Comparison between Strand-seq-based and assembly-based (PAV) inversion phasing. Comparison of phased heterozygous (het) SNP identity between calls made by Strand-seq (StrandPhaseR) and PAV. With the two phasing approaches conducted independently, we expect het SNPs per chromosome to be either close to 100% (h1StrandSeq = h1PAV) or 0% (h1StrandSeq = h2PAV), which suggests concordance or discordance concerning phasing. Inversion genotypes derived from Strand-Seq data on discordant chromosomes were flipped in phase to match the haplotype assignment used for PAV. Chromosomes with a het-SNP identity between 20-80% were considered outliers and were not phase-adjusted.

4.2.2 Experimental validation and assembly-based breakpoint refinement

With the seemingly complete, a set of ten inversions was subjected to validation via a specialized approach based on polymerase chain reaction (PCR) (Fig. 4.5). For this purpose, I chose random inversions of various size ranges (0.5 kbp-366 kbp) and selected samples so that each inversion could be tested once in reference, heterozygous and homozygous state. Next, I used a custom script *design_primers.sh* by T. Rausch [EMBL Heidelberg] to define sets of primers for each inversion according to a 'Four primer' strategy (Fig. 4.5A,B). Our technicians P. Hasenfeld and E. Benito-Garragorri conducted the bench experiments.

However, PCR products for 8/10 inversions were inconclusive or not indicative of an inversion. After possible failure modes in the experimental procedure had been excluded, the only explanations for this result remained problems with the breakpoint locations or genotypes.

Assembly-based breakpoint refinement and validation

In order to identify the reasons for the failure of inversion validations, all 418 inversion loci were subjected to manual review by evaluating dotplot alignments

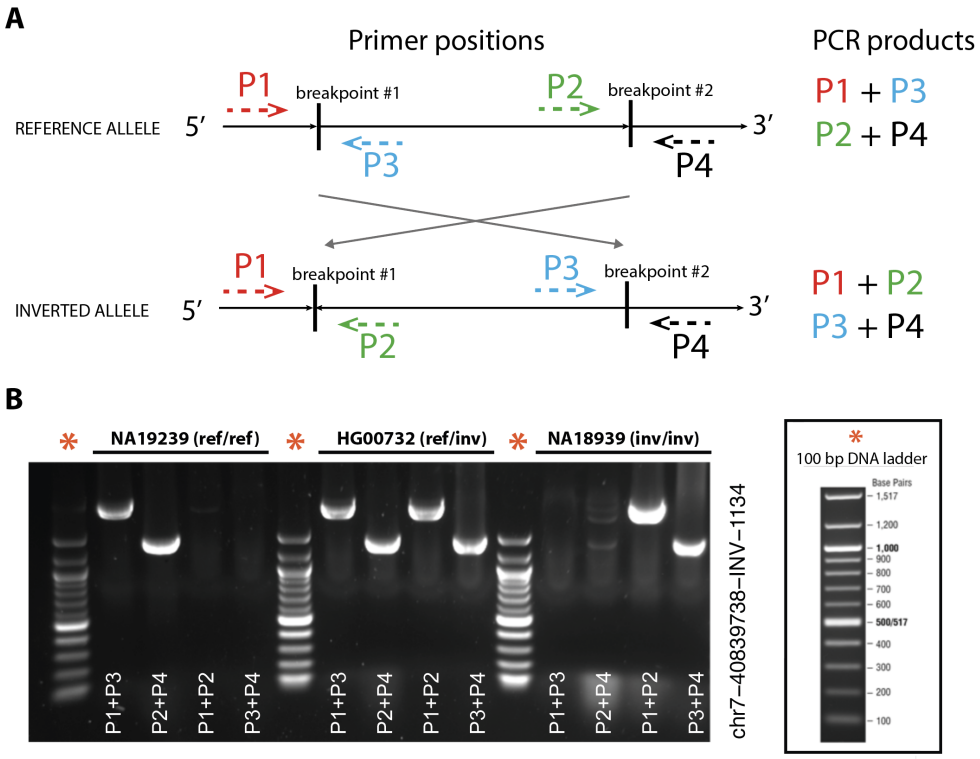

Figure 4.5: Inversion validation via PCR. A Schematic view of the expected primer positions and orientation in reference and inverted alleles. Each inversion breakpoint is spanned by a pair of primers on opposing ends, leading to expected PCR products for primer pairs "P1/P3" and "P2/P4" in reference haplotypes and "P1+P2" and "P3+P4" in inverted haplotypes. **B** Example of a successful PCR validation of an inversion in reference (left), heterozygous (middle), and homozygous (right) state.

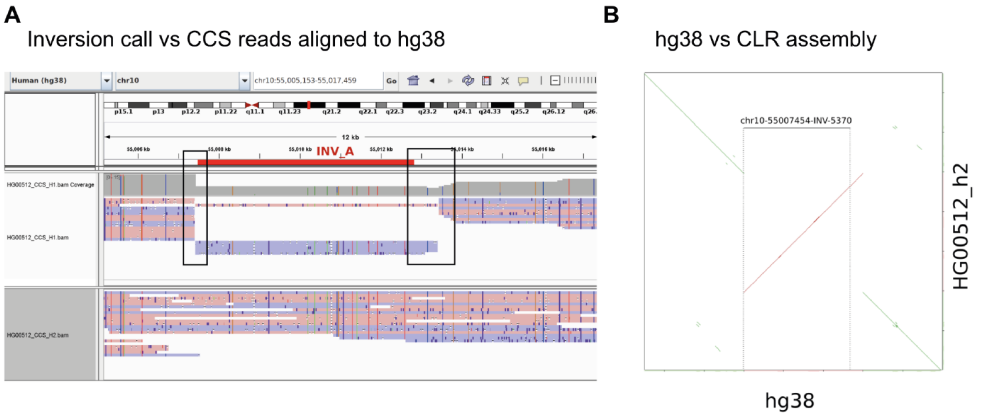

Figure 4.6: Correction of an inaccurate inversion call using long reads and genome assemblies. A) Aligned CCS PacBio reads vs. inversion position predicted by the *PAV* caller. **B)** A dotplot comparing the inversion sequence region in hg38 to the homologous region in the inversion-carrying sample HG00512_h2. The inversion prediction is overlayed in black. CCS reads and the genome assembly agree on a second inversion breakpoint that deviates 613 bp from the prediction.

made by alignments of phased genome assemblies (taken from [Ebert et al., 2021]) against the GRCh38 reference genomes. For this purpose, I designed a workflow that takes as input a sequence on the reference genome (e.g., the region of a putative inversion) and uses the *minimap2* aligner [Li, 2018] to extract the homologous sequence in another assembly by transforming the coordinates of the neighboring 'anchor' regions. Per inversion – sample pair, the two sequences, reference, and alternative assembly, were then compared visually using a dotplot matrix, implemented in the *dotplotly* package (https://github.com/tpoorten/dotPlotly, unpublished) (Fig. 4.6B). This final verification step resulted in rejecting 19 likely false positive inversion loci. The procedure furthermore enabled the refinement of 183/418 (44%) of inversions, each entirely spanned by a single contig, to near-basepair precision (50 bp, microhomology not considered). Additionally, adjacent SVs such as indels and duplications were annotated in this process, substantially improving the scope of the callset. Finally, all inversions with modified breakpoint annotations were subjected to a second round of genotyping in ArbiGent. Final PCR tests confirmed all refined breakpoints in all 10/10 selected inversions (one example shown in Fig. 4.5).

Callset overview

An overview plot created by D. Porubsky reveals critical features of the callset (Fig. 4.7), which recapitulates previous observations in the literature. First, while

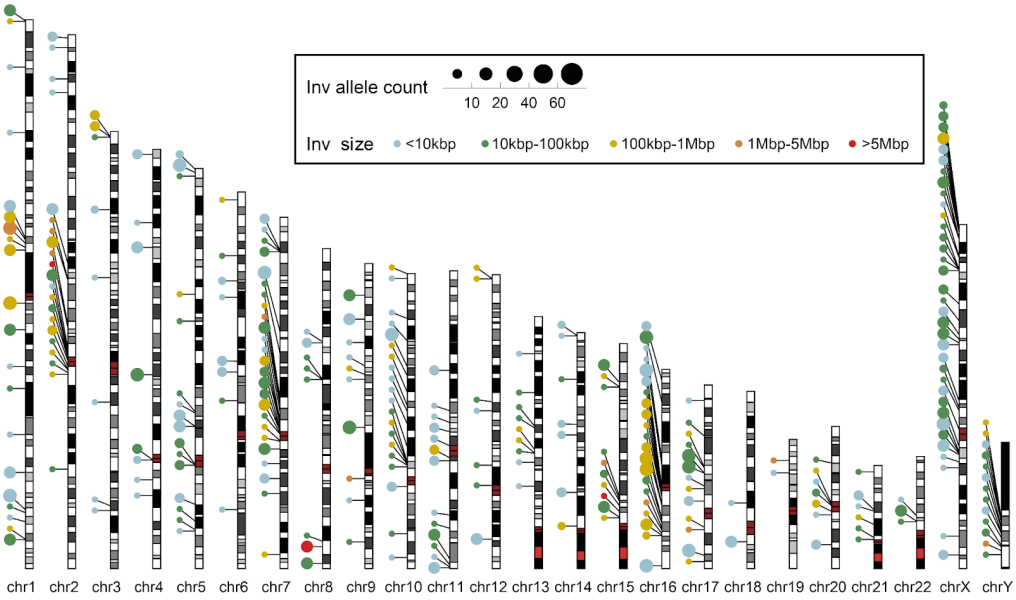

Figure 4.7: Overview plot over the position, size and allele frequency of all balanced inversions in the callset. Inversions can be seen to cluster in hotspots (e.g., on chromsosomes 2, 7 and 16). The figure was created by David Porubsky, one of the co-leading authors of the project.

we identified inversions on all 24 chromosomes, specific chromosomes seem to be more prone to inversions, such as chromosomes 2, 7, 16, and X (shown, e.g., in great apes, [Porubsky et al., 2020] and humans [Chaisson et al., 2019]). Second, the callset includes several inversions spanning multiple Mbp. Though known previously, this is a remarkable feature of inversions, given that other SVs of such length are very rare, especially in the germline of healthy individuals (long inversions reported, e.g., in [Giner-Delgado et al., 2019, Chaisson et al., 2019]). Third, the callset reveals 'hotspots' of inversions, which often focus, e.g., around the centromeres of chromosomes 2 and 7. This finding is consistent with inversion formation through NAHR, mediated by SDs which are enriched in these regions (compare e.g., [Sanders et al., 2016]).

4.3 Identified inversions cluster into three distinct classes

The manual curation of 183 inversions through pairwise sequence alignments provided a basis for examining the inversion breakpoint structures of these inversions, revealing three distinct classes of inversions (Figure 4.8A). 101/183 loci (55%) displayed segmental duplications (SDs) of >90% sequence identity

at their flanks. Inversions flanked by long (>10 kbp), highly identical SD are thought to form through non-allelic homologous recombination [Bailey and Eichler, 2006]. This notion suggests *NAHR* as the predominant mechanism for balanced inversion formation in our dataset, accounting especially for almost all large events (>50 kbp) (Fig. 4.8A). Inversion calls additionally display a positive correlation between repeat length and inversion length (Fig. 4.8B), mirroring previous findings in inversions in great apes [Porubsky et al., 2020].

Further 31/183 loci (17%) were flanked by highly similar repeats which mapped to mobile element sequences (L1: n=22, Alu: n=9). Most (21/22, 95%) inversion-flanking L1 pairs display >90% pairwise sequence identity (median: 97.2%), in sharp contrast to Alu pairs, where this is the case for only 1/9 (11%) (Fig. 4.9A). Additionally, pairwise alignments revealed that six out of nine Alu/Alu-flanked inversions show nearby sequence gains or losses of 35–701 bp in size (Fig. 4.9B). This observation suggests that Alu-flanked inversions may form through a different rearrangement process, as described for Alu-mediated deletions [Morales et al., 2015].

The remaining n=51 inversions did not display large repeats at their breakpoints but were frequently (23/51, 45%) flanked by or nested in adjacent insertions and deletions. Such inversions, lacking large homologous sequences at their breakpoints, are likely to have formed through canonical or alternative non-homologous end-joining (c-NHEJ, a-NHEJ/MMEJ) [McVey and Lee, 2008], or replication-based mechanisms such as microhomology-mediated break-induced replication (MMBIR) [Carvalho and Lupski, 2016]. For future reference, these inversions present well-suited targets for closer mechanistic study via a fine examination of breakpoints at basepair precision to categorize microhomology, templated insertions, and other DNA signatures at the breakpoints and thus determine likely formation mechanisms per event.

4.3.1 Analysis of class-specific overlap with genes and genomic elements

Inversion loci were next overlapped with gene annotations from gencode v35 [Frankish et al., 2019] to assess the potential of different classes of inversions to disrupt genes. This analysis reveals class-specific differences, with SD-mediated inversions displaying more frequent overlap with genes, in contrast to L1/Alu mediated and non-repeat mediated inversions, primarily found in intergenic

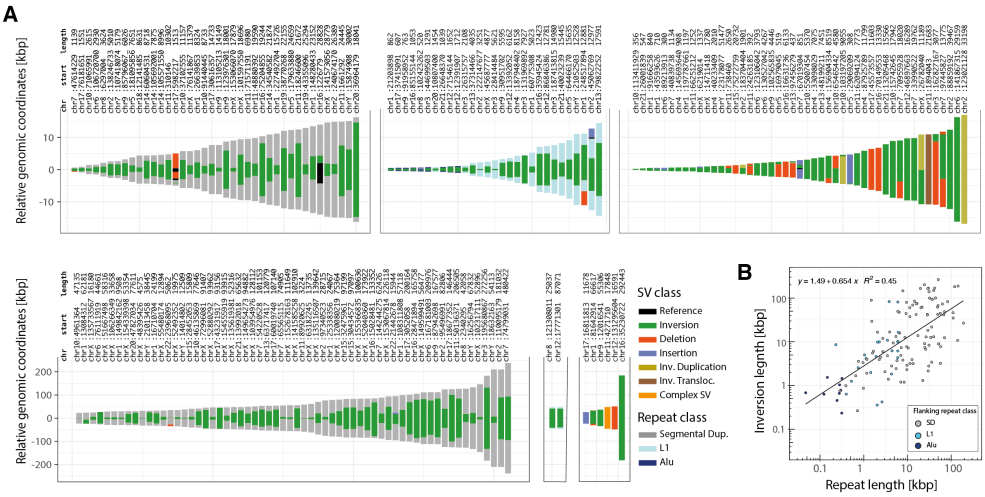

Figure 4.8: Inversion architecture of 183 breakpoint-resolved inversions. **A** Structural conformations for balanced inversions and their flanks. Inversions are grouped by likely formation mechanism (left-to-right: SD-mediated, mobile element-mediated, non-repeat-mediated) and vertically by total event length (top-to-bottom: <40 kbp, R40 kbp in size). Inv. Transloc., inverted translocation. **B** Flanking inverted repeat length correlates with event size.

and intronic regions (Fig. 4.8). Furthermore, there exists the potential for inversion-mediated fusion transcripts between the gene-pairs *RSRP1-RHCE*, *CTRB2-CTRB1* and *RHOXF1AS1-NKAPP1* (The two former ones being famous cases of genes near functionally relevant inversions [Wang et al., 2020] [Rosendahl et al., 2018]). Additionally, several gene disruptions were noted, affecting all or some transcripts from 6 coding and 5 non-coding genes (coding: *OR2G6, IFITM2, PDXDC1, CCDC144B, CCDC200, RBL1*).

4.4 New inversion eQTLs revealed by gene expression analyses

Inversions can potentially affect gene expression (Giner-Delgado et al., 2019) by disrupting genes, disturbing regulatory associations, or facilitating de-novo variation by suppressing homologous recombination [Giglio et al., 2001]. Utilizing bulk deep transcriptome data created for this purpose from lymphoblastic cell lines for 33/41 samples, inversions were systematically tested for associations with gene expression changes. RNA-seq preprocessing was performed following a protocol described in a previous publication from the HGSVC consortium [Ebert et al., 2021]. eQTL association tests were conducted using the LIMIX tool [Lippert et al., 2014], complementing the analysis dataset with additional >16

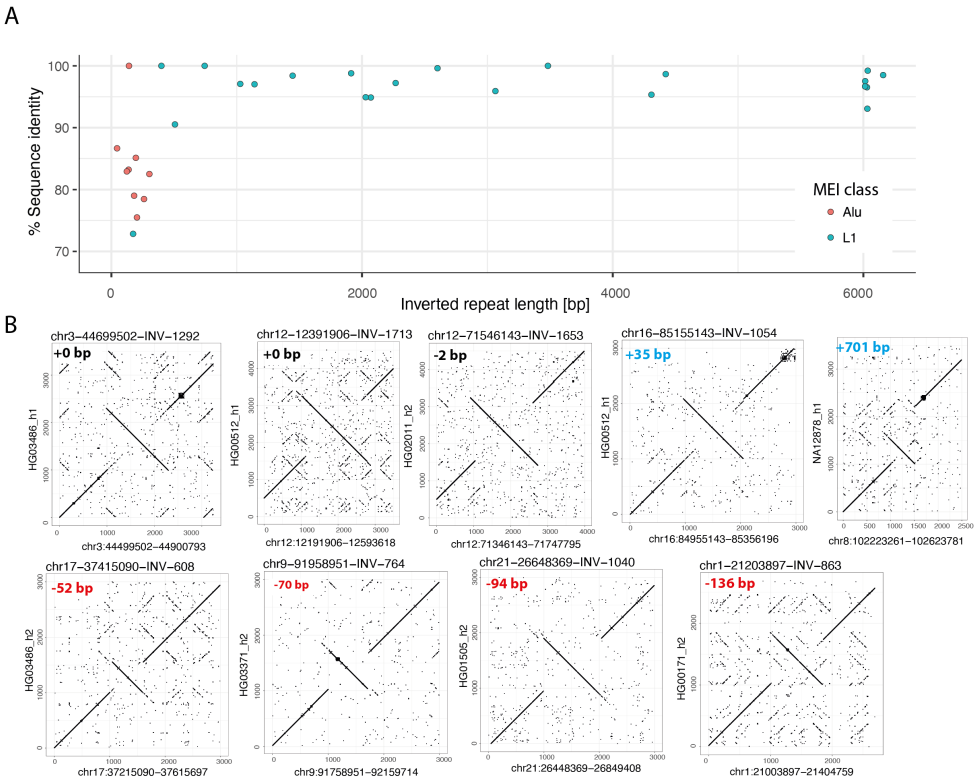

Figure 4.9: Length and sequence identity of balanced inversion-flanking repeats mapping to mobile elements. **A** 31 inverted repeat pairs flanking balanced inversions were found to map to mobile elements (MEI). Compared with pairs of flanking Alu repeats, pairs of flanking L1 repeats are significantly longer (median: 2,435 bp and 181 bp, respectively, p = 3.1e-06, one-sided t-test) and display higher sequence identity (median: 97.2% vs. 82.92%, p = 0.00049, one-sided t-test). **B** Dotplot visualization of the nine inversions flanked by Alu elements. The number of base pairs gained (blue) or lost (red) across the whole inversion locus in the inverted haplotype is indicated in the top left corner of each dotplot.

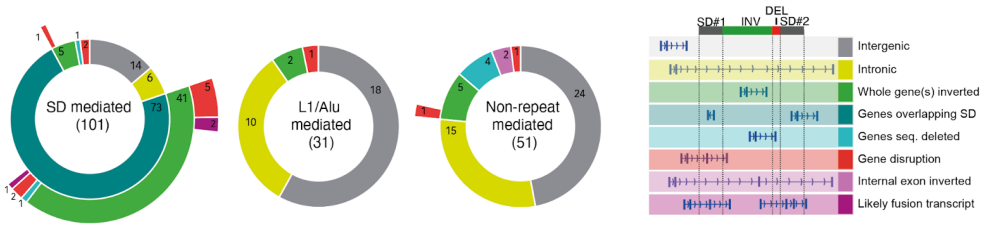

Figure 4.10: Overlap for inversion types established in Fig. 4.8 with functional annotations of the genome. overlap between inversions and genes is reported separately for inversions mediated by SDs (left circle), Mobile elements (middle circle), and non-repeat mediated inversions (right circle). The definition of overlap categories, such as 'Gene disruption', is illustrated in the box on the right.

Million SNPs and >100,000 SNVs taken from Ebert et al. [Ebert et al., 2021]. Despite the relatively low sample number and the limitation to only one cell type, several known and unknown associations have been identified with this approach, including two eQTLs where the inversion is the lead variant (*MAPK8IP1P2, AC126544.2*). Figure 4.11 displays identified gene-inversion associations involving the genes *ATP13A2*, *OR4C6*, *MAGEH1* and *RP11-460N20.4.2*.

