Inference and analysis of recurring genomic variation in human populations

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Hufsah Ashraf
aus Chakwal

Düsseldorf, April, 2025

aus dem Institut für Medizinische Biometrie und Bioinformatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Tobias Marschall

2. Prof. Dr. Alexander Dilthey

Tag der mündlichen Prüfung: 12.12.2025

Statement

I declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the “Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf”.

Düsseldorf, April 2025

Hufsah Ashraf

Abstract

Genomic variation is fundamental to understanding human biology, from population-level diversity to disease susceptibility. Recent advances in both sequencing technologies and computational methods have markedly improved our ability to detect and interpret variation across human genomes, yet challenges persist. This dissertation presents a series of computational approaches aimed at improving the identification, characterization, and analysis of genomic variants, with a particular focus on inversions and their recurrence across human populations.

The first part of this thesis introduces *k*-merald, a method developed to improve allele detection accuracy while using error-prone sequencing data. By leveraging platform-specific error profiles, *k*-merald enhances alignment reliability and substantially reduces genotyping error rates, especially under low-coverage conditions.

The second part presents ArbiGent, a tool for genotyping inversions and copy number variants using Strand-seq data. ArbiGent corrects for alignment artifacts by normalizing Strand-seq read counts based on locus-specific mappability, improving genotyping reliability in repetitive regions. Its role in generating a high-confidence inversion callset, as part of a project under the Human Genome Structural Variation Consortium (HGSVC), is also presented in this part. This callset revealed that inversions affect a larger portion of the genome than other variant types and are enriched in highly repetitive and disease-associated regions of the genome.

The third part introduces the new concept of toggling-indicating SNPs (tiSNPs) and describes an inversion recurrence detection approach that analyzes allelic patterns of within-inversion SNPs across haplotypes to distinguish between single and recurrent inversion events. This approach, supported by orthogonal validation, revealed widespread inversion recurrence, including events that overlap known disease-associated loci.

The fourth part introduces Pivot, a tool for detecting recurrent inversions within a graph-based pangenomic framework, eliminating reference bias. By incorporating all within-inversion variants into the analysis, Pivot displays high sensitivity for recurrence detection. Applied to diverse haplotype panels from the HGSVC and the Human Pangenome Reference Consortium (HPRC), Pivot detected novel recurrent inversions and revealed recurrence evidence in multiple disease-relevant regions.

The final part of this dissertation extends the investigation of recurrence beyond inversions and presents an approach to detect recurrent deletions. This analysis identified several

candidate recurrent deletions in a cohort of 1,019 samples from the 1000 Genomes Project (1KGP), flanked by transposable elements, implicating non-allelic homologous recombination as a potential mechanism.

Kurzfassung

Genomische Variation ist von grundlegender Bedeutung für das Verständnis der menschlichen Biologie, von der Vielfalt auf Bevölkerungsebene bis zur Prädisposition von Krankheiten. Jüngste Fortschritte sowohl bei den Sequenzierungstechnologien als auch bei den Berechnungsmethoden haben unsere Fähigkeit, Variationen im menschlichen Genom zu erkennen und zu interpretieren, deutlich verbessert, doch es gibt immer noch Herausforderungen. In dieser Dissertation werden eine Reihe von Berechnungsansätzen vorgestellt, die darauf abzielen, die Identifizierung, Charakterisierung und Analyse genomischer Varianten zu verbessern, wobei ein besonderer Schwerpunkt auf Inversionen und deren Rekurrenz in menschlichen Populationen liegt.

Im ersten Teil dieser Arbeit wird *k*-merald vorgestellt, eine Methode, die zur Verbesserung der Erkennungsgenauigkeit von Allelen bei der Verwendung von fehleranfälligen Sequenzierungsdaten entwickelt wurde. Durch die Nutzung von plattformspezifischen Fehlerprofilen erhöht *k*-merald die Zuverlässigkeit des Alignments und reduziert die Fehlerraten bei der Genotypisierung erheblich, insbesondere unter Bedingungen mit geringer Abdeckung.

Im zweiten Teil wird ArbiGent vorgestellt, ein Tool zur Genotypisierung von Inversionen und Kopienzahlvarianten unter Verwendung von Strand-seq-Daten. ArbiGent korrigiert Alignment-Artefakte, indem es die Anzahl der Strand-seq *reads* auf der Grundlage lokusspezifischer *Mappability* normalisiert und so die Zuverlässigkeit der Genotypisierung in repetitiven Regionen verbessert. Die Rolle von ArbiGent bei der Generierung eines Inversions-Callsets mit hoher Zuverlässigkeit aus *Human Genome Structural Variation Consortium (HGSVC)* wird ebenfalls in diesem Teil vorgestellt. Dieses *Callset* hat gezeigt, dass Inversionen einen größeren Teil des Genoms betreffen als andere Variantenarten und in stark repetitiven und krankheitsassoziierten Regionen des Genoms angereichert sind.

Im dritten Teil wird ein auf *tiSNPs (toggling-indicating SNPs)* basierender Ansatz vorgestellt, der allelische Muster von SNPs innerhalb von Inversionen über Haplotypen hinweg analysiert, um zwischen einzelnen und rekurrenten Inversionsereignissen zu unterscheiden. Dieser Ansatz, der durch eine unabhängige Methode validiert wird, zeigte eine weit verbreitete Rekurrenz von Inversionen, einschließlich Ereignissen, die sich mit bekannten krankheitsassoziierten *Loci* überschneiden.

Im vierten Teil wird Pivot vorgestellt, ein Tool zur Erkennung von rekurrenten Inversionen innerhalb eines graphbasierten pangenomischen Rahmens, das *reference bias* eliminiert. Durch die Einbeziehung aller Varianten innerhalb der Inversion in die Analyse zeigt

Pivot eine hohe Sensitivität bei der Erkennung von Rekurrenz. Bei der Anwendung auf verschiedene Haplotyp-Panels von HGSVC und *Human Pangenome Reference Consortium (HPRC)* entdeckte Pivot neuartige rekurrente Inversionen und zeigte Hinweise auf ein Rekurrenz in mehreren krankheitsrelevanten Regionen.

Im letzten Teil dieser Dissertation wird die Untersuchung von Rekurrenz über Inversionen hinaus erweitert und ein Ansatz zur Erkennung rekurrenter Deletionen vorgestellt. Diese Analyse identifizierte mehrere Kandidaten für rekurrente Deletionen in einer Kohorte von 1.019 Proben aus dem *1000 Genomes (1KG) Project*, die von *transposable elements* flankiert werden, was auf nicht-allelische homologe Rekombination als möglichen Mechanismus hindeutet.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Tobias Marschall, for his support, insightful guidance, and constant encouragement throughout my doctoral journey. His mentorship has been instrumental in both the development and completion of this work. I am especially grateful for his willingness to take on the huge administrative and bureaucratic responsibilities that come with supervising a foreign doctoral student.

I am also thankful to my colleagues for creating a positive working environment, which made it possible to endure the challenging and demanding times of my PhD. The lunch discussions, debates over very trivial things, and celebration of small victories made the journey feel less daunting and far more enjoyable.

Big thanks to Wolfram Höps, with whom I spent countless hours developing a tool that often made us question our life choices, but in the end, we made it through.

I am also sincerely grateful to Jana Ebler, Hugo Magalhães, Peter Ebert, Mir Henglin, Rebecca Serra Mari, Vithusan Suppiyar and Samarendra Pani, who generously took the time to proofread parts of this thesis.

Above all, I am deeply grateful to my parents for their constant support and encouragement. Their love and sacrifices have played a vital role in helping me reach this milestone. This work is as much theirs as it is mine.

Contents

List of Figures

List of Tables

Introduction

The human genome—comprising roughly three billion base pairs per haplotype and organized into chromosomes—is structurally complex. Just like many fields of science, the more we delve into the human genome, the more complexity we uncover. The publication of the first drafts of the human reference genome in 2001 [1, 2] marked a significant milestone in genomics. This foundational resource opened new avenues for understanding human ancestry, evolution, and disease mechanisms. Yet, it also unveiled a vast array of new questions. One of these lingering questions stemmed from the fact that some parts of the genome were still missing or unresolved. Over the following years, numerous efforts were made to close these gaps, culminating in 2022 when the Telomere-to-Telomere (T2T) consortium delivered a complete human reference genome [3]. This T2T reference not only filled the remaining 8% gap in the genome but also corrected previous inaccuracies and resolved some of the most complex regions.