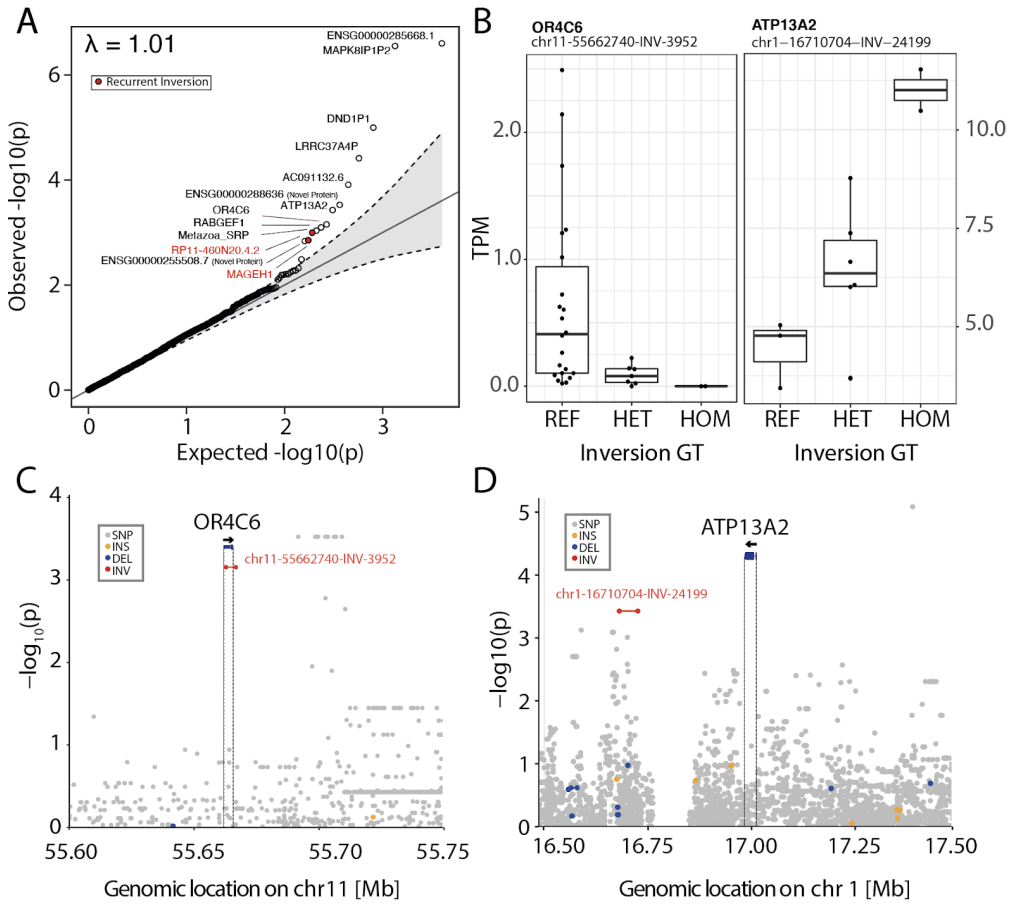

Figure 4.11: eQTL analysis of inversions outside of L1-internal sequences and gene expression in 33 samples. A qq-plot showing the highest observed vs. expected p-values and the genes most strongly affected by an inversion. Genes associated with a recurrent inversion are highlighted in red. **B** Expression values for two inversion eQTLs *OR4C6* and *ATP13A2*. **C** Manhattan plot for textitOR4C6 locus. The gene is overlapped by an inversion (red). **D** Manhattan plot for the ATPase Cation Transporting protein *ATP13A2*, located roughly 250 kbp downstream of an inversion at [hg38]chr1:16710704-16734903. The inversion overlaps an enhancer-rich region with further high-scoring SNP variants.

4.5 Identification of widespread inversion recurrence

Section 1.2.1 has introduced the concept of inversion recurrence. In comparison with other SV classes (taken from Ebert et al. [Ebert et al., 2021]), our inversion callset displays an enrichment of common (minor allele frequency, MAF > 5%) alleles (67% of inversions vs. 48% of other SVs, $p = 2.6 \times 10^{-11}$). Such enrichment can signify recent inversion recurrence as this process spawns new instances of both alleles, effectively acting as a balancing force towards 50% inverted allele frequency, as has been noted previously [Aguado et al., 2014, Zody et al., 2008]. My colleagues H. Ashraf, P. Hsieh and M. Steinrücken have devised two complementary computational frameworks to discover recurrent inversions systematically. These approaches based on genomic *coalescence*, a framework that provides models for the evolution of alleles from a common ancestor [Arenas and Posada, 2014]. Both algorithms exploit the principal observation that recurrent inversions (re)appear in diverse genomic contexts and thus avoid co-segregation with SNPs. The two complementary approaches differ in the data type used (haplotype-resolved Strand-seq reads vs integrated Strand-Seq and PacBio data) and many conceptual details. Fig. 4.12 (created by P. Hsieh) illustrates the concept on phylogenetic trees of a recurrent and a non-recurrent inversion. This analysis eventually led to 40 inversions being confidently labeled as 'recurrent' by both methods, a number 2.5 to 3-times higher than previous estimates [Giner-Delgado et al., 2019, Puig et al., 2020]. Additionally, one of the methods was designed to calculate recurrence rates, which was estimated to range between 3.4×10^6 and 2.7×10^4 per locus per generation. All recurrent inversions together cover >20 Mbp of sequence, or 0.6% of the human genome.

In principle, inversion breakpoints can also be informative of recurrence, as (1) separate instances of inversions might occasionally utilize different breakpoints inside their SDs, and (2) inversion events might spawn particular genomic signatures at the breakpoints ("scars"). However, initial analysis performed by David Porubsky suggests that inversions between segmental duplications are typically accompanied by local gene conversion at the flanks, leaving extended break "regions" rather than a clear-cut breakpoint behind. The nature of these break regions still needs to be understood better, and new methods will have to be developed to study those in more detail.

As a third approach to identifying inversion recurrence, I manually examined pairwise dotplot visualizations of all inversion regions in de-novo-assembled genomes, searching for patterns indicative of recurrence. Two inversion loci – both marked as 'recurrent' by the coalescent-based approach – displayed such clues: First, a 166 kbp inversion on chr16p12.1-p11.2 displayed an 11 kbp deletion

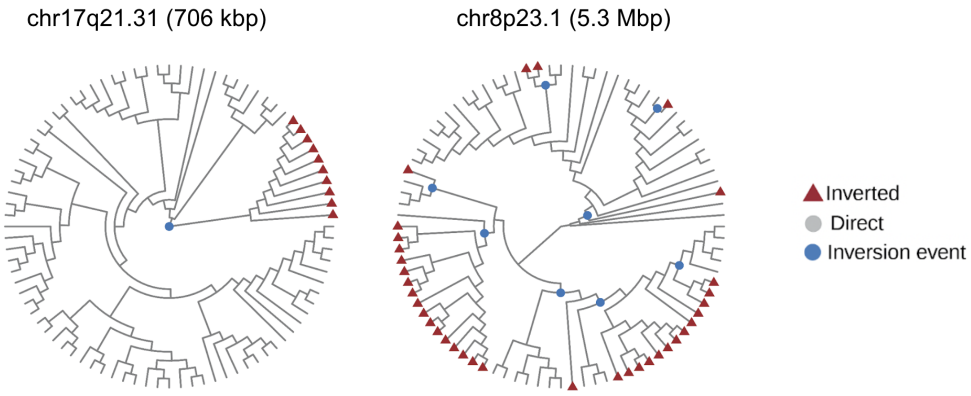

Figure 4.12: Cladograms of two loci harboring a non-recurrent (left) and a recurrent (right) inversion based on haplotype-resolved PacBio data and Strand-Seq derived inversion genotypes. Blue dots indicate putative inversion events. The figure and underlying data were created by P. Hsieh.

in a flanking SD, which was found both present and absent in reference and inverted alleles, violating a temporal order of these two events and thus suggesting at least one recurrence event (Fig. 4.13A). Although the deletion could represent a scar from a previous inversion event, it is equally possible to be a mere 'passenger' event. A nested inversion provided another case of visible signs for inversion recurrence on chrXq28, which displayed haplotypes carrying all four possible combinations of inversion genotypes (ref/ref, ref/inv, inv/ref, inv/inv), suggesting a minimum of one recurrence event in the region (Figure 4.13B).

4.6 Polymorphic inversions associate with morbid CNVs

Inversions in the great apes display enrichment of inversions that overlap recurrent large-scale copy number variants, with 17/36 (47%) of CNV loci displaying >50% reciprocal overlap with a mapped inversion [Porubsky et al., 2021].

This section describes novel forms of interplay between inversions and de-novo copy-number variations, which become apparent by utilizing the unprecedented resolution in highly complex inversion loci provided by high-confidence de-novo genome assemblies [Ebert et al., 2021]. Furthermore, novel mechanistic insight into the formation of two CNV-related diseases - 3q29 microdeletion and 15q13.3 deletion - is discussed, providing novel links between inversions and CNV-related diseases.

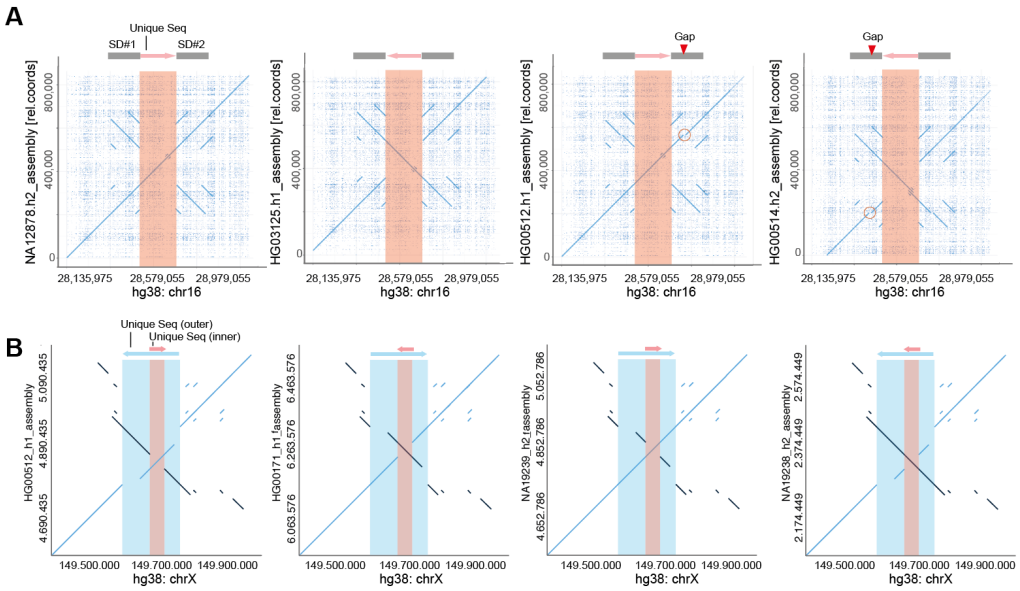

Figure 4.13: Dotplots showing evidence for inversion recurrence. A Series of dotplots of an inversion on chr16, surrounded by a complex SD pattern. Some samples carry a gap that can serve as a marker for sequence ancestry. Sequences with and without the gap are seen in reference and inverted orientation, suggesting the independent formation of the two inversions. **B** Dotplots of a nested inversion on chrX. The outer inverted region is highlighted in blue, the inner one in red. All four possible combinations of nested inversion (inv) states (ref/ref, ref/inv, inv/ref, inv/inv; with ref. for reference orientation) are observed across samples, suggesting at least one instance of inversion recurrence.

4.6.1 Inversions co-locate with known CNV hotspots

My colleague D. Porubsky has examined the overlap between (recurrent and nonrecurrent) inversions and known copy number variants from the decipher database [Bragin et al., 2014], revealing a strong enrichment of such overlaps compared to randomized loci (2-fold enrichment for single inversions, 5-fold for recurrent inversions). This enrichment included polymorphic inversion regions with >90% reciprocal overlap with morbid CNVs at genomic loci 2q13, 7q11.23 (Williams-Beuren Syndrome), 16p13.11, 15q13.3 and 15q11.2-15q12.

Pairs of SDs can predispose regions for pathogenic microduplications and microdeletions occurring via NAHR if oriented in the same direction or mediate inversions if oriented in inverse orientation. In the case of nested SD pairs, this yields a possible mechanistic connection between inversions and CNVs (Figure 4.14): while balanced inversions may not be deleterious themselves, they can provide an opportunity for other SD pairs to 'flip' their relative orientation, enhancing or reducing the risk for subsequent deleterious NAHR events.

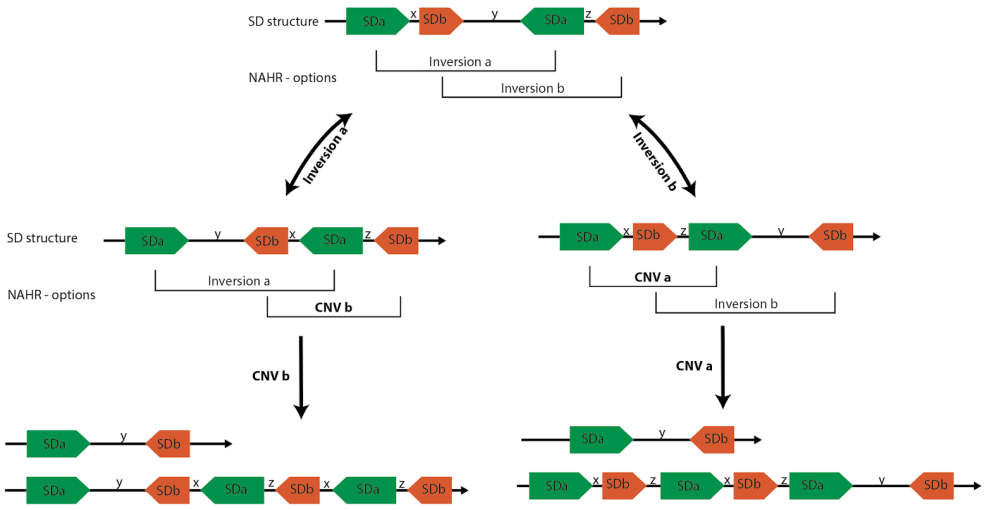

Figure 4.14: Conceptual drawing illustrating how balanced inversions may change the pre-mutational landscape to foster morbid CNV formation. The scheme illustrates a hypothetical locus with two overlapping pairs of identical SDs (designated SDa and SDb). In this example, starting from a pre-inverted locus (top row), an NAHR-mediated inversion, using the SD architecture would change the orientation of an SD pair (middle row) to now potentially allow subsequent morbid CNV formation by NAHR (bottom row).

4.6.2 Systematic identification of CNV-predisposing inversions

Primed by the observations described in chapter 4.6.1, I conducted a genome-wide analysis to identify inversions that affect the relative orientation of pairs of segmental duplications. Utilizing an annotated set of 7,672 SD pairs (>10 kbp and >90% sequence identity, partners on the same chromosome) obtained through the UCSC table browser [Karolchik et al., 2004], I identified 1,094 SD pairs which are expected to see a change of relative orientation as a consequence of to adjacent inversion events. In particular, 29 inversions were identified, which all lead almost exclusively to directly oriented SDs (termed 'potentially risk inducing' N=20) or indirectly oriented SDs ('potentially protective'). Collectively, these 29 inversions overlapped with 10 morbid CNVs.

4.6.3 Inversions display molecular links to CNVs in three genomic loci

INV-CNVs at genomic region 3q29

Utilizing haplotype resolved assemblies based on HiFi-Pacbio reads, I characterized the structure of an inversion at the upstream flank of the 3q29 microdeletion syndrome region associated with schizophrenia (Fig. 4.15). Dotplot alignments revealed that the inversion directly reorients a 21 kb SD which can explain the 3q29 deletion, and that, additionally, non-inverted haplotypes carry directly oriented duplications of another SD capable of triggering the mCNV via NAHR in >50% of cases. In contrast, such duplications are absent from the inverted haplotype ($p < 1.6*10\text{-}5$). This observation provides, for the first time, a mechanistic explanation for the interplay between this inversion and the 3q29 mCNV, suggesting a protective role of the inverted haplotype and explaining, in particular, the lack of inverted haplotypes (0/18) in carriers of the 3q29 mCNV in a recently published study [Yilmaz et al., 2021].

INV-CNVs at genomic region 15q13.3

An analogous study yielded novel insights into the SD structure surrounding a 1.5 Mbp recurrent inversion overlapping the 15q13.3 microdeletion region. The region has previously been implicated in evolutionary instability driven by highly identical copies of the *GOLGA8* core duplicon [Antonacci et al., 2014]. Among 5 distinct haplotype configurations identified (Fig. 4.16), the dotplot alignments indicated independent inversion polymorphisms of either copy of the CNPβ SD (denoted INV-β and INV-β'), which is presumed to provide the substrate for the 15q13.3 deletion via NAHR [Antonacci et al., 2014]. The risk for CNV is therefore likely dependent on the state of both inversions β and β': Inversion of either one, in isolation, leads to directly oriented copies of CNPβ and thus to a risk for microdeletion. Inversion of none, or both of the SDs, in turn, likely acts as a protective allele by yielding inversely oriented SDs. In line with this model, analysis of INV-β alone has not yielded a significant correlation with 15q13.3 morbid CNV formation [Antonacci et al., 2014].

INV-CNVs at genomic region 7q11.23

As a third example, I examined a long inversion spanning the Williams-Beuren-Syndrome critical region of 7q11.23 4.17. The region displayed a significant

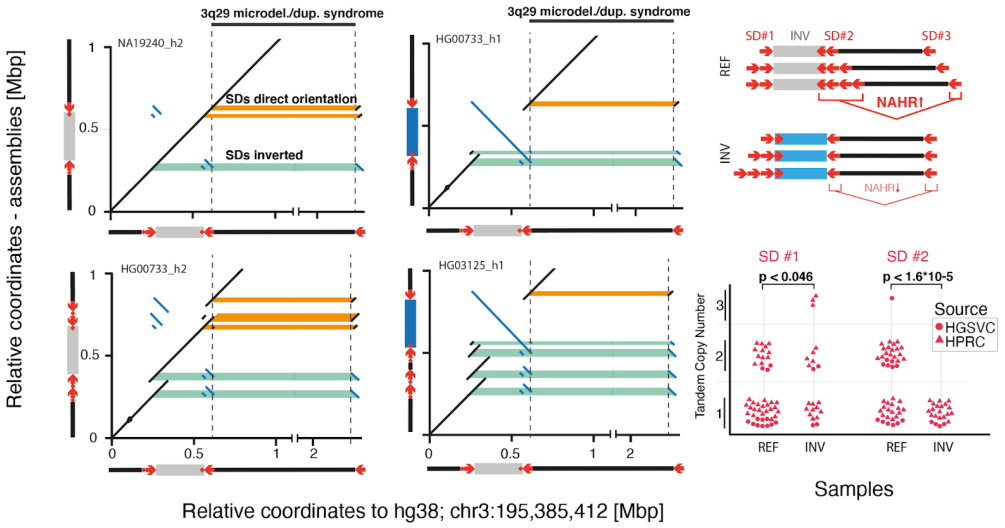

Figure 4.15: Annotated dotplots of four assembled haplotypes in the genomic locus containing a large inversion at the 3q29 microdeletion/microduplication critical region. SD pairs spanning the 3q29 microdeletion/microduplication and their relative orientation are highlighted in yellow (direct orientation) and green (inverse orientation). The relative orientation (direct/inverse) of an SD pair is flipped in inverted haplotypes since one SD is contained within the inversion (top left, bottom left). Additionally, tandem duplications of the inversion-mediating SDs (2nd row) are observed in >50% of haplotypes. Tandem duplications of SD 2, putatively posing a risk for morbid CNVs, are common in reference but absent in inverted haplotypes (bottom right).

variability of haplotypes, including a nested pair of polymorphic inversions (which we denote inversion α and β). Given the presence, absence, or orientation of SDs in the region, we propose that several haplotypes might be at higher risk (cases III and IV, Fig. 4.17) or lower risk (cases V, VI, Fig. 4.17) for subsequently developing of the WBS-associated copy-number variation.

4.7 Discussion

Chapter 4 has presented a comprehensive analysis of the most extensive set of polymorphic inversions in humans described to date, resulting in new computational methodologies and contributing majorly to the understanding of inversion polymorphisms.

Using a multi-technology approach, my colleagues and I identified a set of 398 inversion loci found across 86 diverse human haplotypes. Using de-novo assembled genomes, I could also determine the exact breakpoints of 183 inversions, allowing

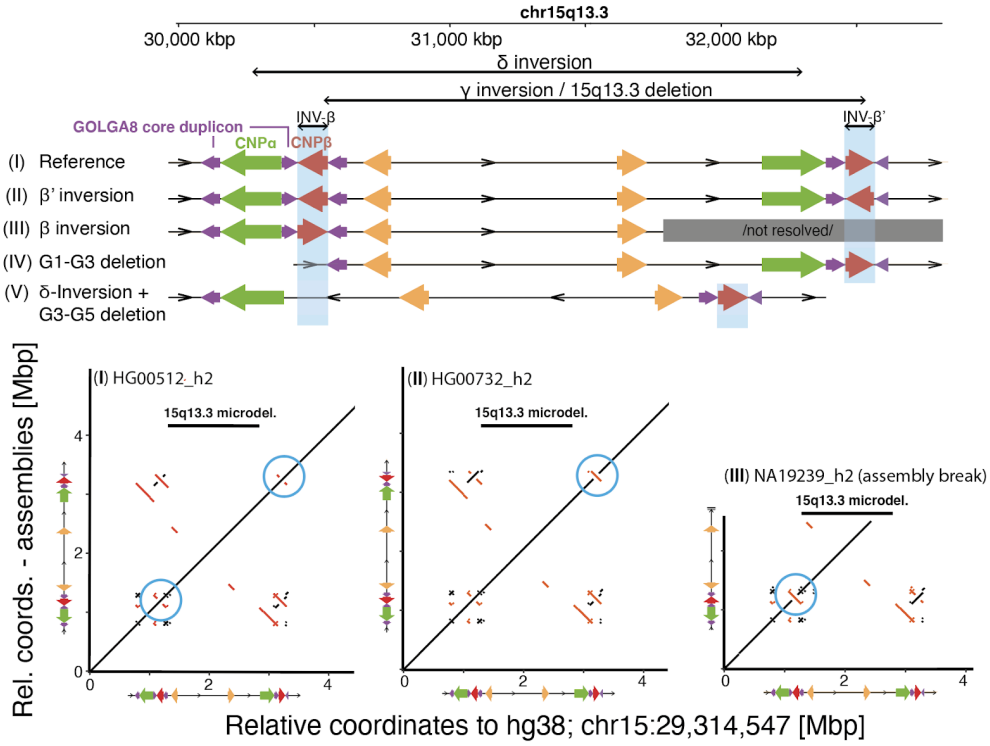

Figure 4.16: Schematic view of the repeat structure and structural variations found in the 15q13.3 region in different haplotypes. Three phased assembly-based dot plots illustrate structural haplotypes containing INV-β and INV-β'. Both inversions are mediated by GOLGA8-duplicons (purple arrows) and contain a copy of the 210 kbp CNPβ-repeat each (red arrows), which we predict serves as a template for morbid CNV or recurrent inversion formation, depending on the combined inversion status of INV-β and INV-β'. We find additional haplotypes (IV, V) containing deletions, which are putatively protective against both inversion recurrence and morbid CNV formation.

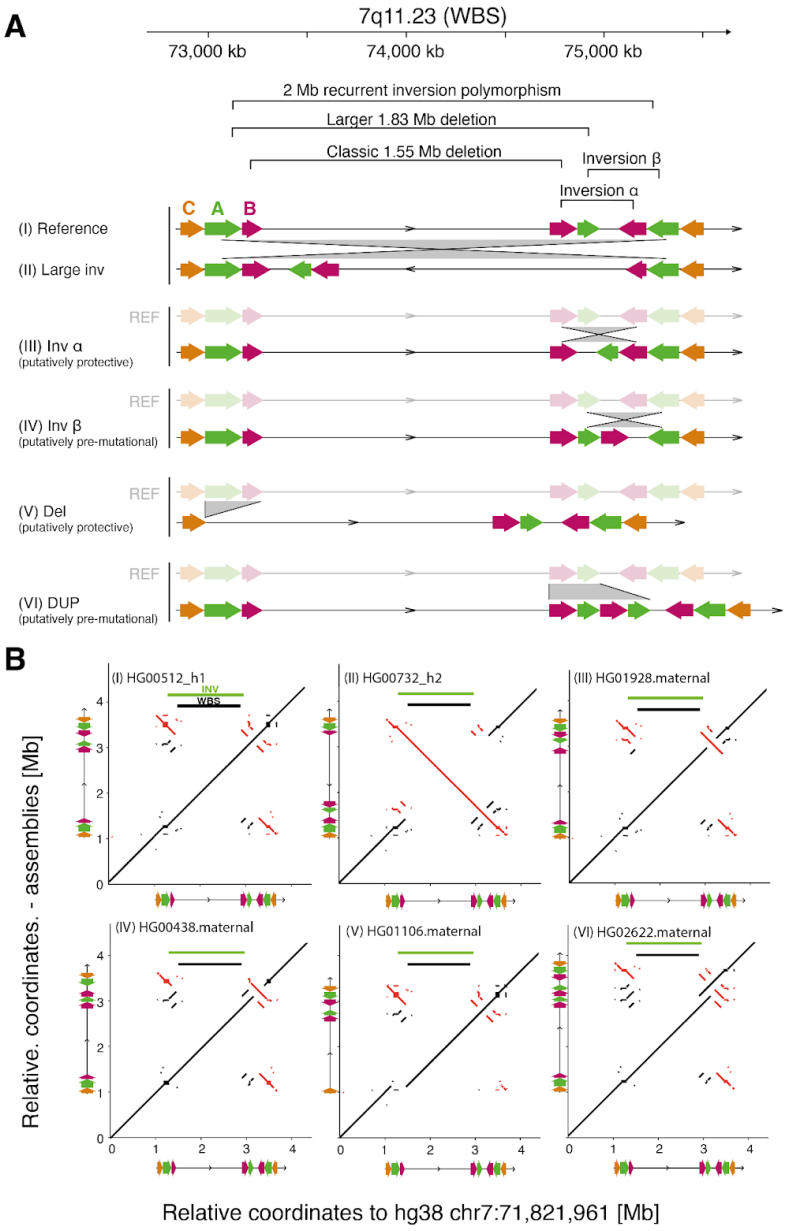

Figure 4.17: Detailed view of haplotypes in the 7q11.23 WBS region. A Overview of genomic rearrangements described in the region and organization of SDs in six samples derived from phased whole-genome assemblies. The nested inversions α and β affect the SD architecture of the region, each producing potentially protective or pre-mutation structural haplotypes with respect to subsequent inversions and CNVs (haplotypes III and IV). We find other haplotypes with a likely protective deletion (haplotype V) and a potentially CNV-enabling duplication (haplotype VI). **B** Dotplots of the region in the samples described in (A). Inferred loci and SD structures are indicated along the axes.

me to characterize specific features of at least three diverse classes of inversions. The majority of long events (> 100 kbp) were flanked by segmental duplications and are thus likely to have arisen through non-allelic homologous recombination (NAHR). In contrast, most short (<10 kbp) inversions were associated with additional insertions or deletions, indicative of other mechanisms such as MMBIR or FoSTES. A significant limitation of the callset is the relatively small number of samples considered (86 haplotypes), which leads to a potentially large number of rare inversions in the human genomes which may be absent from our callset. Another consideration revolves around the incomplete breakpoint resolution of more than half of the inversions. Such lack of resolution is due to technological limitations that are difficult to overcome and may sometimes obscure complexity observed near the breakpoints. With the pace of improvements in long-read sequencing and genome assembly, it can be expected that such gaps will be closed more routinely in future studies, revealing the associated inversions in full detail in subsequent studies.

My colleagues and I demonstrated a larger-than-expected extent of inversion toggling in the human genome, with 40 inversions, covering 0.6% of the human genome, flipping their orientation repeatedly. Furthermore, 6/40 recurrent inversions overlap with recurrent inversions identified in great apes [Porubsky et al., 2020], highlighting that inversion recurrence is not limited to human genomes and likely plays a role in sequence evolution that has been underestimated so far. This finding is also relevant for future population genetics studies, which will have to consider the concept of inversion recurrence. Again, due to the low sample size, the survey is likely still underestimating the number of recurrent inversions, as inversions can only be identified as 'recurrent' after they have been observed in at least two different genomic contexts. As an alternative to population-based identification of recurrence, it would be desirable to determine the exact inversion breakpoint positions inside SDs. Demonstrating such shifts in breakpoint positions would be a more direct way of proving recurrence. However, even when perfectly assembled sequence assemblies were available, this proved to be an unexpectedly difficult task due to interspersed gene conversion events which appear to be a side product of inversion formation. My colleagues and I are interested in developing additional analysis tools to understand better the processes occurring at the breakpoints during *NAHR*-mediated inversion formation. However, we have still observed cases of alternative breakpoint usage in cases where different SDs were involved, leading to the 'same' genomic region being inverted by 'different' events.

The number of inversions disrupting genes is relatively low, with all breakpoint-

resolved inversions disrupting a mere one and two genes, respectively, and only three potential fusion genes were observed. This notion highlights that the overall impact of inversions on gene bodies is likely lower than that of other SVs like insertions or duplications. Likewise, the number of inversion eQTLs remains limited in this project. However, this analysis proved to be underpowered due to the low sample number and the fact that only expression data from LCL cell lines were considered. Furthermore, all data included in this study has been sampled from healthy individuals. While this dataset is well suited for studying natural variation found in healthy members of the human population, the callset may thus be biased against SVs conveying strong negative phenotypes.

Lastly, our project brought novel insights into the relationship of (long) inversions with disease-causing copy-number variants (CNVs), a long-standing association whose molecular underpinnings are unclear in many cases [Antonacci et al., 2009a, Antonacci et al., 2014, Koolen et al., 2016, Maggiolini et al., 2020b]. Our inversion callset confirms the co-segregation of disease-causing CNVs and inversions, which we find to cluster in hotspots around these regions. Inversions and long CNVs have been described before to be linked at the molecular level, as both are typically mediated by segmental duplications (SDs) via non-allelic homologous recombination (NAHR). Indeed, our analysis has provided evidence for such association in three individual CNV-associated regions – 3q29, 15q13.3, and 7q11.23 –, in which inversions may act as facilitators or protectors for CNV formation by affecting the local landscape of SDs. From a broader perspective, inversions may even be viewed as "switches" that coordinate the local SD landscape and regulate the risk of subsequent inversion or CNV formation. We speculate that such machinery may play a role in forming many CNVs, and future studies of CNVs will have to take the role of adjacent inversions more directly into account.

In summary, this chapter has presented evidence that inversions are much more frequent in human populations than previously thought, with at least 0.6% of the human genome subject to inversion recurrence. Furthermore, many of these inversions are frequently associated with disease-causing CNVs. Our analysis suggests that inversions facilitate – or prevent – such events by restructuring the local SD landscapes. More studies will be required to gain molecular insights into specific classes of difficult-to-study inversions. Potential foci for subsequent studies include the identification of rare and disease-causing inversions, providing full breakpoint resolution even of very long and complex inversions, and developing approaches to determine exact breakpoint positions within segmental duplications.

5

Nested repeats promote clusters of complex and highly dynamic SVs

This chapter covers the development and application of a novel tool, NAHRwhals, intended to resolve complex series of genomic rearrangements. Most of the work presented here provides the core of a manuscript I plan to finish and submit for publication shortly. The section on chromosome Y represents a contribution to a collaborative project led by P. Hallast and P. Ebert, for which submission is in preparation. Besides my mentor J. Korbel, the project has been supported massively by F. Sedlazeck [Baylor College, US], who kindly agreed to co-supervise the project and whom I want to thank warmly for his skillful and personal engagement.

Contents

5.1 Introduction: Sequential SVs promote complex rearrangements

The previous chapter has revealed a tight relationship between segmental duplications, inversions, and copy-number variations, with direct influences observed between each. In particular, the repeat- and SV-architecture of three genomic loci, *3q29*, *15q13.3* and *7q11.23*, has suggested that different haplotypes may have an increased or decreased propensity to develop disease-causing copy-number variations. Likely, this effect is driven by variations in their SD composition caused by secondary SVs such as inversions. One of the clearest examples for this concept, discussed only briefly in section 4.6.3, is provided by a complex haplotype in the 15q13.3 locus, which was likely created by two overlapping, temporally distinct SV events (Fig. 5.1). As the following chapter will illustrate, such 'multi-stage' SVs have only been described anecdotally to date, despite likely occurring in conjunction with large, disease-causing CNVs. As an initial step to formalizing the study of such events, I will refer to them as *Serial Structural Variation* (or *sSV* for short) throughout this chapter.

Results presented in [Porubsky et al., 2022b] suggest that > 1000 SD pairs may be subject to relative orientation changes due to inversion events. In particular, 29 inversions affect almost exclusively directly or indirectly oriented SD pairs. Given that more than one-third of these (10/29; 34%) overlap disease-causing morbid CNVs, it might be expected that close inspection of such events will reveal more pre-mutative states, and consequently, more (morbid) CNVs might eventually become ascribed to *sSVs*.

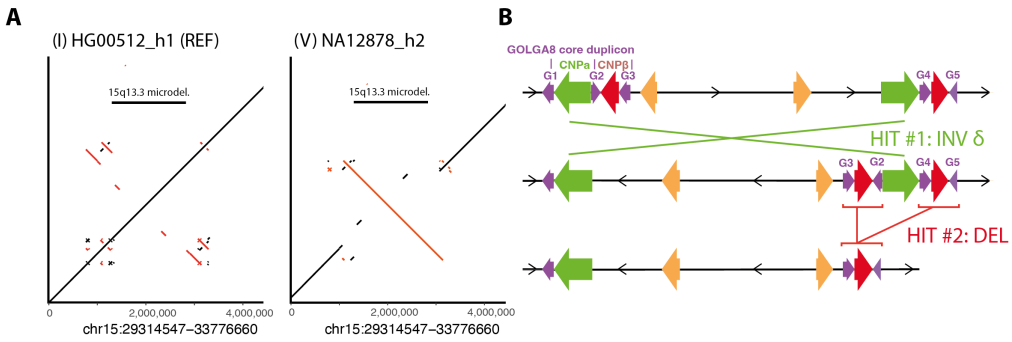

Figure 5.1: Detailed view of an SV event in the 15q13.3 inversion/microdeletion region. A dot plot view of the region 15q13.3 (discussed in Fig. 4.16) in reference state (left) and the inverted state with a deletion (right). **B** A two-hit model that may explain the formation of haplotype (V) as a series of (1) NAHR-mediated inversion δ, followed by (2) *NAHR*-mediated deletion using the directly oriented copies of CNPβ.

sSV-like patterns have also been described in other studies. A remarkably diverse example is the *TCAF1/2* locus. This region has undergone multiple rounds of recurrent SD-driven structural mutations over ∼1.7 million years [Hsieh et al., 2021]. Starting from a single non-duplicated ancestral haplotype, initial duplication and rapid structural change event has led to the emergence of at least five haplogroups represented in modern humans, which diverge from each other through a series of *NAHR*-mediated mutation. A similar process was likely at play in the *TBC1D3* gene expansion, which also displays an abundance of SDs and remarkable human diversity achieved through repetitive rearrangements [Vollger et al., 2022]. Another survey of complex structural variants based on short-read WGS of 1,324 undiagnosed rare disease patients revealed four pathogenic complex SVs. These include complex resolved events described, e.g., as 'duplication-inversion-inversion-deletion' and 'deletion-inversion-duplication'[Sanchis-Juan et al., 2018]. For three such events, hypothetical intermediate states were proposed to explain the transition from the reference state to a derivative, resembling our definition of an *sSV* (though not necessarily mediated by SDs). Moreover, the recently finished first human pangenome has allowed new insights into structurally complex regions, such as the *RHD*, *HLA-A*, *C4*, *CYP2D6* and *LPA* loci, and several loci display similar SV patterns [Wang et al., 2020].

In summary, dozens of human genomic loci have been documented to undergo serial rearrangements in selected populations or individuals. However, a systematic survey of *sSVs* is still lacking, and a significant fraction of SD-mediated complexity has likely remained undetected to date while slowly becoming accessible through the more routine use of long or ultra-long read technology [Ebert et al., 2021].

5.1.1 Computational identification of *Serial SVs*

Apart from technological obstacles, the identification of *sSVs* is also impeded by a lack of conceptual frameworks and computational tools. Even when a serially rearranged locus is fully sequence-resolved, calling of such an event is non-trivial, as the computational objective differs from 'classical' SV calling (Fig. 5.2). For example, a complex event like 'Del-Inv-Del' can be identified routinely using traditional SV callers like *Sniffles* [Sedlazeck et al., 2018], which would correctly report 1) a deleted, 2) an inverted, and 3) another deleted segment (Fig. 5.2B). However, while this call is entirely accurate in a descriptive sense, the call may

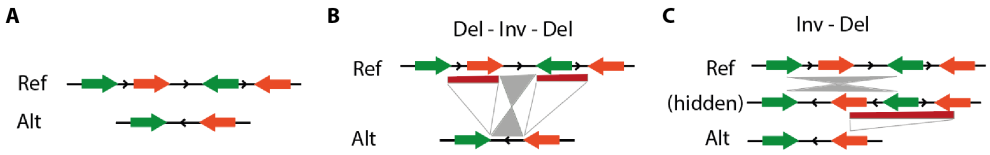

Figure 5.2: Illustration of problems associated with identifying serial SVs. **A** An example locus harboring two SD pairs in the reference state and an alternative allele missing one copy of each SD. **B** In conventional, descriptive SV calling, the explanation 'Del + Inv + Del' is an optimal result, which fully describes the observed SV. **C** An alternative, mechanistic description of the SV considers the likely series of overlapping events in the region: the alternative state can also be explained by an SD-mediated inversion, followed by an SD-mediated deletion.

also be misleading from a mechanistic point of view, as the actual series of events may have been "1) inversion; 2) deletion". This alternative interpretation is not reflected in the descriptive solution (Fig. 5.2 C).

However, complex rearrangements can also result from single complex SVs, e.g., formed by FosTeS or MMBIR [Carvalho and Lupski, 2016], and the distinction to *sSVs* is not always trivial. Owing to the lack of a mechanistic SV calling framework, *sSVs* described previously had to be inferred in a manual way [Porubsky et al., 2022b, Sanchis-Juan et al., 2018, Wang et al., 2020], often preceded by extensive haplotype analysis [Vollger et al., 2022, Hsieh et al., 2021]. In these manual operations, two concepts have proven helpful to distinguish *sSVs* from their complex counterparts and to reconstruct chains of simple SVs:

1. **Intermediate states** may be observed when surveying more than two haplotypes, making certain SV series more plausible. In the example above, an observation of the 'hidden' haplotype, carrying a simple inversion, would provide the missing link between the reference and alternative state and make this explanation favorable to the complex alternative.

2. Pairs of **segmental duplications** serve as breakpoints for *NAHR*-mediated events and can 'suggest' intermediate states even if they are not directly observed. In the example above, the SD pairs can explain the 'hidden' inversion state and subsequent inversion, while the alternative, 'Del-Inv-Del' is also spanning SDs, but not explicitly predicting SVs between any pairs.

The study presented in this chapter revolves mainly around large (>10 kbp) *sSVs*, which have the potential to associate with morbid CNVs. Since (1) large SVs are most frequently formed by *NAHR* [Porubsky et al., 2022b, Ebert et al., 2021], and (2) SDs are a crucial element for identifying *sSV* chains, the remainder

of this chapter focussess on *NAHR*-mediated *sSVs*. The terms *NAHR-mediated sSV* and *sSV* are thus used interchangeably in this chapter.

A part of the work presented here represents a contribution to a project on variation in the Y chromosome. This chromosome is particularly susceptible to NAHR for at least two reasons: First, chrY exhibits an unusually high content of segmental duplications [Vollger et al., 2022], echoing the chromosome's history of progressive degradation ([Charlesworth and Charlesworth, 2000] and section 1.2.2). Second, the lack of a homologous partner likely promotes intrachromosomal ectopic recombination in non-recombining regions [Cáceres et al., 2007]. Accordingly, chrY exhibits an enrichment of recurrent NAHR-driven inversions [Porubsky et al., 2022a]. The high content of segmental duplications and other repeats makes chrY the most difficult chromosome to assemble, and the first complete de-novo assembly of this chromosome was only released recently [Nurk et al., 2022]. Accordingly, a significant fraction of genomic variation on this chromosome has likely been missed to date. In response to this notion, a project by members of the *HGSV* Consortium is currently attempting to generate more than 40 high-quality assemblies of chrY to investigate the variation of this chromosome comprehensively.

This chapter describes a systematic analysis of *serial structural variations* in the human genome. Initially, the development of a new 'mechanistic' *sSV* caller termed *NAHRwhals* is highlighted, which performs detection and genotyping of *sSV* loci from phase-resolved genome assemblies. Subsequently, the results of applying this tool to a set of 56 assembled haplotypes are presented, where n=20 *sSV* loci were identified and analyzed. After discussing new likely associations of several *sSVs* with morbid CNVs, the last section of this chapter describes the application of *NAHRwhals* to a set of new assemblies of chromosome Y.

5.2 Key steps of the NAHRwhals sSV detection routine

The workflow of *NAHRwhals* (short for '***NAHR**-directed **W**orkflow for catc**H**ing seri**AL** **S**tructural variations*') consists mainly of four steps, which will be separately explained in the following sections. The steps are briefly indicated below (see also Fig. 5.3).

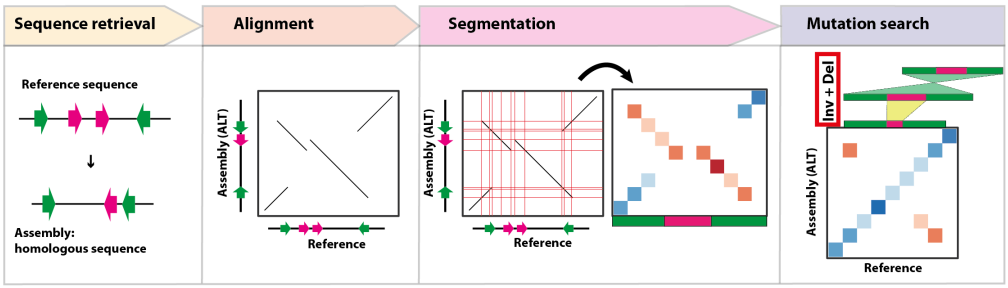

Figure 5.3: Overview over the *NAHRwhals sSV* detection method. Flowchart showing the key steps of the algorithm. Given a reference region of interest, the homologous region is first extracted from the assembly. Pairwise alignments between ref and alt are created and segmented into a condensed, two-dimensional representation. Based on this condensed dot plot, an exhaustive search is employed to examine possible chains of NAHR-mediated rearrangements explaining the structural differences.

1. **Sequence retrieval**: Based on a reference genome, region-of-interest coordinates on the reference genome, and a custom haplotype assembly, the most likely homologous region on the assembly is identified using *minimap2*.

2. **Pairwise alignments**: After isolating a locus on the reference (REF) and its homologous counterpart (ALT), a custom mapping pipeline again based on *minimap2* is invoked a second time to produce an accurate local pairwise alignment.

3. **Alignment segmentation** A custom segmentation algorithm then enables compression of the original alignment, producing a 'compressed dot plot of pre-specified size (default: 50x50 squares) or compression factor (default: 1-10 kbp)

4. **Exhaustive mutation search** On the basis of a condensed dot plot, the mutation space is explored in a depth-first search approach to identify NAHR-based SV chains capable of transforming the reference sequence into a structure equivalent to ALT.

5.2.1 Sequence retrieval

Minimap2 [Li, 2018] is initially invoked to custom-liftover locus coordinates from a reference (REF, typically hg38 or chm13-T2T) to an 'alternative' (ALT) assembly (typically a de-novo assembled genome), identifying and extracting the

most similar homologous region in the assembly. This step can be performed with a pre-computed alignment file (conceptually similar to .chain files) to save computation time. This step can be skipped if extracted sequences are already available or simulated sequences are used.