Human genomes are not identical; they vary significantly across individuals, contributing to the diversity of traits, disease susceptibility, and evolutionary adaptation [4, 5]. This variation spans from single-nucleotide differences to large, complex structural alterations across the genome [4, 5]. The identification and characterization of this variation largely rely on the alignments of sequencing reads to a reference genome. Advances in sequencing technologies have greatly improved the resolution and accuracy of these steps. Short-read sequencing platforms such as Illumina offer high throughput and accuracy [6], making them highly effective for detecting single-nucleotide polymorphisms (SNPs) and short insertions and deletions (indels), though they struggle to resolve complex structural variation and repetitive regions [7]. Long-read technologies, including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), span larger genomic regions and are better suited for detecting structural variants and phasing haplotypes, but come with higher sequencing error rates impacting the accuracy of downstream analyses [7–9]. Integrating data from both short- and long-read sequencing is increasingly common, providing a more comprehensive and accurate landscape of human genomic variation [10–15].

Among the various types of genomic variation, structural variation (SV) represents a significant source of genetic diversity, involving large-scale changes such as deletions, duplications, insertions, inversions, and translocations [16–18]. These structural alterations can profoundly impact gene function and regulation, can influence normal biological processes, and can play a vital role in disease development [19, 20]. SVs can be recurrent,

often arising through non-allelic homologous recombination (NAHR) and other mutational mechanisms [21]. This recurrent SV formation has been implicated in the development of numerous genetic disorders and phenotypic traits [22–25]. Among recurrent SVs, inversions represent a particularly intriguing class. Although copy neutral, inversions can disrupt gene function or regulatory architecture and may predispose to further disease-associated structural rearrangements in the genome [26–34]. The detection of inversions has historically been challenging due to the limitations of short-read sequencing technologies, but advances in long-read sequencing and strand-specific sequencing technologies have greatly improved our ability to resolve inversion breakpoints and study their functional consequences [11–14].

The availability of a complete human genome [3] has played a crucial role in enhancing the overall characterization of genomic variation. However, considering a single genome as "reference" for variant discovery introduces biases, as it cannot capture the huge amount of genomic diversity that exists across populations. This limitation hinders the accurate detection and interpretation of variants, especially those specific to underrepresented populations. An effort to mitigate this problem was carried out by the Human Pangenome Reference Consortium (HPRC), which led to the release of the first draft of a human pangenome reference in 2023 [35].A pangenome provides a more detailed and unbiased view of genomic diversity by incorporating the genetic sequences of individuals from diverse populations.

Outline

Chapter 1 lays the groundwork by providing essential background information and explaining key biological concepts relevant to the content presented in the following chapters.

Chapter 2 describes *k*-merald—a variant allele detection approach that builds platform-specific sequencing error profiles using *k*-mer frequencies observed in aligned sequencing reads. It employs an alignment algorithm that uses these profiles to calculate alignment costs during re-alignment of reads to reference and alternative alleles at candidate variant sites. By incorporating platform-specific error profiles, *k*-merald is able to distinguish true variants from sequencing errors, improving allele detection accuracy, particularly for SNPs and indels. Its effectiveness is especially notable while using long-read sequencing data, where high error rates often confound traditional variant allele detection methods. This work was published in *Bioinformatics Advances* [36].

Chapter 3 presents ArbiGent—a tool for genotyping predefined inversion and copy number variant loci using Strand-seq data. It builds on the framework of MosaiCatcher [37], which uses a Bayesian model to compute genotype likelihoods across single-cell Strand-seq libraries, by analyzing strand-state inheritance patterns. ArbiGent converts these single-cell genotype likelihoods to a consensus sample genotype. To improve genotyping accuracy, ArbiGent normalizes Strand-seq read counts using regional mappability, accounting for alignment biases in repetitive regions. ArbiGent was introduced in the Human Genome Structural

Variation Consortium (HGSVC) study, published in *Science* [12]. Its performance validation and the inversion callset generation, discussed in this chapter, were later published in *Cell* [13].

Chapter 4 explores the interesting yet under-studied phenomenon of inversion recurrence, highlighting the toggling-indicating SNPs (tiSNPs)-based approach for inversion recurrence detection. This method utilizes haplotype-resolved Strand-seq reads to identify within-inversion SNPs that exhibit allele patterns inconsistent with a single inversion origin. By evaluating each SNP independently, the tiSNPs-based approach is largely unaffected by the effects of recombination. The recurrence status of an inversion locus is determined by aggregating evidence from SNPs observed at that locus. The work discussed in this chapter was published in *Cell* [13].

Chapter 5 delves further into the phenomenon of inversion recurrence and extends it into the domain of pangenomics through the introduction of Pivot—a tool designed to analyze inversion recurrence using pangenome graphs. Pivot adopts the core theoretical framework of the tiSNPs-based approach but enhances it by evaluating all variants within the inverted region and its flanks, thereby improving sensitivity in detecting recurrence signals. Operating within a pangenomic context, Pivot also mitigates reference bias. Additionally, it does not require any variant genotyping information, instead utilizing only the pangenome graph as input.

Finally, Chapter 6 shifts the focus to recurrent deletions. To explore these events, we devised a strategy inspired by the tiSNPs-based inversion recurrence detection method, adapting it to identify recurrence-indicating SNPs within the regions flanking deletions. To complement this analysis, we applied phylogenetic validation using hierarchical clustering to determine the recurrence status of each locus under investigation. This approach was applied to a cohort of 1,019 samples from the 1000 Genomes Project (1KGP), and the work presented in this chapter forms part of a study that has been provisionally accepted for publication in *Nature* [38].

The publications underlying this thesis are listed below. First authorship is denoted by *.

- **H. Ashraf***, J. Ebler, and T. Marschall. Allele detection using k-mer-based sequencing error profiles. *Bioinformatics Advances*, 3(1):vbad149, 2023.

- P. Ebert*, P. A. Audano*, Q. Zhu*, B. Rodriguez-Martin*, D. Porubsky, M. J. Bonder, A. Sulovari, J. Ebler, W. Zhou, R. Serra Mari, F. Yilmaz, X. Zhao, P. Hsieh, J. Lee, S. Kumar, J. Lin, T. Rausch, Y. Chen, J. Ren, M. Santamarina, W. Höps, **H. Ashraf**, N. T. Chuang, X. Yang, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537):eabf7117, 2021.

- D. Porubsky*, W. Höps*, **H. Ashraf***, P. Hsieh, B. Rodriguez-Martin, F. Yilmaz, J. Ebler, P. Hallast, F. A. M. Maggiolini, W. T. Harvey, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*,

185(11):1986–2005, 2022.

- D. Porubsky*, W. T. Harvey, A. N. Rozanski, J. Ebler, W. Höps, **H. Ashraf**, P. Hasenfeld, Human Pangenome Reference Consortium (HPRC), Human Genome Structural Variation Consortium (HGSVC), B. Paten, et al. Inversion polymorphism in a complete human genome assembly. *Genome Biology*, 24(1):100, 2023.

- G. A. Logsdon*, P. Ebert*, P. A. Audano*, M. Loftus, D. Porubsky, J. Ebler, F. Yilmaz, P. Hallast, T Prodanov, D. Yoo, C. A. Paisie, W. T. Harvey, X. Zhao, G. V. Martino, M. Henglin, K. M. Munson, K. Rabbani, C. Chin, B. Gu, **H. Ashraf**, O. Austine-Orimoloye, et al. Complex genetic variation in nearly complete human genomes. *bioRxiv*, 2024. (This article has been provisionally accepted for publication in *Nature*.)