5.2.2 Pairwise alignments

A custom pipeline was built around the *minimap2* aligner to obtain high-fidelity pairwise alignments even in highly repetitive genomic regions. Before aligning, the query sequence is split into chunks of 1 kbp (if length(query) $<$ 50 kbp), 10 kbp (if length(query) is between 50 kbp and 5 Mbp), or 100 kbp (if length(query) $>$ 5 Mbp). The 'chunks' are then aligned to the target sequence separately, reducing the need for read-splitting, which is known to be error-prone in minimap2. In a post-processing step, alignment pairs are concatenated whenever the endpoint of one alignment falls close to the start point of another (base pair distance cutoff: 5% of the chunk length). If multiple alignments 'compete' for the same partner (e.g., two alignments ending close to the beginning of another), only the longest 'competitor' gets selected for merging.

5.2.3 Alignment segmentation

Pairwise alignments are retrieved from minimap2 in .paf format, representing pairwise alignments as two-dimensional vectors from start- (query-start/target-start) to end (query-end/target-end) coordinates. In order to prepare subsequent compression steps, alignments are pre-processed in multiple ways: First, alignments are filtered by a minimum length threshold (*min_aln_len*), removing very short alignments. Second, alignment breakpoint coordinates are rounded in the x and y directions to the closest multiple of a rounding parameter (*rounding_parameter*). Finally, alignment vectors are shortened along the x or y axis in the rare case that they do not have a slope of exactly 1 or -1 until they do so.

Following noise-reduction, borders, or 'gridlines', separating unique sequence blocks, are inferred in an iterative way which can be interpreted visually (Fig. 5.4B). In the first iteration, horizontal and vertical gridlines are drawn starting from each start-and endpoint of any alignment. In every subsequent step, overlaps

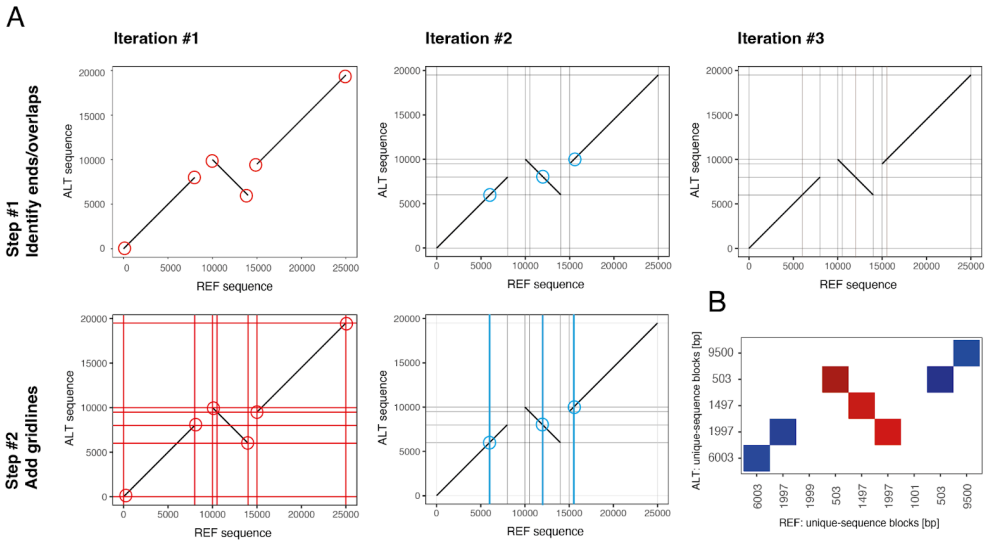

Figure 5.4: Visual representation of the iterative dot plot segmentation algorithm. **A** Starting from a pairwise alignment, all start-and endpoints are identified, and the x- and y-values are noted as the first set of horizontal and vertical vectors ('gridlines') separating unique alignments. In each subsequent step, novel overlaps between existing gridlines and pairwise alignments are identified, and new gridlines are inserted horizontally and vertically at the intersections. Once the grid has converged, each field is, by design, traversed by zero or one alignment vector diagonally, intersecting with exactly two opposite corners. Grids that do not converge after ten iterations are rejected, and the dot plot pre-processing is repeated with another parameter set until a converging representation is found. **B** A "condensed dot plot" is derived, where each field of the new dot plot represents a sector of the grid. The value represents the length of the traversing alignment, and the sign of the value corresponds to the direction of the alignment (blue: positive values: direct orientation; red: negative values: inverse orientation).

between existing gridlines and alignments are determined, with the points of overlaps serving as a new source for spawning a new gridline in a perpendicular direction. This process is repeated until no new gridlines are spawned. Once the 'grid' is established, the length and directionality of each alignment passing a cell are calculated and transferred into a simplified matrix (Fig. 5.4B)

min_aln_len and *rounding_parameter* represent free parameters that influence the level of detail retained in the pairwise alignment and consequently influence the size of the condensed dot plot. These parameters are chosen using an explorative strategy to minimize the size distance between the observed dot plot and a user-specified desired size (default: 50 * 50 squares) while ensuring *min_aln_len* > *rounding_parameter*. In practice, a first condensed dot plot is created by starting from initially random parameters. Then, based on the dimensions of the resulting matrix, the subsequent parameter pair is chosen to increase or decrease gridsize compared to the previous iteration. This process is repeated 20 times, and the best-performing parameter pair is retained for the final compression.

5.2.4 Exhaustive mutation search

The reduction of potentially multi-Mb alignments to a matrix of much smaller dimensions allows NAHRWhals to employ an exhaustive search strategy to identify chains of SVs capable of transforming the REF-configuration into ALT (Fig. 5.5). Condensed dot plots enable the immediate detection of duplicative sequences (i.e., rows or columns with >1 colored square) and associated *NAHR*-mediated SVs (del/dup between similarly colored squares; inv between opposites). Such SVs are systematically explored in a recursive depth-first tree-search algorithm, where SVs are also simulated in dotplot space. Mutated matrices are scored using a customized Needleman-Wunsch algorithm [Needleman and Wunsch, 1970] in which the program treats the condensed matrix like a regular pairwise sequence alignment. SV-chains producing a pairwise alignment of >98% sequence identity are considered successful. Implausible chains are abandoned if the alignment score is lower than 70% of the original alignment after two subsequent mutations to reduce computational load. Still, due to the exponential growth of mutations to simulate, the search depth is limited to 3 SVs.

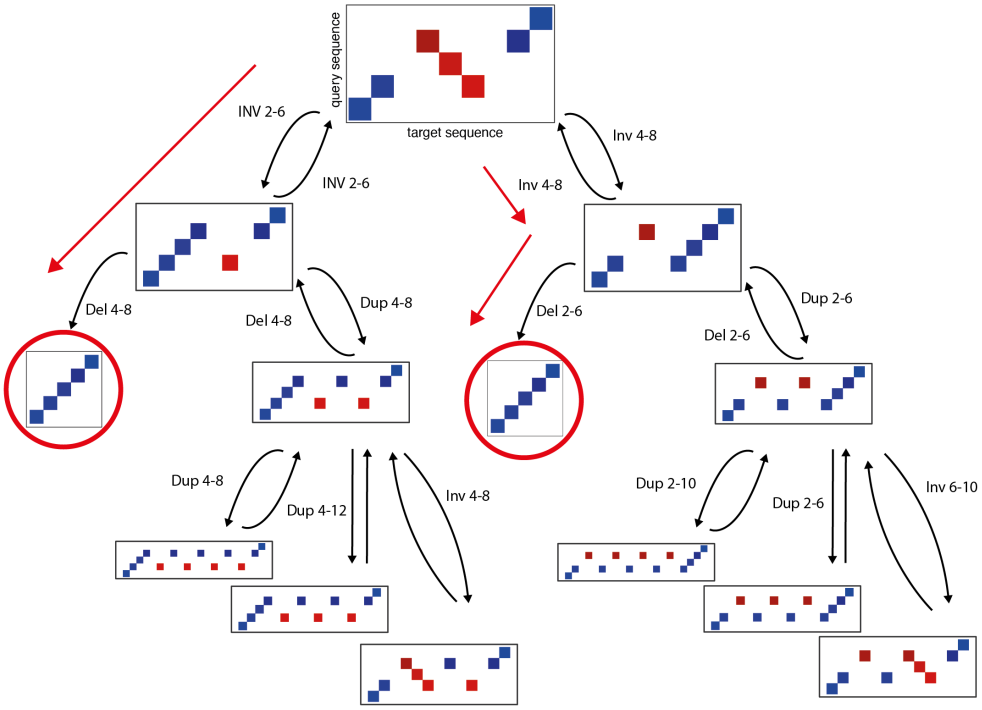

Figure 5.5: A mutation search tree of depth 3 for a simple condensed dot plot. In segmented space, any pair of segments in the same row corresponds to a repeat pair. The search algorithm identifies such repeat pairs and derives possible mutations. Repeat pairs of the same color prime for CNVs, differently colored for inversions. Mutations are always executed on the target sequence (x-axis, typically Reference genome) until the reference resembles the query (red circles).

5.3 Benchmark on simulated and real data

NAHRWhals includes a framework to simulate and mutate SD-containing sequences. The performance of the algorithm was tested on simulated reads as follows. First, 50 genomic sequences with two pairs of non-overlapping SDs of randomized length (100 bp - 10.000 bp), position, orientation, and sequence similarity (90-99%) were initially created. Subsequently, a set of sequence derivatives was created, which contained all NAHR-concordant combinations of INV, DEL, and DUP up to depth 2. Finally, these mutated sequences and their unmutated 'ancestors' were given as input to *NAHRwhals* for SV calling, and results were compared with the known background of sequences. As expected, a positive correlation between genotyping accuracy and the length and similarity of repeats emerged, with near-perfect accuracy for repeats larger than 10 kbp in this simulated setting (Figure 5.6).

Next, NAHRwhals was applied to ten inversion loci (taken from [Porubsky et al., 2022a]) to assess its performance, demonstrating accurate representations

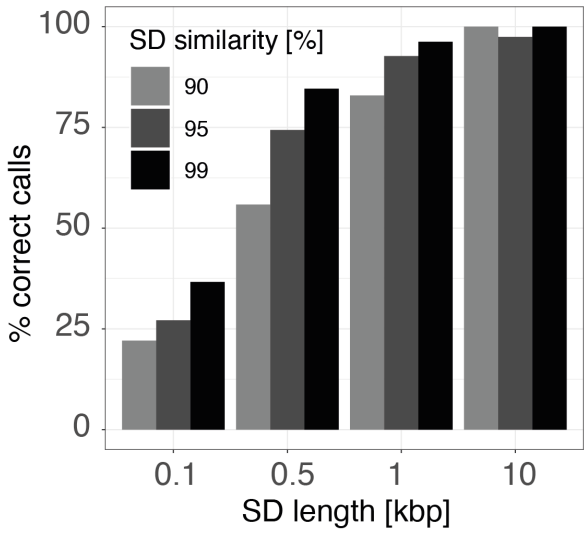

Figure 5.6: Results of simulation experiments. 50 genomic sequences with two pairs of non-overlapping, position- and orientation-randomized SDs were created and subjected to mutations of depth one or two. *NAHRwhals* subsequently was invoked to reconstruct the mutation chains, and calling performance was assessed. In the experimental setting, SD length and similarity both influence the prediction performance. SDs of length 10 kbp were sufficient for SV reconstruction in >95% of cases.

and correct SV genotypes for all ten loci (three of which are depicted in Fig. 5.7).

5.4 Identification of abundant *sSV* patterns in humans.

In chapter 4, my colleagues and I have described an extensive list of inversion polymorphisms in the human genome. I expected that a fraction of these inversions might be associated with additional *NAHR*-mediated events, similar to what my colleagues and I described anecdotally in three loci [Porubsky et al., 2022b]. To test this, I defined foci of interest by merging 398 polymorphic human inversions on hg38 with overlapping SD pairs obtained through the UCSC Table Browser [Karolchik et al., 2004]. Loci were then expanded by 25% of their length towards each direction, yielding a set of 213 inversion-associated, repeat-rich regions between 20 kbp and 35 Mbp in size. Based on the hg38 assembly as a reference, I invoked *NAHRwhals* to call SVs in these loci across 56 assembled human haplotypes, the T2T-CHM13 reference genome [Nurk et al., 2022] and five great ape genomes (chimpanzee, bonobo, gorilla, orangutan, macaque obtained from [Vollger et al., 2022]). When strict filtering criteria of 98% sequence identity

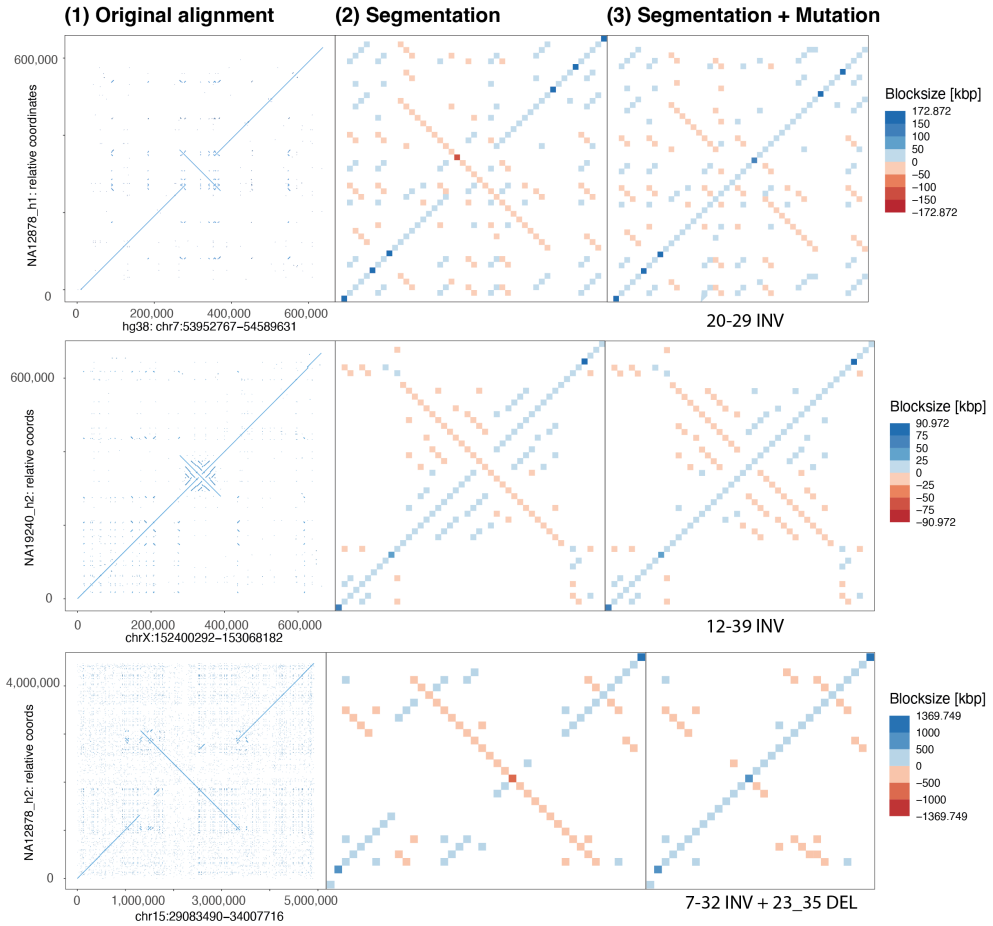

Figure 5.7: Alignment, segmentation, and SV-calling for three out of ten test loci. Pairwise alignments (left), condensed dot plots (middle), and mutation-resolved condensed dot plots for three example loci representing simple repeat-driven inversions. Characteristically for the segmentation, long stretches of unique sequence are condensed by the segmentation algorithm, while repeat-rich regions are resolved in finer detail.

are applied, *NAHRwhals* identifies a set of n=20 high-confidence *sSV* loci which were retained for further analysis (Fig. 5.8A).

The identified *sSVs* span a size range of 45 kbp to 3.5 Mbp (median: 515.5 kbp) and show high variability regarding the classes and depth of their predicted mutation chains. Comparisons with other datasets revealed that 6/20 (30%) of *sSVs* overlap with recurrent inversions [Porubsky et al., 2022b], corresponding to a statistically insignificant 1.59-fold enrichment of recurrent inversions among *sSV* sites (one-sided fisher's test N.S., p=0.1471). Further 6/20 *sSVs* overlap with members of core duplicon gene families such as *GOLGA* and *NPIP* which map to rapidly expanding SD regions [Johnson et al., 2006]. Finally, 7/20 *sSVs* overlap a set of 72 morbid CNVs (collected from [Bragin et al., 2014, Coe et al., 2014, Cooper et al., 2011]), corresponding to a 2.47-fold enrichment when compared to all 213 considered regions, where 31/213 matched mCNVs (p=0.0271, one-sided fisher's test).

On the level of predicted mutations, the most frequent haplotype states were 'reference' (n=238), 'inversion' (n=181), 'inversion+deletion' (n=45), and 'deletion' (n=43). 10/20 *sSVs* harbor at least one depth-three mutation, with a total of n=33 such events, the most common prediction being 'inversion+duplication+inversion' (n=17, 51% of depth-three-sSVs). One example of a relatively simple *sSV* is the 1p11.2-1p12 region (Fig. 5.8 B,C). I devised a new visualization based on the conceptual basis of 'sankey' plots to display inferred chains of mutations (Fig. 5.8 B,D). The 1p11.2-1p12 region contains two pairs of nested, inversely oriented SD pairs. The hg38-like state is carried by five haplotypes, a simple inversion of the inner pair by 17 haplotypes, and another 18 haplotypes harbor a simple inversion followed by a deletion along a then-directly oriented SD pair. The locus 7q35 (486 kbp) displayed even more complex rearrangements, in which different haplotypes follow widely branching rearrangement 'paths' (Fig. 5.8 D,E). The locus is outstanding in its variability, with 15 instances of 6 different depth-three events.

5.4.1 *sSVs* likely influence the risk of CNVs in disease-relevant regions

Given that 7/20 *sSVs* overlap with disease-relevant CNV regions, I suspected that some of the *sSVs* might also be causally linked to disease-relevant regions. Indeed, close inspection of the dot plot views of all 7 CNV-associated *sSV* regions

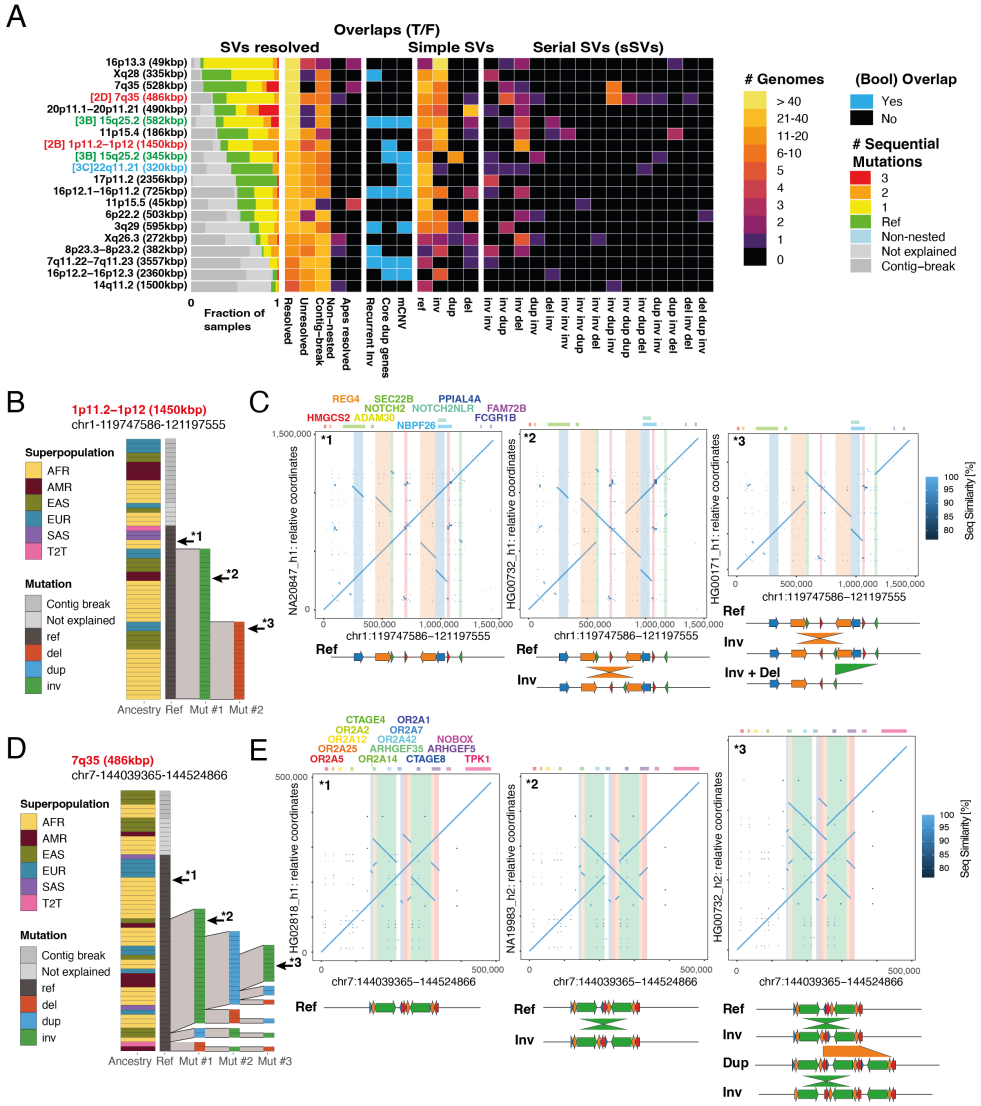

Figure 5.8: Inversion regions identified as sSVs. **A** Overview over the full callset of 20 inversion-containing loci in which *sSVs* were discovered in at least one sample. The diagram shows the prediction performance in humans and apes ('SVs resolved'), the presence of recurrent inversions, core duplicon-mapping genes, and morbid CNV regions in the genomic region, as well as genotypes for each locus. **B** Three distinct sequence configurations observed in the 1p11.2-1p12 sSV. 18/38 samples harbor a deletion preceded by an inversion compared to hg38. **C** Dot plots and SD schematics illustrating examples of all three configurations. **D** A 486 kbp region on 7q35 showing complex patterns of nested SVs leading to extreme diversity in the region explicable by NAHR. **E** Dot plot and SD schematics of three examples of various complexity.

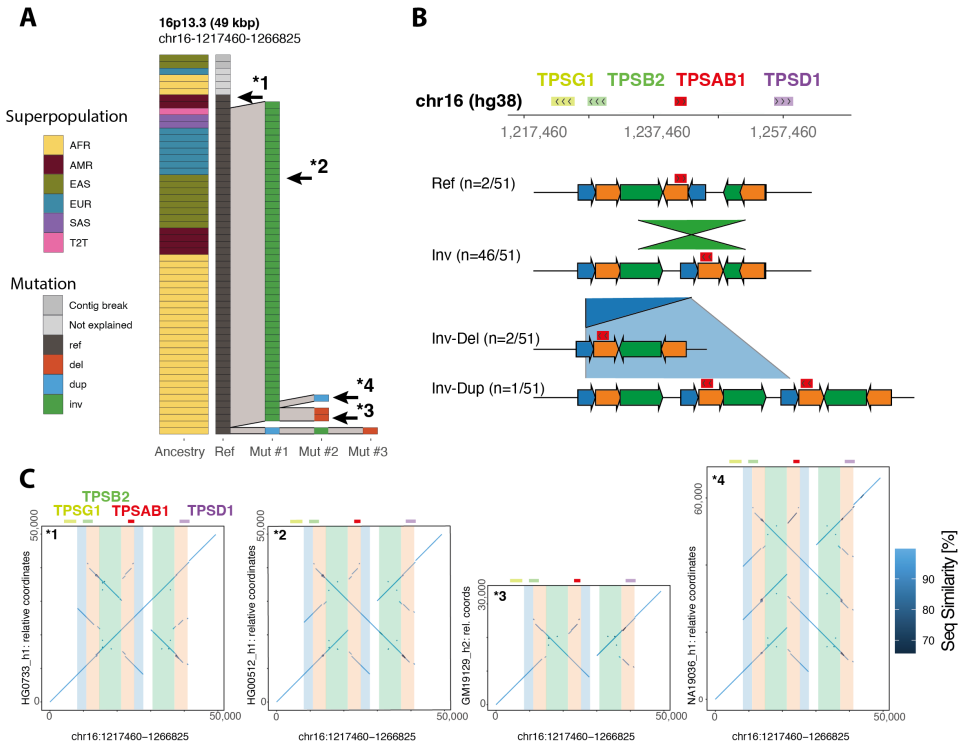

Figure 5.9: Four haplotype variations found in the human tryptase locus. **A** Mutation predictions in all haplotypes identified, with hg38 considered the 'reference' state. Four example haplotypes, indicated by black arrows, are shown in panel C. **B** View of the segmental duplications and the associated serial SVs that lead to variable copy numbers of the *TPSAB1* gene. **C** Dot plot views of examples of all four haplotype structures which were identified.

across all assembled haplotypes revealed at least four complex rearrangements that I predict to be associated with subsequent risk for copy-number variations (Figs 5.9, 5.10, 5.11).

For one, *NAHRwhals* identified two *sSV* events leading to deletion and duplication, respectively, of inverted haplotypes of a 16 kbp segment containing the *TPSAB1* gene (Fig. 5.9 A,B,C). One haplotype, being predicted dup-inv-del, does not show an overall copy number aberration and is structurally similar to an inverted haplotype (see Discussion). Copy-number variants in the tryptase locus have been associated with *Alpha Tryptasemia*, a non-lethal hereditary disease affecting 4-6% of the population [Lyons et al., 2016]. The genetic underpinnings of this disease are still poorly understood. It has been noted recently that classical genotyping approaches such as droplet digital PCR–based genotyping may be unsuited to capture the variation of this region due to the presence of inverted repeats [Lyons, 2021]. The results presented here agree with this notion and

provide evidence for gene dosage aberrations which are the product of *sSVs* between these repeats.

Several more loci show signs of a mechanistic association with CNVs. First, chromosomal region 22q11.2 can harbor local duplications and deletions associated with the DiGeorge Syndrome: The region contains a network of segmental duplications which is highly variable across humans and forms the mediator of the 22q11.2 deletion syndrome [Vervoort et al., 2021]. One *sSV* locus maps to a 618 kbp block of SDs flanking the region (Fig. 5.10A). The *sSV* features two overlapping pairs of inversely oriented SDs. Three divergent complex haplotypes emerge, with two showing either an increase or decrease in sequence content. Deletion or duplication of this region via *NAHR* are only possible from the inverted, not from the reference state of this locus. With this *sSV* block sitting at one breakpoint of the 22q11.2 CNV, I predict that the deletion and duplication may lower or increase, respectively, the risk for subsequent CNV formation (Fig. 5.10B).

The second example is located in chromosomal region 5q35. This region can harbor a ca. 2 Mbp CNV associated with the *SOTOS* syndrome likely caused by haploinsufficiency of the *NSD1* gene contained in this region [Tatton-Brown et al., 2005]. While repeats at the flanks of this region have been noted before as potential substrates of *NAHR* [Tatton-Brown et al., 2005], the exact formation has not been described to date. SDs are indeed present in the region, but the longest repeats are oriented inversely with respect to each other, priming for inversions in the region but not for CNVs (Fig. 5.10C). However, an inspection of SVs identified by NAHRWhal reveals a 200 kbp inversion of one copy of a segmental duplication present in 1/40 haplotypes (2.5%). This inversion flips another SD pair into direct orientation, making this inversion a likely pre-mutative state (Fig. 5.10D).

Lastly, complex rearrangements associated with chromosomal region 15q25.2 were also identified. Two highly variable *sSV* blocks in the region were identified (Figs. 5.8A and 5.11). Manual inspection reveals another inversion across the second *sSV* block and an adjacent region. The inverted segment contains segmental duplications, which are transferred to the adjacent region, likely turning it into a third *sSV* block. This process likely raises the risk for subsequent CNV formation between *sSV* blocks #1 and #3 and highlights yet another level of sequence dynamics, in which inversions can act as a means of SD 'transaction' between regions.

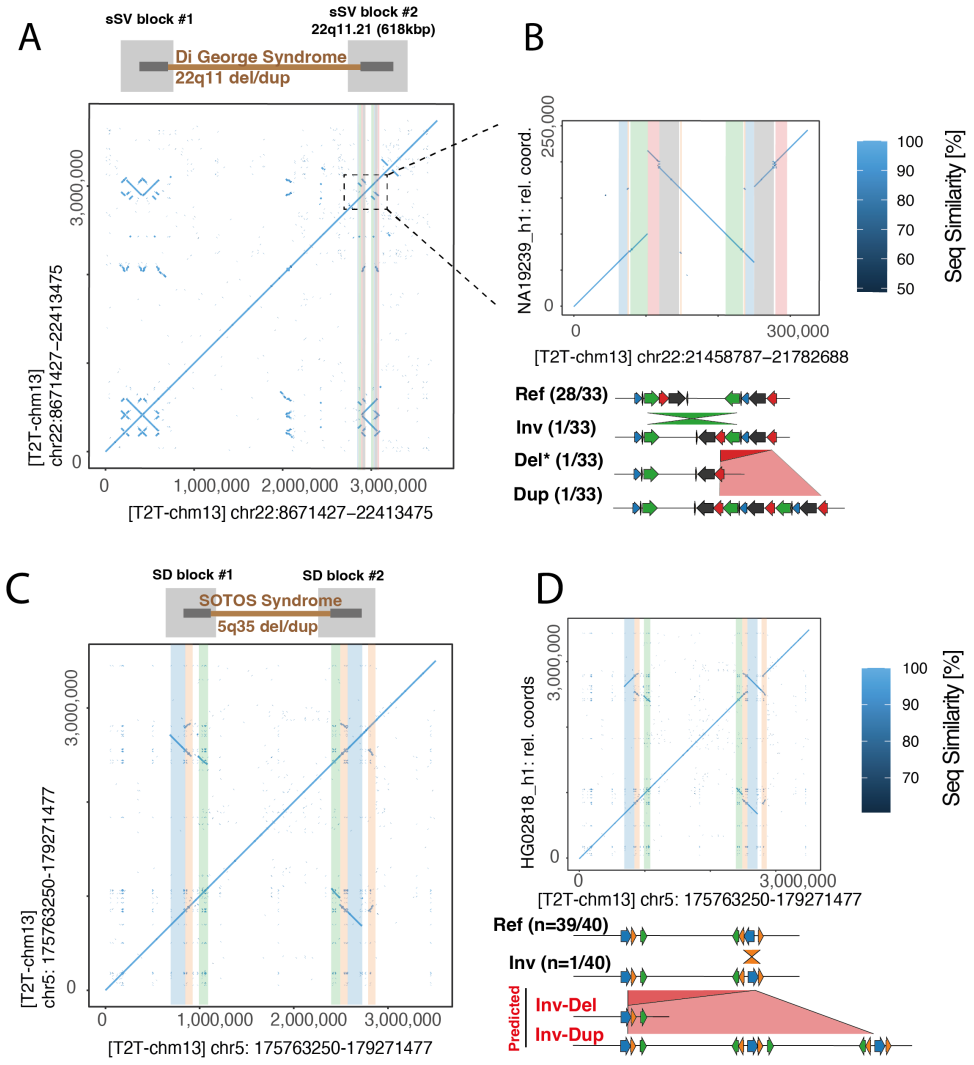

Figure 5.10: *sSVs* in disease-relevant regions. **A** *sSVs* flanking the 22q11 del/dup region. Both breakpoint regions show various NAHR- and non-NAHR rearrangements. **B** A nested SV mediates the removal of potentially CNV-predisposing SDs on one breakpoint of the 22q11 del/dup region. **C, D** Inversion of one breakpoint of the SOTOS deletion creates a long pair of directly oriented SVs likely predisposing to subsequent CNV formation.

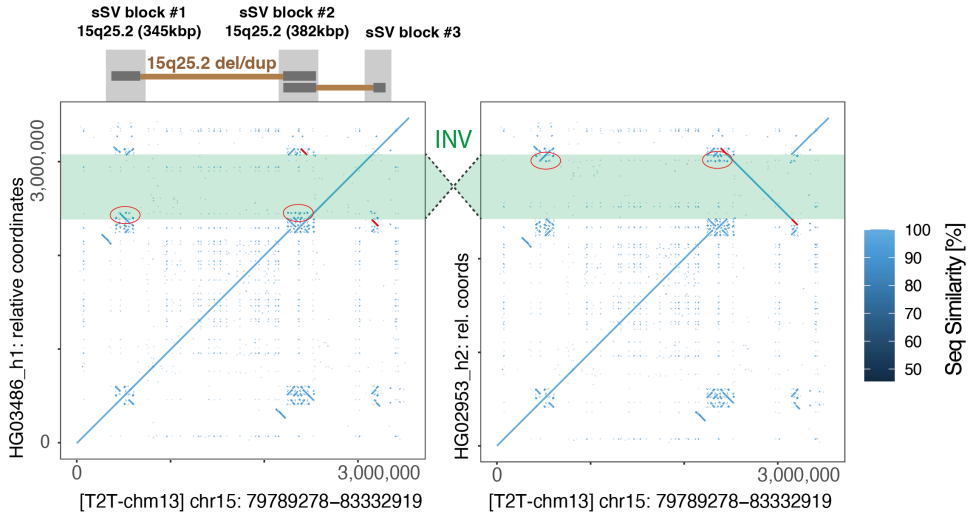

Figure 5.11: *sSVs* in the 15q25.2 microduplication/deletion region. Two *sSVs* map to two ends of the 15q25.2 deletion region, making both breakpoints susceptible to individual rearrangements. An inversion in one sample leads to the transfer of SD sequence to a third *sSV* block. Inversion-mediating SDs are highlighted in red.

5.4.2 sSV loci in great apes display additional forms of variation

As part of the work on NAHRWhals, I intend to examine human sSV loci in the genomes of great apes. While this part of the study is still in early development, an initial view of example sSV loci in great apes reveals unexpected rearrangements exceeding those typically seen in humans in such loci. For example, long inversions spanning the sSV locus and its surrounding are observed in an Orangutan haplotype of the sSV locus at chromosomal region 16p12.1-p11.2 (Fig. 5.12). A systematic evaluation of sSV loci in great apes is planned for the near future in order to investigate sSV locus evolution over greater evolutionary distances,

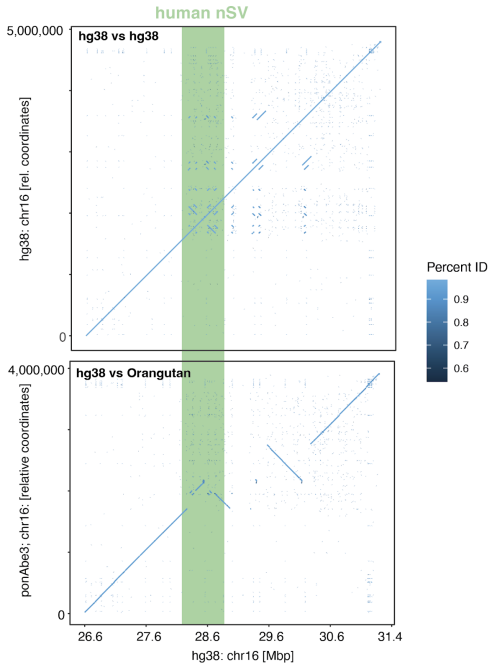

Figure 5.12: Dotplot view of the sSV locus in chromosomal region 16p12.1-p11.2 in humans (top) and orangutan (bottom). The locus has undergone a complex series of long-range SVs, which can not be explained by NAHR alone.

5.5 Complex rearrangements in chrY

Parallel to work on *NAHRwhals* presented so far, I became interested in a project on the diversity of the human Y-chromosome, which has been led by two members of the *HGSVC*, P. Hallast (The Jackson Laboratory, Farmington, US) and P. Ebert (Heinrich-Heine Universität Düsseldorf, Germany). In this project, 43 diverse human Y chromosomes were deeply sequenced with PacBio HIFI and Oxford Nanopore Technologies (ONT) long read data and subsequently applied for de-novo assembly of chrY using the Verkko assembler [Rautiainen et al., 2022]. The result of this effort was high-quality assemblies for Y chromosomes across 43 samples, with gapless assemblies (including heterochromatin) for 3/43 samples and continuous assemblies of 17/24 subregions across 41/43 samples. P. Hallast and I presumed that the *NAHRwhals* framework could likely help explain complex variation found across chrY haplotypes, thus spawning synergies between the two projects. Consequently, I attempted to apply *NAHRwhals* to identify potential multi-stage *NAHR* events that may have played a role in the evolution of this chromosome and which can currently not be identified by other methods.

5.5.1 Large-scale structural variations found across 43 chrY assemblies

NAHRwhals was used to generate views of pairwise sequence alignments for each assembled euchromatic region mapping to itself and its homologs in the hg38 and chm13-T2T reference sequences (example shown in Fig. 5.13 A, B). This approach unravels dozens of structural variation sites of varying size and complexity. These comprise very large variations, out of which three sites of multi-Mbp inversion variants are highlighted that span (1) the IR3 palindrome (3 carriers, 38 non-carriers, 3 incompletely resolved), (2) the P1 palindrome (6 carriers, 25 non-carriers, 13 incompletely resolved) and (3) a previously undescribed 9 Mbp inversion between the AMPL6 and AMPL7 region (2 carriers, 41 non-carriers) (Fig. 5.13C).

Additionally, I identified several instances of more complex large-scale structural variants using *NAHRwhals*. Focussing initially on the AMPL7 region, one large deletion variant spanning the first copy of the P1 LCRs in one sample becomes apparent, as well as a highly rearranged haplotype characterized by an *sSV* containing multiple overlapping inversion events (Fig. 5.14). The presence of directly oriented repeats at all breakpoints suggests *NAHR* as the driving mechanism behind both deletion variants, while two underlying inversion variants appear to alter the SD landscape, predisposing various inversion alleles to different copy number variants.

Lastly, an inverted duplication was identified that affects roughly two-thirds of the 161 kbp unique sequence in the P3 palindrome, spawns a second copy of the TTTY5 gene, and effectively elongates the LCRs in this region. In line with its unique nature among our assemblies, a detailed sequence view reveals a high sequence similarity between the duplication and its template, suggesting recent emergence of this variant (Figure 5.15). However, the mechanism behind this expansion is unknown, and further analysis might bring new insights into the formation and expansion of segmental duplications in the future.

5.6 Discussion

This chapter has explored the phenomenon of sequential *NAHR*-mediated SVs, a concept that has not been systematically explored before and which I have coined *sSV*.

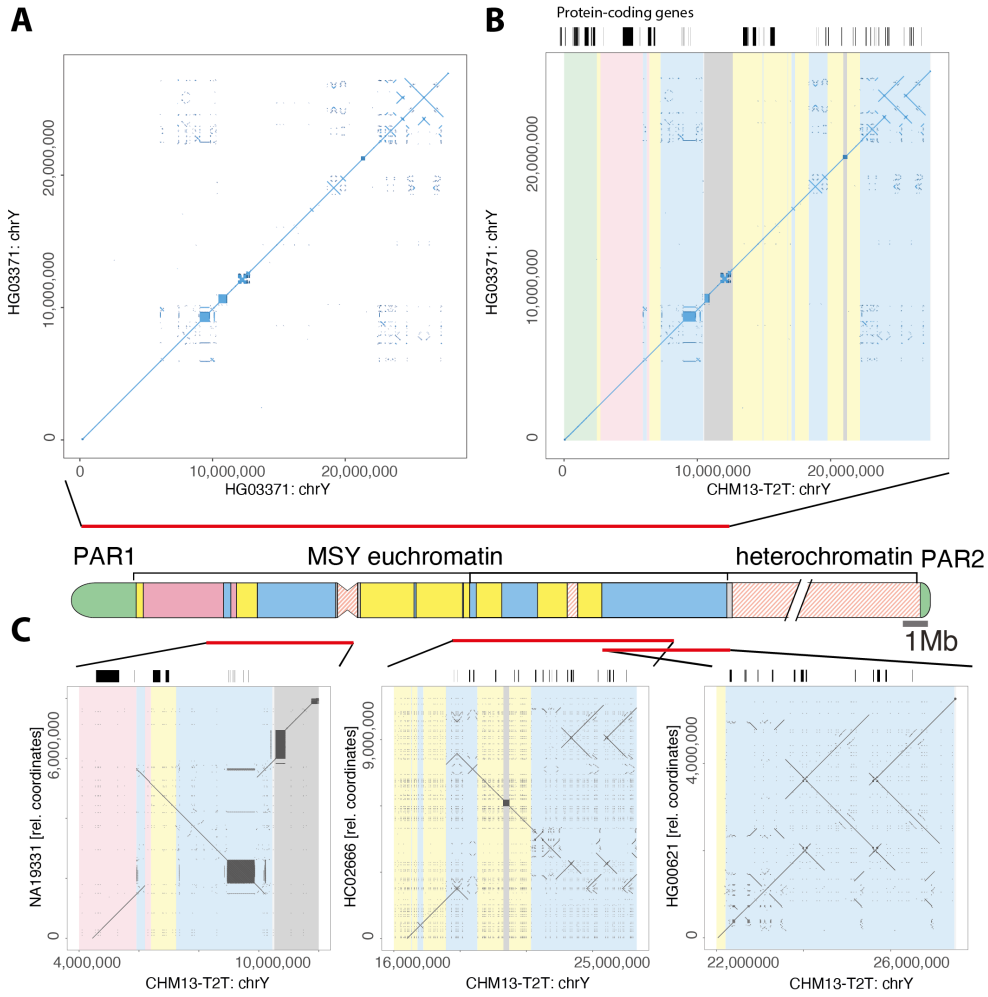

Figure 5.13: dot plot views of chrY assemblies and large SVs. A Visualization of a self-alignment of the euchromatic region of fully assembled chrY in sample HG03371. **B** The same chrY assembly aligned against the T2T reference assembly. **C** individual views of three large inversion variations identified in various samples. Background colors correspond to chromatic regions (green: pseudoautosomal (PAR), yellow: X-degenerate (XDR), red: X-transposed (XTR), blue: ampliconic (AMPL), grey: heterochromatic.

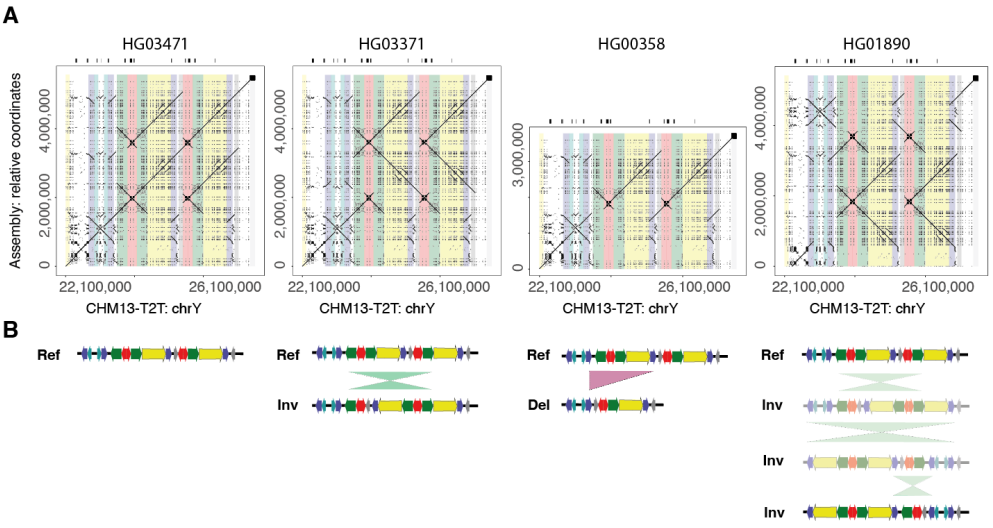

Figure 5.14: *sSV* rearrangements in the AMPL7 region on chrY. A dot plot views of four structurally distinct haplotypes of the AMPL7 region compared to the T2T reference. **B** Visualization of *NAHRwhals*-based *sSV* predictions. Two haplotypes can be explained via simple *NAHR* rearrangements, three sequential inversions are predicted in sample *HG01890*.