- S. Schloissnig*, S. Pani*, B. Rodriguez-Martin, J. Ebler, C. Hain, V. Tsapalou, A. Söylev, P. Hüther, **H. Ashraf**, T. Prodanov, et al. Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 genomes project. *bioRxiv*, pages 2024–04, 2024. (This article has been provisionally accepted for publication in *Nature*.)

Chapter 1

Background

This chapter provides essential background information and describes key biological concepts relevant to the topics that will be explored in detail in the following chapters.

1.1 Genomic variation

The DNA sequences of two individuals are not identical, and the level of genetic variation tends to be higher when comparing individuals from different populations. There are multiple sources that can give rise to this variation, including mutations, genetic recombination, genetic drift and so on. A genomic position exhibiting DNA sequence variation is termed as a "variant" position while the multiple versions of DNA sequence observed at a variant position are termed as "alleles". There can be multiple types of genetic variants (Figure 1.1), for example:

- **SNPs:** single nucleotide polymorphisms characterized by alterations of single nucleotide bases in the genome. SNPs are the most frequently observed type of genetic variation, for example, a human genome on average carries a SNP once every 1000^{th} base pair, thus affecting 3-4 million base pairs of the genome [39].
- **Indels:** a collective term used to represent **in**sertions and **del**etions of DNA sequence.
- **Inversions:** reinsertion of a broken segment of DNA at the same position but in reverse orientation.
- **Translocations:** reinsertion of a broken segment of DNA at a different place on the same chromosome (intrachromosomal) or on a different chromosome (interchromosomal).

Genomic alterations that result in changes to the number of copies of specific DNA segments, such as insertions, deletions, and duplications, are collectively referred to as "copy number variations" (CNVs). All variation subtypes affecting at least 50 bp of nucleotide sequence are broadly classified as "structural variants" (SVs).

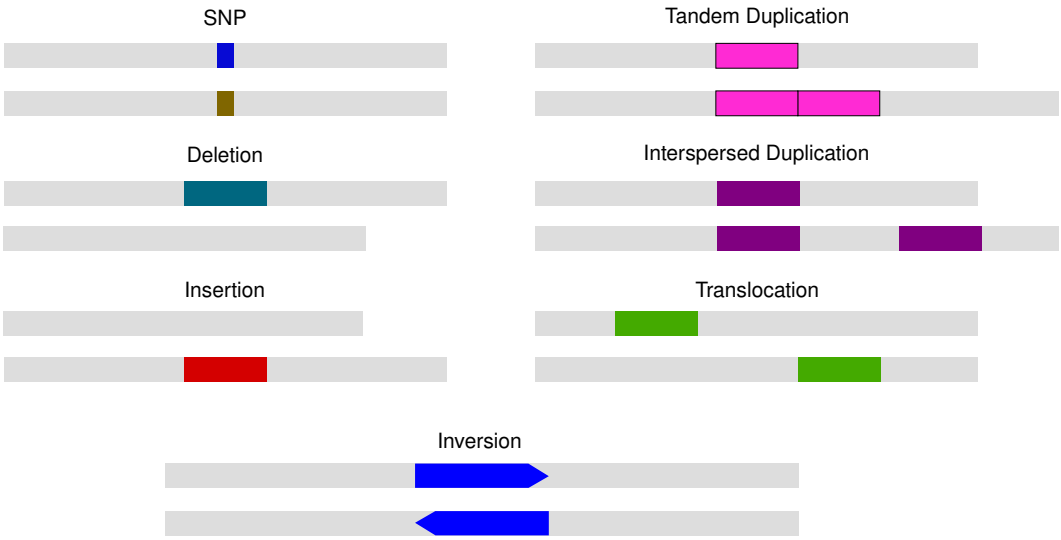


Figure 1.1: Genetic variation.

1.2 Variant calling, genotyping, and phasing

When studying genomic variation, the initial step is the *de novo* identification of varying regions of the genome—a process known as "variant calling". This step pinpoints both the variant positions and the respective alleles, with respect to a reference genome. The allele carried by the reference genome at a particular variant position is referred to as the "reference allele" while those differing from the reference allele are termed as "alternative alleles". Typically, alleles are denoted numerically: "0" represents the reference allele, while "1,2,3...n" are used to represent each of the alternative alleles.

Following variant calling, the next step is to determine the specific alleles carried by an individual on each of its chromosomal copies, a process known as "genotyping". The resulting genotypes are usually represented with allele notations separated by a slash, for example, a diploid organism can have "0/0", "0/1", or "1/1" genotype for a bi-allelic variant.

Phasing, or haplotype estimation, involves inferring which alleles are inherited together from each parent based on an individual's genotype. Phased genotypes use the same numeric allele notation but employ a pipe symbol ("|") instead of a slash to indicate the separation between maternal and paternal alleles, for example, "0|1".

1.3 Genome sequencing, alignment and associated challenges

Genome sequencing is the process of determining the nucleotide sequence of an individual. This typically involves breaking the genome into smaller fragments, sequencing each fragment individually, and storing the resulting sequences as "sequencing reads". These reads form the foundation for a range of downstream analyses, such as genotyping and phasing, and are also used to reconstruct the original DNA sequence in a process known as "genome assembly". Initially, sequencing reads are simply sequence fragments with no

contextual information about their place of origin in the genome. This crucial information is typically retrieved using "sequence alignment" which involves identifying regions of similarity between sequences, either globally (aligning entire sequences end-to-end) or locally (matching subsequences).

The properties of sequencing reads differ significantly depending on the sequencing technology used. Second-generation sequencing platforms, such as Illumina, produce short (150–300 bp), paired-end reads (generated from sequencing DNA fragments from both ends) with very high accuracy (approximately 99.9%) [6]. While this high accuracy improves the reliability of downstream analyses like genotyping, the length of sequencing reads is equally important. Genome sequences are highly repetitive, for example, 50% of the human genome is comprised of repeat sequences [1]. Short reads often fail to span long repetitive regions, making accurate alignment difficult and reducing the reliability of any analysis dependent on these alignments. To overcome this limitation, third-generation sequencing technologies were developed. Platforms such as Pacific Biosciences (PacBio) produce long reads in the range of 15–20 kbp [40], while Oxford Nanopore Technologies (ONT) generate reads ranging from 10–100 kbp, with ONT Ultra-long (ONT-UL) reads exceeding 100 kbp [6, 8]. These reads span much longer regions making them more informative particularly for complex and repetitive regions of the genome. PacBio HiFi sequencing technology has a very high accuracy (about 99.9%) [9]. However, the most commonly used long-read sequencing technologies typically have higher error rates compared to their short-read counterparts. For instance, PacBio's Continuous Long Reads (CLR) exhibit error rates between 8–15%, while ONT reads have error rates ranging from 2–13% [8].

The presence of both genetic variants and sequencing errors complicates the task of sequence alignment. To address this, various approximate string matching algorithms have been developed—many of which are based on the Needleman-Wunsch algorithm for global alignment [41] and the Smith-Waterman algorithm for local alignment [42]. Most of the conventional read aligner still operate in a single, linear reference space, ignoring any allelic variants. This limitation introduces reference bias in the alignments [43], which, combined with sequencing errors, makes sequence alignments in the variant regions of the genome unreliable for the correct allele determination.

One approach to mitigate the issue of reference bias, as employed by WhatsHap [44–46], is the local re-alignment of reads. WhatsHap is a widely used tool for phasing genomic variants, primarily SNPs and indels. It extracts the read sequence within a window around the variant position and independently re-aligns it to the reference and each of the alternative allele sequences (Figure 1.2). Edit-distance-based alignment costs are then used to determine the correct allele carried by the read. While WhatsHap effectively reduces reference bias and thereby enhances allele detection accuracy, it does not account for sequencing errors which can lead to wrong allele detection as shown in Figure 1.3. This open problem served as the motivation for developing an improved allele detection method, k-merald, discussed in detail in Chapter 2. k-merald extends WhatsHap's re-alignment framework by

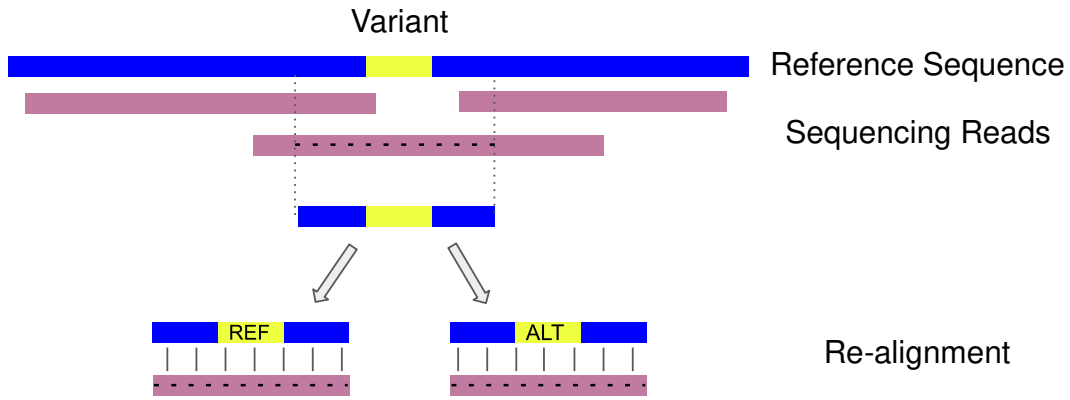


Figure 1.2: Read re-alignment minimizes reference bias.

incorporating additional modules specifically designed to correct for the technology-specific sequencing errors. By addressing both reference bias and sequencing errors, k-merald improves allele detection accuracy, ultimately leading to better genotyping results for both SNPs and indels.

1.4 Structural variation

Structural variants—mutations affecting $\geq$50 bp of nucleotide sequence—broadly include large-scale CNVs like duplications as well as copy-neutral inversions and translocations [24]. Insertions and deletions affecting long stretches of nucleotide sequence are also typically categorized as SVs [24]. For the longest time, single SNPs, being the most frequent variants were considered to be the primary drivers of genomic diversity [5]. Combined with short indels, they encompass $>99\%$ of the variant sites observed in human genomes [18]. The remaining variation is typically attributed to structural variants, which although much less frequent than SNPs, affect about $10\times$ more nucleotides [11, 39]. Based on the global reference of genetic variation released by the 1000 Genomes Project (1KGP) in 2015 [5], a typical human genome contains an estimated 2,100 to 2,500 SVs, predominantly including large deletions, CNVs, mobile element insertions and inversions, affecting about 20 million base pairs of nucleotide sequence. Several studies have highlighted the significant role of SVs in shaping genomic diversity across human populations [5, 10–12, 15]. SVs have also been implicated as key drivers of evolution and speciation events [47–49], and they play a crucial role in disease mechanisms. [17, 50–52].

1.4.1 Genomic architecture and structural variation-mediating mechanisms

Structural variations can arise through a range of mutational mechanisms involving recombination, replication, or DNA repair. The genomic architecture of the variant locus—especially its repeat content—plays a crucial role in determining both the type of SV and the underlying mutational mechanism. Repetitive sequences such as mobile elements, simple

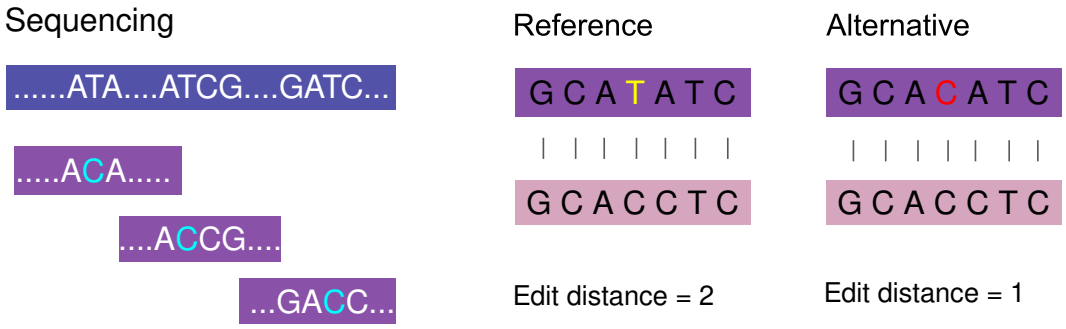


Figure 1.3: Systematic sequencing errors vs alternative alleles.

repeats, tandem repeats, and low-copy repeats (LCRs) comprise approximately 50% of the human genome [1]. Among these, LCRs, particularly segmental duplications (SDs), have been identified as major contributors to human genomic diversity [53]. SDs are DNA segments $\geq$1 kbp in length with $\geq$90% sequence identity [54], constituting about 7% of the human genome [55]. SDs are believed to have arisen from the duplication of genomic segments, giving rise to paralogous regions [23].

Several studies have shown that SDs emerged during primate speciation and continue to evolve, contributing to their complex and dynamic architecture [23]. Vollger et al. observed that SD-rich regions show approximately 10-fold enrichment for CNVs [55]. Notably, large SDs ($>$10 kbp) frequently overlap with regions known for disease-associated genomic rearrangements [54]. Concurrently, studies suggest that genomic regions flanked by SDs are highly unstable and prone to rearrangements [23, 24].

The most commonly implicated DNA rearrangement mechanism in SV formation, particularly in regions flanked by long SDs, is non-allelic homologous recombination (NAHR), also known as ectopic homologous recombination. A subclass of NAHR, known as transposable element-mediated rearrangements (TEMRs), refers to SVs formed via recombination between homologous mobile elements. In addition to NAHR, other SV-mediating mechanisms, especially in contexts of little to no sequence homology, include non-homologous end joining (NHEJ) and replication-based mechanisms (RBMs) [24]. The role of SV-mediating mechanisms, particularly in the formation of inversions and deletions, is discussed in detail in Chapters 4 and 6, respectively.

1.4.2 Recurrent and non-recurrent structural variation

Based on their evolutionary origin, SVs are broadly classified into two categories—recurrent and non-recurrent (Figure 1.4). Recurrent SVs are those that arise independently and repeatedly in the population, typically displaying consistent sizes and shared breakpoint locations across different individuals. These SVs often have breakpoints clustered within flanking SDs [56], and due to their independent origin, they can be found in diverse haplotype backgrounds [31]. The surrounding genomic architecture, particularly the presence of homologous sequences, plays a critical role in their formation. Recurrent SVs are primarily believed

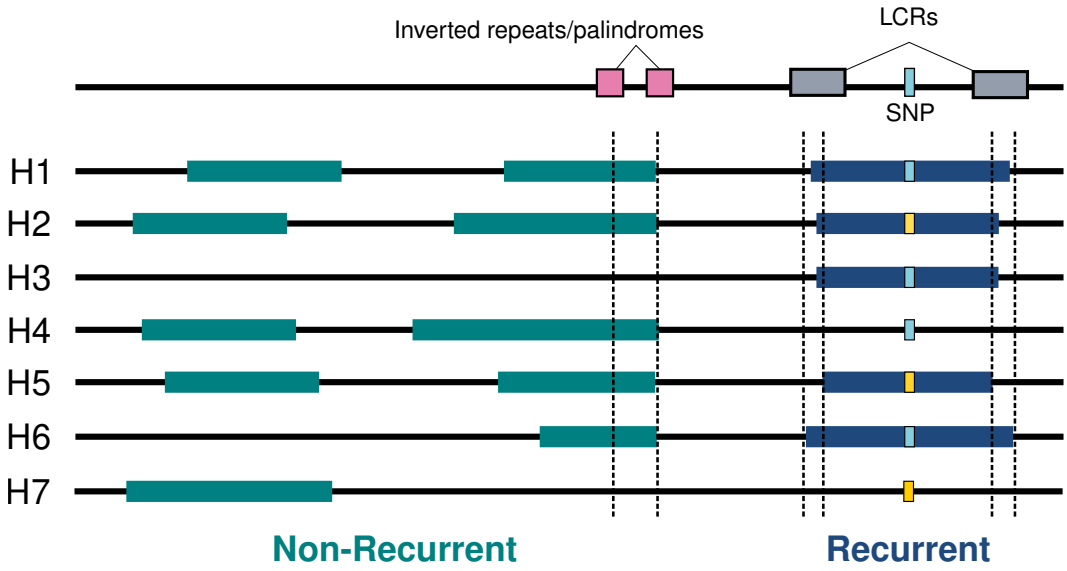


Figure 1.4: Recurrent and non-recurrent genomic rearrangements. Non-recurrent rearrangements exhibit variable lengths and genomic content with no breakpoint clustering. In some cases (as shown in the middle), breakpoints group on one end in regions typically associated with inverted repeats or palindromes. Recurrent rearrangements exhibit breakpoint clustering within the flanking SDs. In some cases, recurrent rearrangements occur independently in different haplotype backgrounds, as shown by both SNP alleles present in haplotypes with and without the rearrangement.

to be mediated by NAHR between flanking SDs [21, 24, 56, 57].