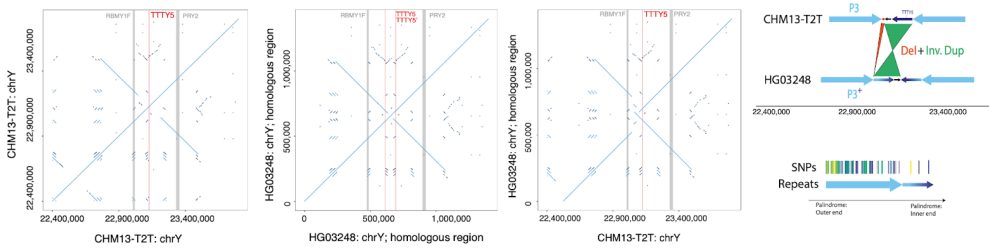

Figure 5.15: Identification of an inverted duplication of the TTTY5 gene. Three dot plots illustrating the reference and inversely-duplicated state. On the right side, the inferred underlying SV event is depicted. The inverted duplication is accompanied by a deletion, and effectively promotes elongation of the SD.

As part of this study, I have developed a new tool, *NAHRwhals*, which specializes in generating and visualizing long pairwise sequence alignments, and inferring likely strings of sequential mutations. Simulation experiments and examples from real datasets have shown that the tool can generally identify *sSVs* up to depth 3 with high accuracy. While the search depth appears to be sufficient for the majority of complex rearrangement, it might be possible to increase the depth of search by employing a more efficient algorithm for identifying SV chains in future versions of the software. Indeed, similar problems have been solved e.g. with graph based approaches with shorter runtime [Bonnet et al., 2022], and it might be feasible to attempt to integrate these into *NAHRWhals*. Another consideration for software improvement lies in the segmentation algorithm. Conceptually, segmentation algorithms used to compute genome graphs like sibeliaZ [Minkin and Medvedev, 2020] and PGGB (https://github.com/pangenome/pggb, unpublished) solve a similar task to the NAHRWHals-based segmentation, and a formal comparison with these approaches is still pending.

Applying *NAHRwhals* to a set of 213 SV-associated regions in 56 samples, I have identified 20 loci harboring *sSVs*, calling more than 100 individual events. While this number likely does not yet represent the complete set of human *sSVs*, it shows that *sSVs* are a ubiquitous phenomenon, and this exploratory study provides examples that help study them further in the future. Furthermore, detailed analyses of these regions have unveiled an unexpeceted diversity of haplotypes. Such diversity showcases how the relatively simple 'building blocks' of inversions, deletions, and duplications can synergize to alter, expand or translocate whole genome regions dynamically.

The findings presented here furthermore suggest that CNVs are intricately linked to *sSVs*, and many long CNV regions exhibit an unexpected diversity of sequence comformations near their borders. I found likely *sSV*-associated pre-mutative or protective states in four disease-associated regions: TPSAB1/2 duplications and deletions linked to *Alpha Tryptasaemia*, and three *sSVs* affecting the SD landscapes of CNV regions associated to the DiGeorge-, 15q25.2 Microdeletion- and Sotos syndromes. In the case of 15q25.2, one haplotype appeared to display the creation of a novel *sSV* block through the transfer of low-copy repeats through an inversion, a process that has not been described before.

I have also applied the tool to find SVs in 43 human haplotypes of chrY. This chromosome harbors a particularly large number of segmental duplications. Dot plot visualizations of large regions – up to 30 Mbp in size – proved essential for SV detection, seeing as three large SVs were identified, which had been missed

by classical SV calling approaches. Especially the AMPL7 region has proven to be a hotspot of *sSV* formation due to several long SDs intersecting in the region.

The results presented in this chapter illustrate the prominent role that *sSVs* play in genome evolution and show that complex webs of SDs can change and interact in the genome in unexpected ways. These analyses also highlight the power of high-quality genome assemblies, which are currently the only technology capable of exploring such complexity at a sufficiently high resolution. From the CNVs discussed in this chapter, as well as others from the previous chapter (chapter 4), I have now collected several examples of SVs acting as pre-mutative or protective states. It would be vital to test these hypotheses in patient data, an effort I hope to follow up on in the future.

6

Summary and Concluding remarks

Genomic inversions were first described over one hundred years ago, yet modern genetics can only provide an incomplete picture of the biology of this exceptional mutation class. The preceding chapters have documented several interconnected studies which have used recently emerging technologies to explore the specific properties of genomic inversions in different genomic contexts. Throughout these projects, it has become clear that the prevalence of inversion-mediated variation in humans and their close ancestors has been underestimated previously. Likewise, the field is only beginning to recognize the vast functional implications that these SVs carry for humans and other species. To summarize the key findings of this thesis, this last chapter reiterates the most striking findings presented in this thesis and attempts to place critical findings into a broader context. Finally, an outlook on future challenges and opportunities in the field will be given.

Inversions in the great ape lineage

Section 1.2 has introduced the three known ways in which inversions can affect genomic function: direct disturbance, association with secondary SVs, and suppression of recombination [Feuk et al., 2005]. The first project described in this thesis has offered an opportunity to test the effect of one of these, *disturbance*, in the context of the evolution of humans and non-human primates, specifically assessing the role of TAD disruptions.

Comparing a set of 687 human-ape inversions discovered by Strand-Seq with human TAD locations and bulk-RNA seq data revealed several novel insights into inversion formation. First, the breakpoints of long (> 100 kbp), but not of shorter inversions displayed a tendency to co-locate with TAD boundaries. It is unclear possible that this reflects an avoidance of adverse fitness effects, or that alternatively inversions are more likely to arise in such locations [Krefting et al., 2018]. It must be cautioned here, however, that the utilized TAD breakpoints are likely over-simplifying the real biological phenomena, as TADs are likely variable across cell types and develeopmental stages [Akdemir et al., 2020]. Nevertheless, TADs disrupted by inversions were slightly (1.13 – 1.15-fold) enriched for differentially expressed genes. Subsequent analysis revealed, however, that this enrichment is mainly driven by genes in close proximity (0 – 150 kbp) to the breakpoints of long inversions, rather than by all genes in broken TADs. Again, several limitations have to be taken into account: (1) Strand-Seq and RNA-seq data were obtained from different individuals of the same species. This can be especially problematic for studying recurrent inversions, as samples analyzed with RNA-seq may not display the identical alleles as those analyzed with Strand-Seq. Furthermore (2), inversions were exclusively identified with Strand-Seq, which limits the accuracy of identified breakpoints and may likely have obscured more complex rearrangements.

Taken together, these findings argue against a strong effect of TAD alterations on gene expression. Instead, they support the notion that TAD-based gene regulation is relatively robust towards structural changes, and the disruption of TADs does not necessarily lead to changes in gene expression. This notion is in line with several recent studies which, too, observed only a moderate effect of TAD breaks on gene expression in *drosophila* [Ghavi-Helm et al., 2019, Said et al., 2018] and humans[Schöpflin et al., 2022].

Inversion genotyping with ArbiGent

In response to the challenges associated with inversion detection, a new genotyping algorithm, *ArbiGent*, was presented in chapter 3. This algorithm uses

Strand-Seq data to accurately genotype structural variations across multiple samples by integrating orientation- and depth information from each cell for which Strand-Seq data is available. ArbiGent outperforms previous Strand-Seq-based genotyping methods mainly by improving read count normalization in regions with a high proportion of structural variations and low read mappability. In addition, the tool supports using all Strand-Seq libraries regardless of their 'strand-state' (see chapter 1.3), rather than using composite files, thereby retaining more information and reducing bias towards over-calling heterozygous inversions. ArbiGent also includes additional utilities to phase inversions and filter calls based on population-wide metrics. In accordance with previous expectations, *ArbiGent* displays limitations in sequence resolution due to the sparseness of Strand-Seq data, which limits its ability to genotype SVs shorter than 1-10 kilobases. As a genotyping tool, *ArbiGent* is furthermore unsuited to identify likely inversion regions de-novo, a limitation that can obscure sample-specific local differences, such as alternative breakpoints. Lastly, more complex events, like nested inversions or inversions associated with deletions, can only be classified correctly if the correct individual segments are passed as input. It should be noted that another Strand-Seq-based inversion genotyper was developed for a related task and shares some of its functionalities [Hanlon et al., 2021], providing an opportunity for cross-testing and improving both algorithms in the future. ArbiGent is expected to contribute to Strand-Seq-based studies and may be used as a genotyping tool for large inversion hotspots in regions with low numbers of cells per sample. To facilitate the continued use of *ArbiGent* by the Strand-Seq community, my colleague T. Weber has kindly helped with its integration into the latest release of the *MosaiCatcher* Strand-Seq toolbox.

Full-spectrum analysis of human inversions reveals hotspots of recurrence associated with genomic disorders

Chapter 4 presented the most comprehensive study of human inversion polymorphisms to date, which was enabled by an extensive collaborative effort surrounding a multi-technology inversion calling approach. Firstly, our consortium was able to define a set of 398 inversion loci across 86 diverse human haplotypes. Inversions could be assigned into mechanistically separate classes, and long events (>100 kbp) were typically flanked by segmental duplications, suggesting ectopic recombination (NAHR) as the predominant mechanism behind the formation of such rearrangements, which is in agreement with previous reports. Recurrence was confirmed in 40 individual loci, covering 0.6% of the human genome and highlighting the widespread nature of this feature of NAHR-based inversions. It

should be taken into account, however, that the sample size is still relatively small with only 86 haplotypes, leading to (1) an incomplete representation of rare inversions (MAF <5%) and (2) an under-estimation of inversion recurrence, as recurrence can only be determined when several unrelated samples carry the same inversion. Along these lines, previous inversion-based studies have speculated that almost NAHR-based inversions display some level of recurrence [Giner-Delgado et al., 2019].

Given that inversions are thought to play an essential role in local adaptation by suppressing recombination, recurrence may add another layer of complexity to this concept, as a region's orientation directly determines which alleles recombine. Given the variability of SDs, particular inversion loci may also lose – or gain – their recurrent properties over time, a possibility that has to be considered in future population genetics studies. Inversions are furthermore known to exhibit a close relationship with CNVs in several individual loci (see e.g., [Antonacci et al., 2009b, Koolen et al., 2016]). Drawing from a genome-wide view of inversion loci, one of the study's contributions was to broaden this view by identifying a general potential of inversions to alter the risk for subsequent CNVs by rearranging SD landscapes. While this effect has been clearly demonstrated in the three examples highlighted (3q29, 15q13.3, 7q11.23), the breakpoint architecture of many recurrent inversions still needs to be clarified, and secondary SVs like deletions or duplications at the flanks may still be left for discovery for future studies. Such future studies will likely also discover more examples of such inversion-CNV interplay in other disease-critical regions of the genome. Finally, while exceeding the scope of this study, the integration of primate inversion data would enable a confident identification of ancestral loci and bring more insights into the long-term evolutionary dynamics that (recurrent) inversions exhibit.

As the first of its kind, the inversion callset presented in this chapter poses a vital reference dataset for subsequent studies. Inversion calling is not yet routine and has thus required developing novel computational approaches. These efforts also confirm that no single technology can currently resolve inversions across all size ranges. Instead, multi-platform approaches are the only viable options to capture inversions of all sizes. Consequently, examination of this novel inversion callset has highlighted that especially long inversions have been systematically overlooked in preceding genomic studies [1000 Genomes Project Consortium et al., 2012, Sudmant et al., 2015, Chaisson et al., 2019]. Likewise, our knowledge of inversion recurrence has been advanced by this work, and subsequent studies will be able to expand on our insights by examining more samples at even higher resolutions. While the study of inversions in apes (chapter

2) has found that disruption of genes and TADs is likely not among their most critical evolutionary effects, recurrent inversions in humans do argue towards the importance of another effect, *secondary SV formation*. It is still unclear to what extent the re-organization of the SD landscape facilitates CNV formation. More studies will be needed to investigate all cases and mechanisms by which the rearrangement of SDs affects disease formation or evolutionary processes.

Nested repeats promote clusters of complex and highly dynamic SVs

Segmental Duplications can promote both inversions and CNVs, and complex genomic regions often display several layers of nested SDs in one locus [Vollger et al., 2022]. Accordingly, complex mutation patterns are conceivable in which subsequent rounds of inversions, duplications, and deletions can dramatically transform a locus over time. The analysis of selected sequence-resolved inversions in the context of CNVs in chapter 4 has revealed evidence for a mechanistic link between inversions and CNVs, in which inversions act as switches, promoting or prohibiting subsequent CNV by altering SD distance and orientation. Taking these preliminary findings as the starting point, the final chapter 5 has described the development and application of a new algorithm, *NAHRwhals*, for the systematic detection of such serial, overlapping NAHR-based rearrangements.

Using a custom segmentation algorithm and exhaustive mutation search, NAHRWhals can identify diverse classes of serial SVs (sSVs), as demonstrated on simulated and real data. As the first algorithm aimed at solving the task of mechanistically reconstructing sSVs, NAHRWhals introduces new concepts of sequence analysis. In the future, several aspects of the computational implementation can still be improved, e.g. by using alternative algorithms for SV-chain calling and alignment segmentation. On the level of input data, larger and more refined collections of de-novo assembled genomes will be needed to identify the full spectrum of sSVs and discover the full extent of their association with morbid CNVs.

With these limitations in mind, genome-wide analysis of inversion-associated rearrangements indicates at least twenty such sSV loci, which have undergone various degrees of complex rearrangements in the human population. sSVs are associated with morbid CNVs in many cases (6/20 regions), and such complex rearrangements are likely intimately linked with genomic evolution and disease. Likewise, the analysis of 43 chrY assemblies suggests that the SD-rich landscape of this chromosome has promoted extensive serial SV formation, especially in the AMPL7 region. Finally, comparing human assemblies of the sSV loci to

high-quality primate assemblies has revealed even more complex rearrangement patterns, which suggest that NAHR is not the sole driver of large-scale sequence evolution over long timespans (Mya).

NAHRWhals and the associated analyses provide an essential first step into leveraging the novel sequence resolution of complex, SD-enriched regions to unravel large rearrangements beyond 'simple' SVs, which have constituted the focus of most large-scale studies conducted to identify and characterize SVs.

6.1 Future outlook

The work presented here is embedded in a large number of recent studies which aim to characterize genomic variation in ever greater numbers and detail. Even within the time frame of this thesis, the field's understanding of inversion polymorphisms has progressed rapidly, partly because sequence resolution of large SVs has become more feasible during these years. Still, several open questions remain, some of which will depend on deeper knowledge and more advanced technology than those available to date.

Due to the high intensity in cost and resources to perform inversion detection using a multi-technology approach, inversion detection is currently limited to relatively small sample sizes (such as 43 samples in the study in chapter 4). Low sample numbers are especially problematic in light of rare events (illustrated e.g., by a single 25-Mbp inversion present in one sample in [Porubsky et al., 2022a]). Several technologies might contribute to this endeavor in future studies:

First, the Strand-Seq technology has been a forerunner in inversion detection for many years due to its inherent strength in detecting long inversions. In previous studies, Strand-Seq has been used on a sample-by-sample basis, where one sequencing experiment produces 96 single-cell libraries from one sample. Data from different cells are often integrated across cells in the analysis stage, allowing for an increased resolution in clonal events. However, the single-cell nature of Strand-Seq also allows pooling cells from different samples into one sequencing run, producing 96 cells from a pool of, e.g., 10 samples. This strategy effectively reduces the cost-per-sample by a factor of 10 while still producing an average of 5-10 cells per sample – typically enough to identify inversions longer than ca. 50 kbp. The viability of such a pooled experiment has been demonstrated in [Porubsky et al., 2022a]. By this means of cost reduction, Strand-Seq could be utilized to identify inversions in hundreds up to a few thousand samples, which would be the largest to date systematic survey of long inversions in humans.

In parallel, (ultra) long-read sequencing, e.g., provided by ONT or PacBio HIFI sequencing, is on the rise. These technologies are suitable for detecting most short to medium-sized (10's of kbp) and many longer inversions. Long-read sequencing has been steadily improving in accuracy while becoming more affordable in parallel. Consequently, sizeable population-wide genome sequencing projects increasingly rely on long-read technologies to maximize SV discovery capabilities. Prominent examples today include the human pan-genome reference (HPRC) [Wang et al., 2020], the All-of-Us initiative [The "All of Us" Research Program, 2019], and efforts to sequence large proportions of the Icelandic population [Beyter et al., 2021]. Given the general demand of science and medicine for big data sets (e.g., to advance precision medicine [Suwinski et al., 2019]), the number and scale of such projects will likely continue to increase. Similarly to other classes of genomic variations, more inversions will be identified as such projects continue. Even if the length and accuracy of sequencing data may only sometimes be sufficient to call long inversions accurately, such datasets can later be combined with other approaches, such as Strand-Seq, and will likely provide an invaluable basis for future studies.

Apart from increasing the breadth of inversion discovery, future efforts will also be able to study inversion events at a higher resolution than currently possible, identifying additional complexities such as secondary SVs at the breakpoints. Such analyses might also exhibit new insights into formation mechanisms (e.g., regarding the role of transposable elements [Balachandran et al., 2022]). More importantly, sequence resolution in complex or SD-rich regions will allow accurately describing large, non-trivial rearrangements, alleviating the custom of calling such events' complex' or 'cryptic' SVs. Finally, increased sequencing resolution will also be crucial for reconstructing the evolution of the human genome (or that of other species), in which inversions, CNVs, and SDs appear tightly interlinked in a way that is essentially unclear to date.

Perhaps more so than most other SVs, inversions have retained many unknowns and are yet to be characterized in full. With this perspective in mind, it becomes clear that our current knowledge of inversions – including the work presented in this thesis – provides merely a stepping stone to understanding this peculiar class of SVs. In conclusion – there has not been a better time to study inversions.

Bibliography

[1000 Genomes Project Consortium et al., 2012] 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

[1000 Genomes Project Consortium et al., 2015] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[Aguado et al., 2014] Aguado, C., Gayà-Vidal, M., Villatoro, S., Oliva, M., Izquierdo, D., Giner-Delgado, C., Montalvo, V., García-González, J., Martínez-Fundichely, A., Capilla, L., Ruiz-Herrera, A., Estivill, X., Puig, M., and Cáceres, M. (2014). Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS Genet.*, 10(3):e1004208.

[Akdemir et al., 2020] Akdemir, K. C., Le, V. T., Chandran, S., Li, Y., Verhaak, R. G., Beroukhim, R., Campbell, P. J., Chin, L., Dixon, J. R., Futreal, P. A., PCAWG Structural Variation Working Group, and PCAWG Consortium (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.*, 52(3):294–305.

[Aken et al., 2017] Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Juettemann, T., Keenan, S., Laird, M. R., Lavidas, I., Maurel, T., McLaren, W., Moore, B., Murphy, D. N., Nag, R., Newman, V., Nuhn, M., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Wilder, S. P., Zadissa, A., Kostadima, M., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Cunningham, F., Yates, A., Zerbino, D. R., and Flicek, P. (2017). Ensembl 2017. *Nucleic Acids Res.*, 45(D1):D635–D642.

[Alkan et al., 2011] Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, 12(5):363–376.

[Anderson et al., 2005] Anderson, A. R., Hoffmann, A. A., McKechnie, S. W., Umina, P. A., and Weeks, A. R. (2005). The latitudinal cline in the In(3R)Payne inversion polymorphism has shifted in the last 20 years in australian drosophila melanogaster populations. *Mol. Ecol.*, 14(3):851–858.

[Andolfatto et al., 2001] Andolfatto, P., Depaulis, F., and Navarro, A. (2001). Inversion polymorphisms and nucleotide variability in drosophila. *Genet. Res.*, 77(1):1–8.

[Antonacci et al., 2014] Antonacci, F., Dennis, M. Y., Huddleston, J., Sudmant, P. H., Steinberg, K. M., Rosenfeld, J. A., Miroballo, M., Graves, T. A., Vives, L., Malig, M., Denman, L., Raja, A., Stuart, A., Tang, J., Munson, B., Shaffer, L. G., Amemiya, C. T., Wilson, R. K., and Eichler, E. E. (2014). Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat. Genet.*, 46(12):1293–1302.

[Antonacci et al., 2009a] Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., and Eichler, E. E. (2009a). Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.*, 18(14):2555–2566.

[Antonacci et al., 2009b] Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., and Eichler, E. E. (2009b). Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.*, 18(14):2555–2566.

[Arenas and Posada, 2014] Arenas, M. and Posada, D. (2014). Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. *Mol. Biol. Evol.*, 31(5):1295–1301.

[Bailey and Eichler, 2006] Bailey, J. A. and Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.*, 7(7):552–564.

[Balachandran et al., 2022] Balachandran, P., Walawalkar, I. A., Flores, J. I., Dayton, J. N., Audano, P. A., and Beck, C. R. (2022). Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.*, 13(1):7115.

[Belton et al., 2012] Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276.