In contrast, non-recurrent SVs—also referred to as single-event SVs—arise once in evolutionary history. They have variable sizes and breakpoint locations, distributed across the genome. While most are unique, some non-recurrent SVs may share a small overlapping region among individuals [56–58]. Interestingly, some of these rearrangements may show breakpoint grouping at one end, localized to a small genomic region [59] (Figure 1.4). Although this grouping differs from the breakpoint clustering characteristic of recurrent SVs, it may still reflect specific aspects of the underlying genomic architecture. In particular, such grouping is often associated with inverted repeats or palindromic sequences that are relevant to the mechanisms driving these rearrangements [59]. Breakpoints of non-recurrent SVs often exhibit features such as microhomologies or small insertions at the junctions. Unlike recurrent SVs, the genomic content and architecture surrounding non-recurrent SVs vary considerably, which makes them difficult to classify based solely on genomic architecture [56]. These SVs are typically mediated by mechanisms that do not require extensive sequence homology, such as NHEJ [21]. Additionally, RBMs like microhomology-mediated break-induced replication (MMBIR) and fork stalling and template switching (FoSTeS) may also contribute to their formation in certain contexts [24, 57, 60].

From a phylogenetic perspective, non-recurrent SVs are defined as events that arise once in evolutionary history, whereas recurrent SVs arise independently multiple times across lineages. Importantly, this means that inherited SVs—even if shared among multiple in-

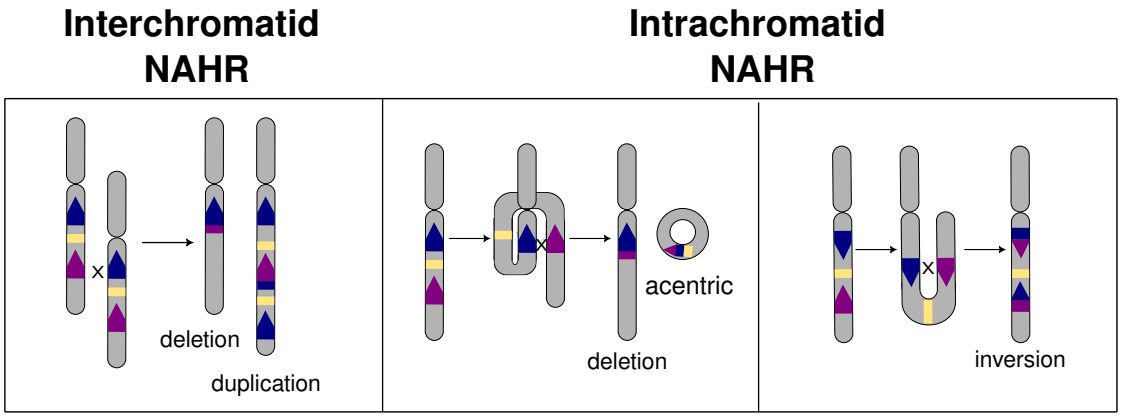


Figure 1.5: NAHR products based on SD orientations and involved chromatids. Interchromatid NAHR between directly oriented SDs results in a deletion and/or a duplication, while intrachromatid NAHR between directly oriented SDs results in a deletion and an unstable acentric chromatid which is lost in subsequent cell divisions. Intrachromatid NAHR between inversely oriented SDs results in an inversion of the fragment they flank. Figure adapted from [61].

dividuals with identical breakpoints—do not qualify as recurrent, since the rearrangement occurred only once, with subsequent inheritance rather than repeated independent formation.

1.4.3 NAHR, SV recurrence, and disease-associations

Multiple instances of DNA rearrangements affecting the same genomic interval have been detected in different individuals, suggesting their recurrent nature [21]. As mentioned above, NAHR is believed to be the predominant process driving recurrent rearrangements in our genome [21, 24, 56, 57]. SDs, highly prevalent across the whole genome [55], provide excellent substrates for NAHR to occur. Cell division—the process by which a parent cell divides into two or more daughter cells—is fundamental for growth, damage repair, and reproduction in living organisms. There are two types of cell division in eukaryotic organisms: mitosis and meiosis. During mitosis, a single parent cell divides once to form two daughter cells [62]. It occurs in somatic cells and is responsible for growth and repair in multicellular organisms. Meiosis, on the other hand, occurs in germ cells and produces gametes—sperm and egg cells. It involves two rounds of cell division and produces four daughter cells, each containing half the number of chromosomes compared to the parent cell [62]. In a diploid cell, meiosis starts with each chromosome replicating to produce two sister chromatids. The homologous chromosomes then pair, forming a structure containing four chromatids. This pairing allows crossover—exchange of genetic material between non-sister chromatids of the two homologous chromosomes. After this step, the homologous chromosomes are pulled to opposite sides, while the sister chromatids remain together. This results in two daughter cells, each with a haploid number of chromosomes, but still containing sister chromatids. Each of these daughter cells then goes through a second cell division (meiosis II)

similar to mitosis where the sister chromatids are separated and distributed to two daughter cells, resulting in the production of four haploid gametes in total [62]. Due to their high sequence identity, homologous SDs are often mistaken by the cellular machinery as allelic counterparts. Crossover between these non-allelic copies leads to genomic rearrangements in the progeny cells. The orientation of involved SD copies determines the type of resulting rearrangement. If the involved SD copies are on the same chromosome and share the same orientation, interchomatid NAHR between them causes duplication and deletion of the intermediate sequence while intrachromatid NAHR only results in deletion of the sequence they flank. In contrast, if the SDs are oriented oppositely, intrachromatid NAHR results in inversion of the fragment flanked by them (Figure 1.5). NAHR between SDs on different chromosomes can lead to chromosomal translocations [57, 63].

The frequency of NAHR is inversely correlated with the intermediate distance between involved SDs, while sequence homology, repeat length, and the presence of specific PRDM9 binding sites—associated with recombination hotspots [64, 65]—are positively correlated with NAHR frequency [24, 66]. However, studies have also shown that the strand exchanges during NAHR do not occur evenly along the SDs, but cluster in narrow "hotspots" [57]. This implies that independent NAHR-mediated recurrences of the same SV event lead to breakpoints converging within these hotspots. In certain cases, more than one SD pair present in close vicinity can be involved in mediating the same event, for example, Smith-Magenis syndrome and Potocki-Lupski syndrome [24, 66]. However, there is always preferential use of one of the SD pairs leading to a distinction between common and uncommon type of the same recurrent SV [24, 66].

NAHR can occur both during mitosis and meiosis, however, only the latter one contributes to genomic variation via germline transmission. For over four decades, meiotic NAHR—and consequently, recurrent SV formation—has been recognized as a major contributor to genomic disorders [22, 23]. These disorders often involve chromosomal microdeletions or microduplications associated with a variety of clinical phenotypes, including developmental delay, intellectual disability, autism, schizophrenia, obesity, and epilepsy [24, 25]. Since the required SD architecture is the same, independent NAHR can give rise to both deletion and duplication of the same genomic segment. Therefore, some genomic disorders associated with deletions have a reciprocal duplication counterpart, with mirror-image phenotypes, for example, individuals with 17p11.2 microdeletion being overweight while those with 17p11.2 microduplication being underweight [24]. These findings suggest that a deeper understanding of NAHR-mediated SV recurrence is crucial for characterizing and uncovering disease-susceptible regions in the human genome and for better understanding the mutational mechanisms behind genomic disorders. Inversions, also predominately mediated by NAHR, though requiring a distinct SD architecture, play a particularly intriguing role, discussed in the next section. Chapters 4 and 5 of this thesis explore the phenomenon of inversion recurrence while Chapter 6 focuses on homology-mediated deletion recurrence.