[Beyter et al., 2021] Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., Gudjonsson, S. A., Magnusdottir, D. N., Jonasdottir, A., Jonasdottir, A., Kristjansson, R. P., Sverrisson, S. T., Holley, G., Palsson, G., Stefansson, O. A., Eyjolfsson, G., Olafsson, I., Sigurdardottir, O., Torfason, B., Masson, G., Helgason, A., Thorsteinsdottir, U., Holm, H., Gudbjartsson, D. F., Sulem, P., Magnusson, O. T., Halldorsson, B. V., and Stefansson, K. (2021). Long-read sequencing of 3,622 icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.*, 53(6):779–786.

[Bonnet et al., 2022] Bonnet, K., Marschall, T., and Doerr, D. (2022). Constructing founder sets under allelic and non-allelic homologous recombination. *bioRxiv*, page 2022.05.27.493721.

[Bragin et al., 2014] Bragin, E., Chatzimichali, E. A., Wright, C. F., Hurles, M. E., Firth, H. V., Bevan, A. P., and Swaminathan, G. J. (2014). DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res.*, 42(Database issue):D993–D1000.

[Brawand et al., 2011] Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

[Cáceres et al., 2007] Cáceres, M., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, Sullivan, R. T., and Thomas, J. W. (2007). A recurrent inversion on the eutherian X chromosome. *Proc. Natl. Acad. Sci. U. S. A.*, 104(47):18571–18576.

[Cantsilieris et al., 2020] Cantsilieris, S., Sunkin, S. M., Johnson, M. E., Anaclerio, F., Huddleston, J., Baker, C., Dougherty, M. L., Underwood, J. G., Sulovari, A., Hsieh, P., Mao, Y., Catacchio, C. R., Malig, M., Welch, A. E., Sorensen, M., Munson, K. M., Jiang, W., Girirajan, S., Ventura, M., Lamb, B. T., Conlon, R. A., and Eichler, E. E. (2020). An evolutionary driver of interspersed segmental duplications in primates. *Genome Biol.*, 21(1):202.

[Capozzi et al., 2012] Capozzi, O., Carbone, L., Stanyon, R. R., Marra, A., Yang, F., Whelan, C. W., de Jong, P. J., Rocchi, M., and Archidiacono, N. (2012). A comprehensive molecular cytogenetic analysis of chromosome rearrangements in gibbons. *Genome Res.*, 22(12):2520–2528.

Bibliography

[Carbone et al., 2002] Carbone, L., Ventura, M., Tempesta, S., Rocchi, M., and Archidiacono, N. (2002). Evolutionary history of chromosome 10 in primates. *Chromosoma*, 111(4):267–272.

[Carvalho and Lupski, 2016] Carvalho, C. M. B. and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.*, 17(4):224–238.

[Catacchio et al., 2018] Catacchio, C. R., Maggiolini, F. A. M., D'Addabbo, P., Bitonto, M., Capozzi, O., Lepore Signorile, M., Miroballo, M., Archidiacono, N., Eichler, E. E., Ventura, M., and Antonacci, F. (2018). Inversion variants in human and primate genomes. *Genome Res.*, 28(6):910–920.

[Chaisson et al., 2019] Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., Hastie, A. R., Antaki, D., Anantharaman, T., Audano, P. A., Brand, H., Cantsilieris, S., Cao, H., Cerveira, E., Chen, C., Chen, X., Chin, C.-S., Chong, Z., Chuang, N. T., Lambert, C. C., Church, D. M., Clarke, L., Farrell, A., Flores, J., Galeev, T., Gorkin, D. U., Gujral, M., Guryev, V., Heaton, W. H., Korlach, J., Kumar, S., Kwon, J. Y., Lam, E. T., Lee, J. E., Lee, J., Lee, W.-P., Lee, S. P., Li, S., Marks, P., Viaud-Martinez, K., Meiers, S., Munson, K. M., Navarro, F. C. P., Nelson, B. J., Nodzak, C., Noor, A., Kyriazopoulou-Panagiotopoulou, S., Pang, A. W. C., Qiu, Y., Rosanio, G., Ryan, M., Stütz, A., Spierings, D. C. J., Ward, A., Welch, A. E., Xiao, M., Xu, W., Zhang, C., Zhu, Q., Zheng-Bradley, X., Lowy, E., Yakneen, S., McCarroll, S., Jun, G., Ding, L., Koh, C. L., Ren, B., Flicek, P., Chen, K., Gerstein, M. B., Kwok, P.-Y., Lansdorp, P. M., Marth, G. T., Sebat, J., Shi, X., Bashir, A., Ye, K., Devine, S. E., Talkowski, M. E., Mills, R. E., Marschall, T., Korbel, J. O., Eichler, E. E., and Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, 10(1):1784.

[Charlesworth and Charlesworth, 2000] Charlesworth, B. and Charlesworth, D. (2000). The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 355(1403):1563–1572.

[Chen et al., 2014] Chen, L., Zhou, W., Zhang, L., and Zhang, F. (2014). Genome architecture and its roles in human copy number variation. *Genomics Inform.*, 12(4):136–144.

[Cheng et al., 2022] Cheng, J., Clayton, J. S., Acemel, R. D., Zheng, Y., Taylor, R. L., Keleş, S., Franke, M., Boackle, S. A., Harley, J. B., Quail, E., Gómez-Skarmeta, J. L., and Ulgiati, D. (2022). Regulatory architecture of the RCA gene cluster captures an intragenic TAD boundary, CTCF-Mediated chromatin looping and a Long-Range intergenic enhancer. *Front. Immunol.*, 13:901747.

[Coe et al., 2014] Coe, B. P., Witherspoon, K., Rosenfeld, J. A., van Bon, B. W. M., Vulto-van Silfhout, A. T., Bosco, P., Friend, K. L., Baker, C., Buono, S., Vissers, L. E. L. M., Schuurs-Hoeijmakers, J. H., Hoischen, A., Pfundt, R., Krumm, N., Carvill, G. L., Li, D., Amaral, D., Brown, N., Lockhart, P. J., Scheffer, I. E., Alberti, A., Shaw, M., Pettinato, R., Tervo, R., de Leeuw, N., Reijnders, M. R. F., Torchia, B. S., Peeters, H., O'Roak, B. J., Fichera, M., Hehir-Kwa, J. Y., Shendure, J., Mefford, H. C., Haan, E., Gécz, J., de Vries, B. B. A., Romano, C., and Eichler, E. E. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.*, 46(10):1063–1071.

[Collins et al., 2020] Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y.,

Brookings, T., Sharpe, T., Stone, M. R., Valkanas, E., Fu, J., Tiao, G., Laricchia, K. M., Ruano-Rubio, V., Stevens, C., Gupta, N., Cusick, C., Margolin, L., Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Taylor, K. D., Lin, H. J., Rich, S. S., Post, W. S., Chen, Y.-D. I., Rotter, J. I., Nusbaum, C., Philippakis, A., Lander, E., Gabriel, S., Neale, B. M., Kathiresan, S., Daly, M. J., Banks, E., MacArthur, D. G., and Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, 581(7809):444–451.

[Collins et al., 2017] Collins, R. L., Brand, H., Redin, C. E., Hanscom, C., Antolik, C., Stone, M. R., Glessner, J. T., Mason, T., Pregno, G., Dorrani, N., Mandrile, G., Giachino, D., Perrin, D., Walsh, C., Cipicchio, M., Costello, M., Stortchevoi, A., An, J.-Y., Currall, B. B., Seabra, C. M., Ragavendran, A., Margolin, L., Martinez-Agosto, J. A., Lucente, D., Levy, B., Sanders, S. J., Wapner, R. J., Quintero-Rivera, F., Kloosterman, W., and Talkowski, M. E. (2017). Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.*, 18(1):36.

[Cooper et al., 2011] Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., Kussmann, J., Shashi, V., Johnson, K., Rehder, C., Ballif, B. C., Shaffer, L. G., and Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nat. Genet.*, 43(9):838–846.

[Cosenza et al., 2022] Cosenza, M. R., Rodriguez-Martin, B., and Korbel, J. O. (2022). Structural variation in cancer: Role, prevalence, and mechanisms. *Annu. Rev. Genomics Hum. Genet.*, 23:123–152.

[Dekker and Heard, 2015] Dekker, J. and Heard, E. (2015). Structural and functional diversity of topologically associating domains. *FEBS Lett.*, 589(20 Pt A):2877–2884.

[Delahaye and Nicolas, 2021] Delahaye, C. and Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS One*, 16(10):e0257521.

[Despang et al., 2019] Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Wittler, L., Vingron, M., Mundlos, S., and Ibrahim, D. M. (2019). Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.*, 51(8):1263–1271.

[Dixon et al., 2012] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[Ebert et al., 2021] Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., Santamarina, M., Höps, W., Ashraf, H., Chuang, N. T., Yang, X., Munson, K. M., Lewis, A. P., Fairley, S., Tallon, L. J., Clarke, W. E., Basile, A. O., Byrska-Bishop, M., Corvelo, A., Evani, U. S., Lu, T.-Y., Chaisson, M. J. P., Chen, J., Li, C., Brand, H., Wenger, A. M., Ghareghani, M., Harvey, W. T., Raeder, B., Hasenfeld, P., Regier, A. A., Abel, H. J., Hall, I. M., Flicek, P., Stegle, O., Gerstein, M. B., Tubio, J. M. C., Mu, Z., Li, Y. I., Shi, X., Hastie, A. R., Ye, K., Chong, Z.,

Sanders, A. D., Zody, M. C., Talkowski, M. E., Mills, R. E., Devine, S. E., Lee, C., Korbel, J. O., Marschall, T., and Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537).

[Ebler et al., 2022] Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., and Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.*, 54(4):518–525.

[Eslami Rasekh et al., 2017] Eslami Rasekh, M., Chiatante, G., Miroballo, M., Tang, J., Ventura, M., Amemiya, C. T., Eichler, E. E., Antonacci, F., and Alkan, C. (2017). Discovery of large genomic inversions using long range information. *BMC Genomics*, 18(1):65.

[Falconer et al., 2012] Falconer, E., Hills, M., Naumann, U., Poon, S. S. S., Chavez, E. A., Sanders, A. D., Zhao, Y., Hirst, M., and Lansdorp, P. M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods*, 9(11):1107–1112.

[Faria et al., 2019] Faria, R., Johannesson, K., Butlin, R. K., and Westram, A. M. (2019). Evolving inversions. *Trends Ecol. Evol.*, 34(3):239–248.

[Farré et al., 2013] Farré, M., Micheletti, D., and Ruiz-Herrera, A. (2013). Recombination rates and genomic shuffling in human and chimpanzee–a new twist in the chromosomal speciation theory. *Mol. Biol. Evol.*, 30(4):853–864.

[Feuk et al., 2005] Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., Khaja, R., and Scherer, S. W. (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.*, 1(4):e56.

[Frankish et al., 2019] Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1):D766–D773.

[Gel et al., 2016] Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., and Malinverni, R. (2016). regioner: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, 32(2):289–291.

[Ghareghani et al., 2018] Ghareghani, M., Porubskỳ, D., Sanders, A. D., Meiers, S., Eichler, E. E., Korbel, J. O., and Marschall, T. (2018). Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics*, 34(13):i115–i123.

[Ghavi-Helm et al., 2019] Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O., and Furlong, E. E. M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.*, 51(8):1272–1282.

[Giglio et al., 2001] Giglio, S., Broman, K. W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R., Weber, J. L., Ledbetter, D. H., and Zuffardi, O. (2001). Olfactory Receptor–Gene clusters, Genomic-Inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.*, 68(4):874–883.

[Giner-Delgado et al., 2019] Giner-Delgado, C., Villatoro, S., Lerga-Jaso, J., Gayà-Vidal, M., Oliva, M., Castellano, D., Pantano, L., Bitarello, B. D., Izquierdo, D., Noguera, I., Olalde, I., Delprat, A., Blancher, A., Lalueza-Fox, C., Esko, T., O'Reilly, P. F., Andrés, A. M., Ferretti, L., Puig, M., and Cáceres, M. (2019). Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.*, 10(1):1–14.

[Goel et al., 2019] Goel, M., Sun, H., Jiao, W.-B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.*, 20(1):277.

[Hanlon et al., 2021] Hanlon, V. C. T., Mattsson, C.-A., Spierings, D. C. J., Guryev, V., and Lansdorp, P. M. (2021). InvertypeR: Bayesian inversion genotyping with strand-seq data. *BMC Genomics*, 22(1):582.

[Harringmeyer and Hoekstra, 2022] Harringmeyer, O. S. and Hoekstra, H. E. (2022). Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat Ecol Evol*.

[Hastings et al., 2009] Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, 10(8):551–564.

[Hnisz et al., 2016] Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J., and Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458.

[Ho et al., 2020] Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural variation in the sequencing era. *Nat. Rev. Genet.*, 21(3):171–189.

[Hsieh et al., 2021] Hsieh, P., Dang, V., Vollger, M. R., Mao, Y., Huang, T.-H., Dishuck, P. C., Baker, C., Cantsilieris, S., Lewis, A. P., Munson, K. M., Sorensen, M., Welch, A. E., Underwood, J. G., and Eichler, E. E. (2021). Evidence for opposing selective forces operating on human-specific duplicated TCAF genes in neanderthals and humans. *Nat. Commun.*, 12(1):5118.

[International Human Genome Sequencing Consortium, 2001] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Itsara et al., 2012] Itsara, A., Vissers, L. E. L. M., Steinberg, K. M., Meyer, K. J., Zody, M. C., Koolen, D. A., de Ligt, J., Cuppen, E., Baker, C., Lee, C., Graves, T. A., Wilson, R. K., Jenkins, R. B., Veltman, J. A., and Eichler, E. E. (2012). Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *Am. J. Hum. Genet.*, 90(4):599–613.

[Johnson et al., 2006] Johnson, M. E., National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Cheng, Z., Morrison, V. A., Scherer, S., Ventura, M., Gibbs, R. A., Green, E. D., and Eichler, E. E. (2006). Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 103(47):17626–17631.

[Jost et al., 2017] Jost, D., Vaillant, C., and Meister, P. (2017). Coupling 1D modifications and 3D nuclear organization: data, models and function. *Curr. Opin. Cell Biol.*, 44:20–27.

[Kannan and Zilfalil, 2009] Kannan, T. P. and Zilfalil, B. A. (2009). Cytogenetics: past, present and future. *Malays. J. Med. Sci.*, 16(2):4–9.

Bibliography

[Karolchik et al., 2004] Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, 32(Database issue):D493–6.

[Kehrer-Sawatzki et al., 2005] Kehrer-Sawatzki, H., Sandig, C., Chuzhanova, N., Goidts, V., Szamalek, J. M., Tänzer, S., Müller, S., Platzer, M., Cooper, D. N., and Hameister, H. (2005). Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (pan troglodytes). *Hum. Mutat.*, 25(1):45–55.

[Kidd and Kidd, 2007] Kidd, K. K. and Kidd, J. R. (2007). *Human genetic variation of medical significance*, volume 2. Oxford University Press.

[Kirkpatrick, 2010] Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biol.*, 8(9).

[Kirkpatrick and Barton, 2006] Kirkpatrick, M. and Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1):419–434.

[Koolen et al., 2016] Koolen, D. A., Pfundt, R., Linda, K., Beunders, G., Veenstra-Knol, H. E., Conta, J. H., Fortuna, A. M., Gillessen-Kaesbach, G., Dugan, S., Halbach, S., and Others (2016). The koolen-de vries syndrome: a phenotypic comparison of patients with a 17q21. 31 microdeletion versus a KANSL1 sequence variant. *Eur. J. Hum. Genet.*, 24(5):652–659.

[Kozel et al., 2021] Kozel, B. A., Barak, B., Kim, C. A., Mervis, C. B., Osborne, L. R., Porter, M., and Pober, B. R. (2021). Williams syndrome. *Nat Rev Dis Primers*, 7(1):42.

[Kragesteen et al., 2018] Kragesteen, B. K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., Annunziatella, C., Bianco, S., Chiariello, A. M., Jerković, I., Harabula, I., Guckelberger, P., Pechstein, M., Wittler, L., Chan, W.-L., Franke, M., Lupiáñez, D. G., Kraft, K., Timmermann, B., Vingron, M., Visel, A., Nicodemi, M., Mundlos, S., and Andrey, G. (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.*, 50(10):1463–1473.

[Krefting et al., 2018] Krefting, J., Andrade-Navarro, M. A., and Ibn-Salem, J. (2018). Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol.*, 16(1):87.

[Kronenberg et al., 2018] Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., Underwood, J. G., Nelson, B. J., Chaisson, M. J. P., Dougherty, M. L., Munson, K. M., Hastie, A. R., Diekhans, M., Hormozdiari, F., Lorusso, N., Hoekzema, K., Qiu, R., Clark, K., Raja, A., Welch, A. E., Sorensen, M., Baker, C., Fulton, R. S., Armstrong, J., Graves-Lindsay, T. A., Denli, A. M., Hoppe, E. R., Hsieh, P., Hill, C. M., Pang, A. W. C., Lee, J., Lam, E. T., Dutcher, S. K., Gage, F. H., Warren, W. C., Shendure, J., Haussler, D., Schneider, V. A., Cao, H., Ventura, M., Wilson, R. K., Paten, B., Pollen, A., and Eichler, E. E. (2018). High-resolution comparative analysis of great ape genomes. *Science*, 360(6393).

[Krueger, 2012] Krueger, F. (2012). Trim galore: a wrapper tool around cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (reduced representation Bisufite-Seq) libraries. *URL http://www. bioinformatics. babraham. ac. uk/projects/trim_galore/. (Date of access: 28/04/2016)*.

[Lahn and Page, 1999] Lahn, B. T. and Page, D. C. (1999). Four evolutionary strata on the human X chromosome. *Science*, 286(5441):964–967.

[Lakich et al., 1993] Lakich, D., Kazazian, Jr, H. H., Antonarakis, S. E., and Gitschier, J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia a. *Nat. Genet.*, 5(3):236–241.

[Lam et al., 2012] Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., and Kwok, P.-Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.*, 30(8):771–776.

[Lam and Jeffreys, 2006] Lam, K.-W. G. and Jeffreys, A. J. (2006). Processes of copy-number change in human DNA: The dynamics of α-globin gene deletion. *Proceedings of the National Academy of Sciences*, 103(24):8921–8927.

[Landeen and Presgraves, 2013] Landeen, E. L. and Presgraves, D. C. (2013). Evolution: From autosomes to sex chromosomes — and back.

[Lazar et al., 2018] Lazar, N. H., Nevonen, K. A., O'Connell, B., McCann, C., O'Neill, R. J., Green, R. E., Meyer, T. J., Okhovat, M., and Carbone, L. (2018). Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.*, 28(7):983–997.

[Li, 2018] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

[Liao et al., 2022] Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., Garg, S., Groza, C., Guarracino, A., Harvey, W. T., Heumos, S., Howe, K., Jain, M., Lu, T.-Y., Markello, C., Martin, F. J., Mitchell, M. W., Munson, K. M., Mwaniki, M. N., Novak, A. M., Olsen, H. E., Pesout, T., Porubsky, D., Prins, P., Sibbesen, J. A., Tomlinson, C., Villani, F., Vollger, M. R., Human Pangenome Reference Consortium, Bourque, G., Chaisson, M. J. P., Flicek, P., Phillippy, A. M., Zook, J. M., Eichler, E. E., Haussler, D., Jarvis, E. D., Miga, K. H., Wang, T., Garrison, E., Marschall, T., Hall, I., Li, H., and Paten, B. (2022). A draft human pangenome reference. *bioRxiv*, page 2022.07.09.499321.

[Liao et al., 2014] Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

[Lippert et al., 2014] Lippert, C., Casale, F. P., Rakitsch, B., and Stegle, O. (2014). LIMIX: genetic analysis of multiple traits. *bioRxiv*, page 003905.

[Liu et al., 2014] Liu, L., Missirian, V., Zinkgraf, M., Groover, A., and Filkov, V. (2014). Evaluation of experimental design and computational parameter choices affecting analyses of ChIP-seq and RNA-seq data in undomesticated poplar trees. *BMC Genomics*, 15 Suppl 5:S3.

[Liu et al., 2011] Liu, P., Lacaria, M., Zhang, F., Withers, M., Hastings, P. J., and Lupski, J. R. (2011). Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *Am. J. Hum. Genet.*, 89(4):580–588.

[Liu et al., 2022] Liu, Z., Roberts, R., Mercer, T. R., Xu, J., Sedlazeck, F. J., and Tong, W. (2022). Author correction: Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol.*, 23(1):198.

[Loose et al., 2018] Loose, M., Rakyan, V., Holmes, N., and Payne, A. (2018). Whale watching with BulkVis: A graphical viewer for oxford nanopore bulk fast5 files. *biorXiv*, 35(312256).

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550.

[Loveland et al., 2021] Loveland, J. L., Lank, D. B., and Küpper, C. (2021). Gene expression modification by an autosomal inversion associated with three male mating morphs. *Front. Genet.*, 12:641620.

[Lozier et al., 2002] Lozier, J. N., Dutra, A., Pak, E., Zhou, N., Zheng, Z., Nichols, T. C., Bellinger, D. A., Read, M., and Morgan, R. A. (2002). The chapel hill hemophilia a dog colony exhibits a factor VIII gene inversion. *Proc. Natl. Acad. Sci. U. S. A.*, 99(20):12991–12996.

[Lupiáñez et al., 2015] Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.

[Lupiáñez et al., 2016] Lupiáñez, D. G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How alterations of chromatin domains result in disease. *Trends Genet.*, 32(4):225–237.

[Lyons, 2021] Lyons, J. J. (2021). On the complexities of tryptase genetics and impact on clinical phenotypes. *J. Allergy Clin. Immunol.*, 148(5):1342–1343.