1.5 The complex universe of inversions

In 1921, Alfred Sturtevant introduced the scientific community to the world of inversions which he discovered by comparing genetic maps of closely related species of Drosophila melanogaster [67]. Sturtevant also observed that inversions, when present in a heterozygous state, suppress recombination due to the production of unbalanced gametes, highlighting their significant potential for shaping evolution. These observations led to inversions becoming some of the earliest genetic markers used to reconstruct phylogenies [68]. Over the next half century, population geneticists focused on studying the inversion polymorphisms in Drosophila, until the research focus shifted more towards biochemical and molecular genetics in the 1970s [69]. However, recent advancements in high-throughput genomic techniques, which allow for the characterization of large and complex structural variants, have brought inversions back into the focus of research [69]. This has led to a growing recognition that inversions contain more genetic variation than collinear regions of the genome [70, 71], are more widespread across taxa than previously acknowledged [72–77], and play a significant role in speciation [72, 78].

Simply put, a chromosomal inversion is the flipping of a genomic segment in place, triggered by a DNA double-strand break which is repaired by cellular repair machinery. However, the actual biology behind this process is far more complicated and diverse partly due to different mediating mechanisms. These mechanisms, previously introduced in Section 1.4.1, can be broadly classified into two categories: non-homology-mediated and homology-mediated mechanisms. Some of the inversions occur in non-homology or micro-homology backgrounds with the most prevalent driving mechanism being NHEJ [13, 24]. Such events are typically short and are accompanied by adjacent >50 bp deletions or insertions and sometimes complex genomic rearrangements. These complex rearrangements are hypothesized to be mediated by RBMs and have been known to be associated with developmental abnormalities [13, 79]. The non-homology-mediated inversions are typically non-recurrent [60]. The most prevalent form of inversion polymorphisms are homology-mediated, with NAHR being the predominant mechanism. As mentioned earlier, NAHR between inversely oriented SDs results in inversion of the segment between them. Homology-mediated inversions are usually longer and, like other NAHR-mediated structural variants, tend to be recurrent.

1.5.1 Inversions and disease associations

As mentioned in Section 1.4.3, many studies have established associations between NAHR driven recurrent SVs and various diseases [22–25]. Particularly in the context of inversions, several studies have provided evidence for orientation changes in disease-associated genomic regions, such as 15q13.3, 15q25, and Xq28 (Hemophilia A locus), throughout primate evolution [26–32]. Interestingly, studies on Williams-Beuren syndrome (WBS) [33] and Koolen-de Vries syndrome (KdVS) [34] have shown inversions predisposing the relevant regions to recurrent pathogenic copy number variations, also known as morbid CNVs

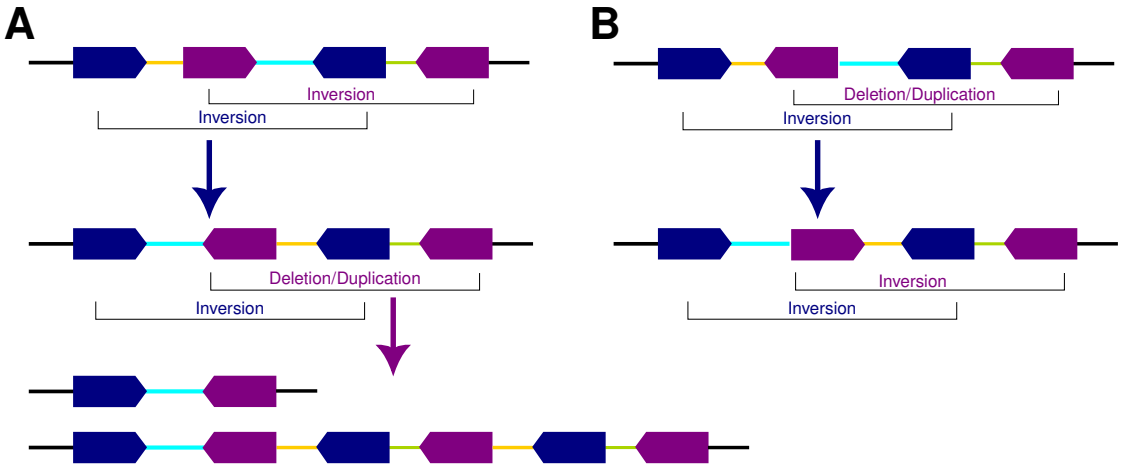


Figure 1.6: Inversions and CNVs. A. NAHR between inversely oriented SD-pairs (blue) flips the orientation of the SD overlapping the intermediate region (purple), bringing the two copies in the same orientation, thus providing a substrate for subsequent NAHR-driven deletion or duplication events. Similarly, NAHR-mediated inversion involving the purple SDs would predispose the region flanked by blue SDs to CNV formation. **B.** NAHR between inversely oriented SD-pairs (blue) flips the orientation of the SD overlapping the intermediate region (purple), bringing the two copies in opposite orientation, thus protecting the region from subsequent NAHR-driven deletion or duplication events.

[13, 80]. In contrast to deletions and duplications, inversions are copy number neutral (unless accompanied by other complex rearrangements) [79]. This raises an important question: if no genetic material is lost, how can inversions lead to or predispose individuals to disease? One aspect of this association is explained by the disruption of protein-coding genes or gene regulatory regions by inversions [81–83]. The most commonly known example of such disruptions is Hemophilia A. About 50% of all severe Hemophilia A cases carry an inversion that flips exons 1-22 of the F8 gene on chromosome X, disrupting the production of clotting factor VIII [81]. This inversion results from NAHR between inverted repeats flanking the affected region.

The other, more prevalent and mechanistically complex aspect of this association involves NAHR and inversion recurrence—also known as “inversion toggling” [32]. In some cases, the inversion flips one or more copies of the inversion mediating or neighboring SDs. This shift in relative orientation of SDs can influence the susceptibility of the region to NAHR-mediated morbid CNV formation. By aligning SDs in the same orientation, the inversion can predispose the region to deletion or duplication events. Conversely, if the inversion brings the SDs in opposite orientation, it can have a protective role [13], as depicted in Figure 1.6. This phenomenon becomes even more interesting if the inversion is recurrent. A recurrent inversion can be protective in some individuals, while in others it may predispose the same region to disease, depending on the location of each event’s breakpoints. Several studies in human cohorts have revealed an unexpected high degree of inversion recurrence, particularly for those mediated by inverted SDs, establishing inversion recurrence to be a

widespread phenomenon with significant implications in disease and evolutionary contexts [32, 84, 85]. Given these insights, investigating inversion recurrence—particularly in patient cohorts—can offer valuable information about disease susceptibility and the underlying mechanisms. However, unlike many other classes of SVs that are becoming increasingly easier to detect because of advances in sequencing technologies, inversions remain exceptionally difficult to identify which has rendered them one of the most underexplored classes of structural variation [13].

1.5.2 Inversion detection: Challenges and solutions

Genome-wide identification of inverted SDs revealed that as much as 12% of the human genome is susceptible to NAHR-mediated inversions [86]. Despite this high susceptibility, the number of known inversion polymorphisms remains highly underestimated [24]. This is largely because inversions are among the most challenging type of SVs to detect and validate [11, 12], primarily due to the following reasons:

- **Copy number neutrality:** Unlike deletions or duplications, inversions do not result in a net loss or gain of genomic material. As a result, sequencing depth based inference is largely uninformative for detecting inversions.
- **Localization in repetitive regions:** Inversions often occur in highly repetitive and difficult-to-map regions of the genome, that is, regions that pose challenges for accurate alignment of sequencing reads, which hinders accurate breakpoint identification using conventional sequencing technologies.
- **Recurrence:** Inversions are highly recurrent [32, 84, 85], arising independently in different haplotype backgrounds. This results in recurrent inversions having low linkage disequilibrium (LD) with nearby variants making their identification via inversion-tagging variants nonviable.

The remainder of this section focuses on the commonly used SV detection techniques, examining their respective strengths and limitations, with a particular emphasis on their effectiveness in identifying inversions.

Short-reads-based SV identification

Despite their limitations in resolving repetitive and structurally complex regions of the genome, short-read sequencing remains the standard technology for SV identification, primarily due to its cost-effectiveness [87]. In genomic regions where sequence alignments are reliable, short reads can help detect SVs in several ways. For example, insert size (the distance between mates of a read pair) can be used to identify deletions and insertions, read depth can indicate copy number variations, and the orientation of aligned reads can reveal

inversions [87]. However, because most of the structural variation occurs in difficult-to-map regions, the effectiveness of short reads in SV identification is limited [11, 87]. Even with substantial methodological advancements, no short-read-based technique is capable of detecting all types or size ranges of SVs [87]. This limitation is especially evident for inversions, largely due to the presence of long and highly identical flanking SDs, which exceed the length of short reads leading to unreliable alignments [11]. This is reflected in the extreme under-representation of inversions in SV call sets that primarily rely on short-read sequencing [10].

Recent developments in the field of pangenomics has made it possible to nearly double the number of SVs genotyped using short reads by a process referred to as "genome inference" [35, 88]. This improvement is achieved by comparing short reads to a highly detailed and linked layout of variation represented in the form of a pangenome instead of a single linear reference [12, 88]. However, this approach depends heavily on LD between variants in the pangenome, which, as noted earlier, is disrupted in the case of recurrent inversions. In summary, short-read sequencing, even in the most optimized settings, is impractical for detecting inversions reliably.

Long-reads-based SV identification

The (ultra-)long reads, especially those from ONT and PacBio, which can span several kilobase pairs of sequence, have enabled the capture of both simple and complex structural rearrangements that are predominantly located in repetitive regions [11, 12, 87]. While long reads enable more accurate alignments and can cover much longer genomic stretches compared to short reads, their higher cost and error rates limit their widespread use in SV discovery and characterization [7, 87]. Furthermore, despite their ability to span longer DNA sequences, some SV alleles remain challenging to capture, for example, several megabase pairs (Mbp) long inversions. Even with the development of local assembly-based methods to address these issues, long-read technologies still struggle to resolve long inversions flanked by >5 kbp repeats [87].

Genome assembly-based SV identification

Genome assembly can be either reference-guided, where reads are aligned to a reference genome and pieced together, or *de novo*, where a genome is assembled from scratch without any reference. *De novo* genome assembly, being exempt from reference bias, is considered more effective at capturing novel variation. In fact, Chen et al. describe whole genome *de novo* assembly as "the ultimate solution for SV characterization" [89]. In recent years, several studies have generated high-quality *de novo* genome assemblies using integrated frameworks, primarily leveraging long-read technologies, and supported by orthogonal sequencing techniques such as single-cell DNA template strand sequencing (Strand-seq) or high-throughput chromosome conformation capture sequencing (Hi-C) [12, 35]. These as-

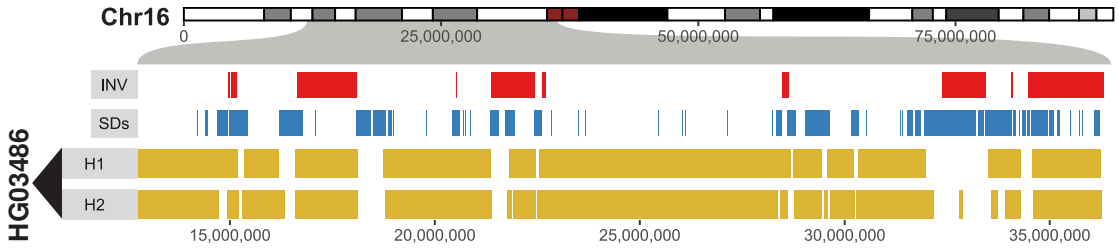


Figure 1.7: Genome assembly truncations around inversions. Genome assemblies for both haplotypes (H1 and H2) of sample HG03486 in an inversion-rich region on chromosome 16 are shown at the bottom. Sequence resolved regions are shown in yellow, while white regions indicate assembly truncations. Figure adapted from [12].

semblies have significantly improved variant characterization, with most pronounced improvement seen for SVs, particularly long insertions where assemblies capture $>85\%$ of the previously unknown variation [11, 12]. However, again, inversions are an exception, primarily because genome assemblies tend to break or collapse in highly repetitive regions that constitute the flanking regions for majority of the inversions [12] (Figure 1.7).

Strand-seq based SV identification

A promising solution to the challenges of inversion discovery and characterization, is the use of Strand-seq [11, 13, 91]. Developed by Falconer et al. in 2012, Strand-seq is a specialized DNA sequencing technique that sequences individual DNA template strands separately, while preserving their directionality, thus making it possible to track the parental origin of each DNA strand [91]. During DNA replication, the two inversely oriented "template" strands—Watson(W, +) and Crick(C,-), based on their 5'–3' orientation—detach from each other followed by the synthesis of two new strands relative to the respective template. The protocol, depicted in Figure 1.8A, begins with the incorporation of a thymidine analog, such as bromodeoxyuridine (BrdU), during a single round of DNA replication. Cells are cultured in the presence of BrdU, which is selectively incorporated into newly synthesized DNA strands, while the original parental DNA strands remain unlabeled. Following cell division, several daughter cells are selected via Fluorescence-activated cell sorting (FACS) and deposited into single wells of a 96-well plate. The newly synthesized BrdU-labeled strands are then selectively degraded through a strand-specific cleavage process, typically using ultraviolet light combined with exonuclease treatment. This results in the retention of only the original parental template strands for each chromatid. The remaining template strands are subsequently prepared for sequencing on an Illumina platform through library construction protocols, which include fragmentation, adapter ligation, and amplification steps [92].

The orientation in which a Strand-seq read aligns to a reference genome indicates the orientation of the strand of origin. For diploid organisms, there can be four equally-likely possible strand states, referred to as the "ground states", for the daughter cells, that is, CC, CW, WC or WW, independently for each chromosome per single-cell. The Strand-seq reads