[Lyons et al., 2016] Lyons, J. J., Yu, X., Hughes, J. D., Le, Q. T., Jamil, A., Bai, Y., Ho, N., Zhao, M., Liu, Y., O'Connell, M. P., Trivedi, N. N., Nelson, C., DiMaggio, T., Jones, N., Matthews, H., Lewis, K. L., Oler, A. J., Carlson, R. J., Arkwright, P. D., Hong, C., Agama, S., Wilson, T. M., Tucker, S., Zhang, Y., McElwee, J. J., Pao, M., Glover, S. C., Rothenberg, M. E., Hohman, R. J., Stone, K. D., Caughey, G. H., Heller, T., Metcalfe, D. D., Biesecker, L. G., Schwartz, L. B., and Milner, J. D. (2016). Elevated basal serum tryptase identifies a multisystem disorder associated with increased TPSAB1 copy number. *Nat. Genet.*, 48(12):1564–1569.

[Maggiolini et al., 2020a] Maggiolini, F. A. M., Mercuri, L., Antonacci, F., Anaclerio, F., Calabrese, F. M., Lorusso, N., L'Abbate, A., Sorensen, M., Giannuzzi, G., Eichler, E. E., Catacchio, C. R., and Ventura, M. (2020a). Evolutionary dynamics of the POTE gene family in human and nonhuman primates. *Genes*, 11(2).

[Maggiolini et al., 2020b] Maggiolini, F. A. M., Sanders, A. D., Shew, C. J., Sulovari, A., Mao, Y., Puig, M., Catacchio, C. R., Dellino, M., Palmisano, D., Mercuri, L., Bitonto, M., Porubský, D., Cáceres, M., Eichler, E. E., Ventura, M., Dennis, M. Y., Korbel, J. O., and Antonacci, F. (2020b). Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Res.*, 30(11):1680–1693.

[Mahmoud et al., 2019] Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome Biol.*, 20(1):246.

[Marks et al., 2019] Marks, P., Garcia, S., Barrio, A. M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., Fiddes, I. T., Galvin, B., Heaton, H., Herschleb, J., Hindson, C., Holt, E., Jabara, C. B., Jett, S., Keivanfar, N., Kyriazopoulou-Panagiotopoulou, S., Lek, M., Lin, B., Lowe, A., Mahamdallie, S., Maheshwari, S., Makarewicz, T., Marshall, J., Meschi, F., O'Keefe, C. J., Ordonez, H., Patel, P., Price, A., Royall, A., Ruark, E., Seal, S., Schnall-Levin, M., Shah, P., Stafford, D., Williams, S., Wu, I., Xu, A. W., Rahman, N., MacArthur, D., and Church, D. M. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.*, 29(4):635–645.

[Marques-Bonet et al., 2009] Marques-Bonet, T., Girirajan, S., and Eichler, E. E. (2009). The origins and impact of primate segmental duplications. *Trends Genet.*, 25(10):443–454.

[McKusick, 1970] McKusick, V. A. (1970). Human genetics. *Annu. Rev. Genet.*, 4:1–46.

[McVey and Lee, 2008] McVey, M. and Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet.*, 24(11):529–538.

[Minkin and Medvedev, 2020] Minkin, I. and Medvedev, P. (2020). Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nat. Commun.*, 11(1):6327.

[Morales et al., 2015] Morales, M. E., White, T. B., Streva, V. A., DeFreece, C. B., Hedges, D. J., and Deininger, P. L. (2015). The contribution of alu elements to mutagenic DNA double-strand break repair. *PLoS Genet.*, 11(3):e1005016.

[Mostovoy et al., 2021] Mostovoy, Y., Yilmaz, F., Chow, S. K., Chu, C., Lin, C., Geiger, E. A., Meeks, N. J. L., Chatfield, K. C., Coughlin, C. R., Surti, U., Kwok, P.-Y., and Shaikh, T. H. (2021). Genomic regions associated with microdeletion/microduplication syndromes exhibit extreme diversity of structural variation. *Genetics*, 217(2).

[Nagarajan and Pop, 2013] Nagarajan, N. and Pop, M. (2013). Sequence assembly demystified. *Nat. Rev. Genet.*, 14(3):157–167.

[Navarro and Barton, 2003] Navarro, A. and Barton, N. H. (2003). Chromosomal speciation and molecular divergence–accelerated evolution in rearranged chromosomes. *Science*, 300(5617):321–324.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453.

[Nurk et al., 2022] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sović, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S., Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H., and Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588):44–53.

[Osborne et al., 2001] Osborne, L. R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L.-C., and Scherer, S. W. (2001). A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat. Genet.*, 29(3):321–325.

Bibliography

[Paigen and Petkov, 2018] Paigen, K. and Petkov, P. M. (2018). PRDM9 and its role in genetic recombination. *Trends Genet.*, 34(4):291–300.

[Pannunzio et al., 2014] Pannunzio, N. R., Li, S., Watanabe, G., and Lieber, M. R. (2014). Non-homologous end joining often uses microhomology: implications for alternative end joining. *DNA Repair*, 17:74–80.

[Park et al., 2014] Park, C.-Y., Kim, J., Kweon, J., Son, J. S., Lee, J. S., Yoo, J.-E., Cho, S.-R., Kim, J.-H., Kim, J.-S., and Kim, D.-W. (2014). Targeted inversion and reversion of the blood coagulation factor 8 gene in human iPS cells using TALENs. *Proc. Natl. Acad. Sci. U. S. A.*, 111(25):9253–9258.

[Perry et al., 2008] Perry, G. H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A. S., Hyland, C., Stone, A. C., Hurles, M. E., Tyler-Smith, C., Eichler, E. E., Carter, N. P., Lee, C., and Redon, R. (2008). Copy number variation and evolution in humans and chimpanzees. *Genome Res.*, 18(11):1698–1710.

[Pombo and Dillon, 2015] Pombo, A. and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.*, 16(4):245–257.

[Porubsky et al., 2021] Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maggiolini, F. A. M., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Human Genome Structural Variation Consortium (HGSVC), Lee, C., Antonacci, F., Steinrücken, M., Beck, C. R., Sanders, A. D., Marschall, T., Eichler, E. E., and Korbel, J. O. (2021). Haplotype-resolved inversion landscape reveals hotspots of mutational recurrence associated with genomic disorders. *bioRxiv*, page 2021.12.20.472354.

[Porubsky et al., 2022a] Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F. A., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Human Genome Structural Variation Consortium (HGSVC), Lee, C., Antonacci, F., Steinrücken, M., Beck, C. R., Sanders, A. D., Marschall, T., Eichler, E. E., and Korbel, J. O. (2022a). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, 185(11):1986–2005.e26.

[Porubsky et al., 2020] Porubsky, D., Sanders, A. D., Höps, W., Hsieh, P., Sulovari, A., Li, R., Mercuri, L., Sorensen, M., Murali, S. C., Gordon, D., Cantsilieris, S., Pollen, A. A., Ventura, M., Antonacci, F., Marschall, T., Korbel, J. O., and Eichler, E. E. (2020). Recurrent inversion toggling and great ape genome evolution. *Nat. Genet.*, 52(8):849–858.

[Porubsky et al., 2022b] Porubsky, D., Vollger, M. R., Harvey, W. T., Rozanski, A. N., Ebert, P., Hickey, G., Hasenfeld, P., Sanders, A. D., Stober, C., The Human Pangenome Reference Consortium, Korbel, J. O., Paten, B., Marschall, T., and Eichler, E. E. (2022b). Gaps and complex structurally variant loci in phased genome assemblies. *bioRxiv*, page 2022.07.06.498874.

[Puig et al., 2015] Puig, M., Castellano, D., Pantano, L., Giner-Delgado, C., Izquierdo, D., Gayà-Vidal, M., Lucas-Lledó, J. I., Esko, T., Terao, C., Matsuda, F., and Cáceres, M. (2015). Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript. *PLoS Genet.*, 11(10):e1005495.

[Puig et al., 2020] Puig, M., Lerga-Jaso, J., Giner-Delgado, C., Pacheco, S., Izquierdo, D., Delprat, A., Gayà-Vidal, M., Regan, J. F., Karlin-Neumann, G., and Cáceres, M. (2020). Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR. *Genome Res.*, 30(5):724–735.

[Rao et al., 2017] Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., Casellas, R., Lander, E. S., and Aiden, E. L. (2017). Cohesin loss eliminates all loop domains. *Cell*, 171(2):305–320.e24.

[Rao et al., 2014] Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

[Rautiainen et al., 2022] Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., Eichler, E. E., Phillippy, A. M., and Koren, S. (2022). Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv*, page 2022.06.24.497523.

[Reardon, 2021] Reardon, S. (2021). A complete human genome sequence is close: how scientists filled in the gaps. *Nature*, 594(7862):158–159.

[Remeseiro et al., 2016] Remeseiro, S., Hörnblad, A., and Spitz, F. (2016). Gene regulation during development in the light of topologically associating domains. *Wiley Interdiscip. Rev. Dev. Biol.*, 5(2):169–185.

[Rosendahl et al., 2018] Rosendahl, J., Kirsten, H., Hegyi, E., Kovacs, P., Weiss, F. U., Laumen, H., Lichtner, P., Ruffert, C., Chen, J.-M., Masson, E., Beer, S., Zimmer, C., Seltsam, K., Algül, H., Bühler, F., Bruno, M. J., Bugert, P., Burkhardt, R., Cavestro, G. M., Cichoz-Lach, H., Farré, A., Frank, J., Gambaro, G., Gimpfl, S., Grallert, H., Griesmann, H., Grützmann, R., Hellerbrand, C., Hegyi, P., Hollenbach, M., Iordache, S., Jurkowska, G., Keim, V., Kiefer, F., Krug, S., Landt, O., Leo, M. D., Lerch, M. M., Lévy, P., Löffler, M., Löhr, M., Ludwig, M., Macek, M., Malats, N., Malecka-Panas, E., Malerba, G., Mann, K., Mayerle, J., Mohr, S., Te Morsche, R. H. M., Motyka, M., Mueller, S., Müller, T., Nöthen, M. M., Pedrazzoli, S., Pereira, S. P., Peters, A., Pfützer, R., Real, F. X., Rebours, V., Ridinger, M., Rietschel, M., Rösmann, E., Saftoiu, A., Schneider, A., Schulz, H.-U., Soranzo, N., Soyka, M., Simon, P., Skipworth, J., Stickel, F., Strauch, K., Stumvoll, M., Testoni, P. A., Tönjes, A., Werner, L., Werner, J., Wodarz, N., Ziegler, M., Masamune, A., Mössner, J., Férec, C., Michl, P., P H Drenth, J., Witt, H., Scholz, M., Sahin-Tóth, M., and all members of the PanEuropean Working group on ACP (2018). Genome-wide association study identifies inversion in the CTRB1-CTRB2 locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis. *Gut*, 67(10):1855–1863.

[Rowley and Corces, 2018] Rowley, M. J. and Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nat. Rev. Genet.*, 19(12):789–800.

[Said et al., 2018] Said, I., Byrne, A., Serrano, V., Cardeno, C., Vollmers, C., and Corbett-Detig, R. (2018). Linked genetic variation and not genome structure causes widespread differential expression associated with chromosomal inversions. *Proc. Natl. Acad. Sci. U. S. A.*, 115(21):5492–5497.

[Sanchis-Juan et al., 2018] Sanchis-Juan, A., Stephens, J., French, C. E., Gleadall, N., Mégy, K., Penkett, C., Shamardina, O., Stirrups, K., Delon, I., Dewhurst, E., Dolling, H., Erwood, M., Grozeva, D., Stefanucci, L., Arno, G., Webster, A. R., Cole, T., Austin, T., Branco, R. G., Ouwehand, W. H., Raymond, F. L., and Carss, K. J. (2018). Complex structural variants in mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med.*, 10(1):95.

Bibliography

[Sanders et al., 2017] Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J., and Lansdorp, P. M. (2017). Single-cell template strand sequencing by strand-seq enables the characterization of individual homologs. *Nat. Protoc.*, 12(6):1151–1176.

[Sanders et al., 2016] Sanders, A. D., Hills, M., Porubský, D., Guryev, V., Falconer, E., and Lansdorp, P. M. (2016). Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.*, 26(11):1575–1587.

[Sanders et al., 2020] Sanders, A. D., Meiers, S., Ghareghani, M., Porubsky, D., Jeong, H., van Vliet, M. A. C. C., Rausch, T., Richter-Pechańska, P., Kunz, J. B., Jenni, S., Bolognini, D., Longo, G. M. C., Raeder, B., Kinanen, V., Zimmermann, J., Benes, V., Schrappe, M., Mardin, B. R., Kulozik, A. E., Bornhauser, B., Bourquin, J.-P., Marschall, T., and Korbel, J. O. (2020). Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat. Biotechnol.*, 38(3):343–354.

[Schöpflin et al., 2022] Schöpflin, R., Melo, U. S., Moeinzadeh, H., Heller, D., Laupert, V., Hertzberg, J., Holtgrewe, M., Alavi, N., Klever, M.-K., Jungnitsch, J., Comak, E., Türkmen, S., Horn, D., Duffourd, Y., Faivre, L., Callier, P., Sanlaville, D., Zuffardi, O., Tenconi, R., Kurtas, N. E., Giglio, S., Prager, B., Latos-Bielenska, A., Vogel, I., Bugge, M., Tommerup, N., Spielmann, M., Vitobello, A., Kalscheuer, V. M., Vingron, M., and Mundlos, S. (2022). Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. *Nat. Commun.*, 13(1):6470.

[Sedlazeck et al., 2018] Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, 15(6):461–468.

[Sharp et al., 2005] Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segraves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., and Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.*, 77(1):78–88.

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

[Stankiewicz and Lupski, 2002] Stankiewicz, P. and Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet.*, 18(2):74–82.

[Stefansson et al., 2005] Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., Olafsdottir, A., Cazier, J.-B., Kristjansson, K., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., Kong, A., and Stefansson, K. (2005). A common inversion under selection in europeans. *Nat. Genet.*, 37(2):129–137.

[Steinmann et al., 2007] Steinmann, K., Cooper, D. N., Kluwe, L., Chuzhanova, N. A., Senger, C., Serra, E., Lazaro, C., Gilaberte, M., Wimmer, K., Mautner, V.-F., and Kehrer-Sawatzki, H. (2007). Type 2 NF1 deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination. *Am. J. Hum. Genet.*, 81(6):1201–1220.

[Stevison et al., 2011] Stevison, L. S., Hoehn, K. B., and Noor, M. A. F. (2011). Effects of inversions on within- and Between-Species recombination and divergence. *Genome Biol. Evol.*, 3:830–841.

[Sturtevant, 1917] Sturtevant, A. H. (1917). Genetic factors affecting the strength of linkage in drosophila. *Proc. Natl. Acad. Sci. U. S. A.*, 3(9):555–558.

[Sturtevant, 1921] Sturtevant, A. H. (1921). A case of rearrangement of genes in drosophila. *Proc. Natl. Acad. Sci. U. S. A.*, 7(8):235–237.

[Sudmant et al., 2015] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Mu, X. J., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., and Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81.

[Suntsova and Buzdin, 2020] Suntsova, M. V. and Buzdin, A. A. (2020). Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species.

[Suwinski et al., 2019] Suwinski, P., Ong, C., Ling, M. H. T., Poh, Y. M., Khan, A. M., and Ong, H. S. (2019). Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Front. Genet.*, 10:49.

[Szustakowski et al., 2021] Szustakowski, J. D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P. G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., Katrina Loomis, A., Mikkilineni, R., Noh, H. J., Wadhawan, S., Bai, X., Hawes, A., Krasheninina, O., Ulloa, R., Lopez, A. E., Smith, E. N., Waring, J. F., Whelan, C. D., Tsai, E. A., Overton, J. D., Salerno, W. J., Jacob, H., Szalma, S., Runz, H., Hinkle, G., Nioi, P., Petrovski, S., Miller, M. R., Baras, A., Mitnaul, L. J., Reid, J. G., and UKB-ESC Research Team (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK biobank. *Nat. Genet.*, 53(7):942–948.

[Tatton-Brown et al., 2005] Tatton-Brown, K., Douglas, J., Coleman, K., Baujat, G., Chandler, K., Clarke, A., Collins, A., Davies, S., Faravelli, F., Firth, H., Garrett, C., Hughes, H., Kerr, B., Liebelt, J., Reardon, W., Schaefer, G. B., Splitt, M., Temple, I. K., Waggoner, D., Weaver, D. D., Wilson, L., Cole, T., Cormier-Daire, V., Irrthum, A., Rahman, N., and Childhood Overgrowth Collaboration (2005). Multiple mechanisms are implicated in the generation of 5q35 microdeletions in sotos syndrome. *J. Med. Genet.*, 42(4):307–313.

[The "All of Us" Research Program, 2019] The "All of Us" Research Program (2019). The "all of us" research program. *N. Engl. J. Med.*, 381(7):668–676.

[Tian et al., 2018] Tian, S., Yan, H., Klee, E. W., Kalmbach, M., and Slager, S. L. (2018). Comparative analysis of de novo assemblers for variation discovery in personal genomes. *Brief. Bioinform.*, 19(5):893–904.

[Trost et al., 2021] Trost, B., Loureiro, L. O., and Scherer, S. W. (2021). Discovery of genomic variation across a generation. *Hum. Mol. Genet.*, 30(R2):R174–R186.

Bibliography

[Tukiainen et al., 2017] Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B. B., Castel, S. E., Karczewski, K. J., Aguet, F., Byrnes, A., Lappalainen, T., Regev, A., Ardlie, K. G., Hacohen, N., and MacArthur, D. G. (2017). Landscape of X chromosome inactivation across human tissues. *Nature*, 550(7675):244–248.

[Valton and Dekker, 2016] Valton, A.-L. and Dekker, J. (2016). TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.*, 36:34–40.

[Ventura et al., 2001] Ventura, M., Archidiacono, N., and Rocchi, M. (2001). Centromere emergence in evolution. *Genome Res.*, 11(4):595–599.

[Vervoort et al., 2021] Vervoort, L., Dierckxsens, N., Pereboom, Z., Capozzi, O., Rocchi, M., Shaikh, T. H., and Vermeesch, J. R. (2021). 22q11.2 low copy repeats expanded in the human lineage. *Front. Genet.*, 12:706641.

[Vollger et al., 2022] Vollger, M. R., Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K. M., Lewis, A. P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K. H., Phillippy, A. M., Timp, W., Ventura, M., and Eichler, E. E. (2022). Segmental duplications and their variation in a complete human genome. *Science*, 376(6588):eabj6965.

[Wang et al., 2020] Wang, Z.-Y., Leushkin, E., Liechti, A., Ovchinnikova, S., Mößinger, K., Brüning, T., Rummel, C., Grützner, F., Cardoso-Moreira, M., Janich, P., Gatfield, D., Diagouraga, B., de Massy, B., Gill, M. E., Peters, A. H. F. M., Anders, S., and Kaessmann, H. (2020). Transcriptome and translatome co-evolution in mammals. *Nature*, 588(7839):642–647.

[Weischenfeldt et al., 2013] Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, 14(2):125–138.

[Wellenreuther and Bernatchez, 2018] Wellenreuther, M. and Bernatchez, L. (2018). Eco-Evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.*, 33(6):427–440.

[Wellenreuther et al., 2019] Wellenreuther, M., Mérot, C., Berdan, E., and Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.*, 28(6):1203–1209.

[Wutz et al., 2017] Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M. J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P., and Peters, J.-M. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.*, 36(24):3573–3599.

[Yilmaz et al., 2021] Yilmaz, F., Gurusamy, U., Mosley, T. J., Mostovoy, Y., Shaikh, T. H., Zwick, M. E., Kwok, P.-Y., Lee, C., and Mulle, J. G. (2021). Multi-modal investigation of the schizophrenia-associated 3q29 genomic interval reveals global genetic diversity with unique haplotypes and segments that increase the risk for non-allelic homologous recombination. *bioRxiv*.

[Yuan et al., 2015] Yuan, B., Liu, P., Gupta, A., Beck, C. R., Tejomurtula, A., Campbell, I. M., Gambin, T., Simmons, A. D., Withers, M. A., Harris, R. A., Rogers, J., Schwartz, D. C., and Lupski, J. R. (2015). Comparative genomic analyses of the human NPHP1 locus reveal complex genomic architecture and its regional evolution in primates. *PLoS Genet.*, 11(12):e1005686.

[Yunis and Prakash, 1982] Yunis, J. J. and Prakash, O. (1982). The origin of man: a chromosomal pictorial legacy. *Science*, 215(4539):1525–1530.

[Zhang et al., 2010] Zhang, F., Potocki, L., Sampson, J. B., Liu, P., Sanchez-Valle, A., Robbins-Furman, P., Navarro, A. D., Wheeler, P. G., Spence, J. E., Brasington, C. K., Withers, M. A., and Lupski, J. R. (2010). Identification of uncommon recurrent Potocki-Lupski syndrome-associated duplications and the distribution of rearrangement types and mechanisms in PTLS. *Am. J. Hum. Genet.*, 86(3):462–470.

[Zhang et al., 2021] Zhang, L., Reifová, R., Halenková, Z., and Gompert, Z. (2021). How important are structural variants for speciation? *Genes*, 12(7).

[Zhao et al., 2017] Zhao, Q., Ma, D., Vasseur, L., and You, M. (2017). Segmental duplications: evolution and impact among the current lepidoptera genomes. *BMC Evol. Biol.*, 17(1):161.

[Zheng-Bradley and Flicek, 2017] Zheng-Bradley, X. and Flicek, P. (2017). Applications of the 1000 genomes project resources. *Brief. Funct. Genomics*, 16(3):163–170.

[Zody et al., 2008] Zody, M. C., Jiang, Z., Fung, H.-C., Antonacci, F., Hillier, L. W., Cardone, M. F., Graves, T. A., Kidd, J. M., Cheng, Z., Abouelleil, A., Chen, L., Wallis, J., Glasscock, J., Wilson, R. K., Reily, A. D., Duckworth, J., Ventura, M., Hardy, J., Warren, W. C., and Eichler, E. E. (2008). Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.*, 40(9):1076–1083.