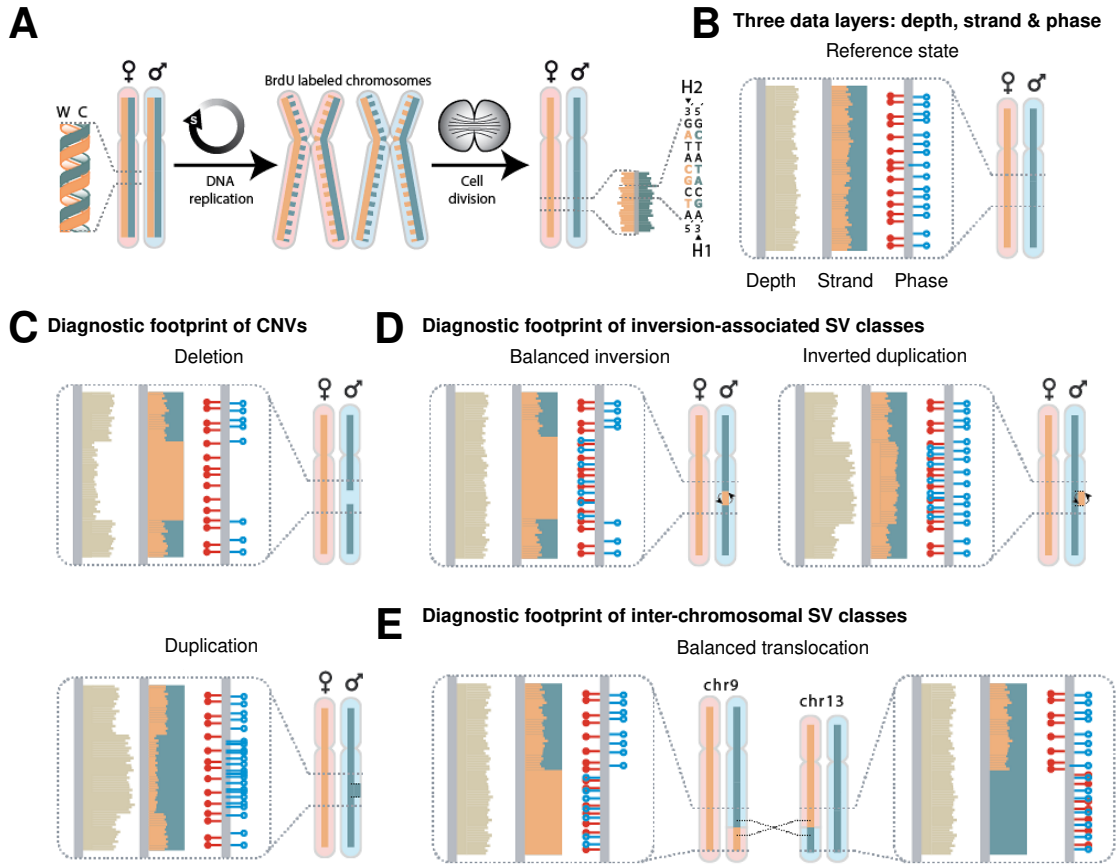


Figure 1.8: Strand-seq based SV diagnostic footprints. **A.** Both maternal and paternal chromosomal copies consist of a positive (Crick, "C", teal) and a negative (Watson, "W", orange) template strand. During DNA replication, BrdU is incorporated exclusively into nascent DNA strands (dashed lines). After cell division, each daughter cell inherits both Crick, both Watson or one Watson and one Crick template strands from the parental chromosomes. **B.** An example Strand-seq depth, strand distribution, and phase pattern expected in case of no structural variation. **C.** Deletion (top): haplotype-specific loss in read coverage with no change in read orientation. Duplication (bottom): haplotype-specific gain in read coverage with no change in read orientation. **D.** Balanced inversion (left): haplotype-specific read orientation flips with unaltered read depth. Inverted duplication (right): haplotype-specific read orientation flips accompanied with a read depth gain of the same haplotype. **E.** Balanced translocation: correlated template strand switches in the same paired genomic regions. Figure adapted from [90].

coming from a Watson strand would map to a reference genome in forward and the ones coming from a Crick strand would map in reverse orientation. Therefore, in WC and CW cells, a clear haplotype-based assignment of the Strand-seq reads is possible, a property that is utilized for Strand-seq-based phasing of genome assemblies [12]. However, sister-chromatid exchanges and other SVs, particularly inversions, can make the strand-states vary within the same chromosome as well. For example, a heterozygous inversion in a CC cell would make the cell CW/WC in the inverted region or a homozygous inversion in a CC cell would switch the stand state to WW in the respective region. Panels B–E in Figure 1.8,

depict the Strand-seq diagnostic signatures for different classes of SVs. Deduced based on read depth, strand and phase information, these footprints help identify different types of SVs at single-cell level [37].

The quality of Strand-seq libraries plays a crucial role in accurate SV characterization. Some of the libraries can be of low quality, that is, they might exhibit low read depth (≤ 100 reads per Mbp), contain no reliable strand-state information because of failed BrdU incorporation, or go through more than one round of BrdU incorporation, leading to patches of missing reads [92]. Such cells can negatively impact the SV breakpoint localization and copy number predictions [92]. Therefore, manual or automated [93] cell selection procedures are employed to select "good quality" Strand-seq libraries for downstream analyses.

In comparison to other standard DNA sequencing technologies that struggle with identifying inversions because of large flanking SDs that cannot be sequence-resolved, Strand-seq enables inversion detection solely by identifying strand switches along the chromosomes [11]. Multiple studies have highlighted this improved performance of Strand-seq particularly in identifying large (>50 kbp) inversions that other contemporary methods fail to detect [11]. In addition to inversions, Strand-seq can analogously help improve the detection accuracy of CNVs, for example, deletions, duplications, and inverted duplications by using the SV diagnostic signatures, as shown in Figure 1.8B–E. However, the technology comes with its own limitations. Strand-seq has extremely sparse read coverage ($0.03\times$) [94], which limits its sensitivity for detecting small SVs (typically <5 kbp) and constrains its ability to resolve complex or nested variation.

In summary, until the advent of a standalone sequencing platform capable of generating telomere-to-telomere (T2T) resolved assemblies, there is no "one-size-fits-all" sequencing technology or detection method capable of fully capturing the extensive structural diversity present in human genomes. The most effective strategy, as demonstrated by numerous studies [11–13, 35], is to leverage the strengths of multiple platforms by employing integrated, multi-technology approaches for structural variant discovery and characterization. This aspect, in the context of inversions, is discussed in detail in Chapter 3.

1.6 Pangenomics

Even with remarkable advancements in technologies and methodologies for studying structural variation, there remain inherent limitations when working with a single linear reference genome. As the name indicates, a single reference genome represents one haplotype, or at best, a consensus sequence derived from a small cohort, thus failing to capture the full breadth of human genetic diversity. While large-scale efforts, such as the 1000 Genomes Project (1KGP) [5], have cataloged millions of variants across global populations, rare and population-specific variants continue to be underrepresented. Solutions to address this limitation include increased sampling from diverse populations, the development of variation-

aware alignment algorithms, and the use of personalized or population-specific reference sequences. A more compelling solution that the scientific community is increasingly adopting is the shift from a linear reference genome to a pangenome. The Human Pangenome Reference Consortium (HPRC) defines a pangenome as "the collection of all of the genetic information of a species" [95]. A pangenome reference consists of multiple individual genome assemblies and their alignments, typically represented as a graphical structure that serves as a coordinate system. In a pangenome graph, genomic variants appear as "bubbles" [96], which are structures representing multiple path traversals between a common start and end graph node. The nodes present between these start and end nodes, termed as "inside nodes", correspond to different variant alleles. The graphical layout complexity of the variants represented in a pangenome graph varies across different variant types. There are "simple bubbles", representing simple variation, for example, SNPs and indels, without any interleaved complexity. On the other hand, there are "superbubbles" which depict complex structural variants or nested variation (overlap of different variants).

The concept of a pangenome was first introduced in microbial genomics, where researchers observed significant differences in gene content even among closely related bacterial strains [97]. The pangenome concept has since expanded across various domains, including plants [98–101], fungi [102], and animals [103–106], shedding light on population-specific genetic variations, adaptive traits, and disease-associated mutations. Human pangenomics has also significantly advanced in recent years, improving our understanding of human genetic diversity by moving beyond the limitations of a single reference genome. One major step in this direction is the work by the HPRC, which released a draft human pangenome in 2023, comprising 47 phased, diploid assemblies from a diverse cohort [35]. This reference integrates over 100 million new basepairs, capturing substantially more population diversity than previous references [35]. In addition to representing known variants and haplotypes, this pangenome revealed new alleles at structurally complex loci [35]. Moreover, pangenomes have demonstrated improved characterization of previously known but unresolved structural variants—that is, variation whose exact nature or structure is not fully determined or characterized [12, 88]. Ongoing efforts to build larger and more representative pangenomes are expected to further enhance the accuracy of genomic analyses and foster more inclusive biomedical research. Consequently, there is a growing need to develop computational methods that operate in the pangenomic space to fully leverage the increased depth and granularity of genomic information it provides. This aspect is explored in Chapter 6 of this thesis, which describes a computational method to detect inversion recurrence using pangenome graphs.

Chapter 2

k-merald: Allele detection using *k*-mer based sequencing error profiles

This chapter introduces k-merald, an approach for allele detection that utilizes technology-specific sequencing error profiles in order to improve the allele detection accuracy using long reads. This work was published in Bioinformatics Advances [36] and this chapter presents an extended version of this publication of which I am the first author. All sections presented in this chapter re-use material from this publication. For publication details and author contributions, please refer to Section C.1.

2.1 Introduction

Genotyping is a process used for detecting the genotypes of an individual, which further helps in detection of haplotypes, a task termed as phasing. These processes are widely used in studying the genetic aspects of different diseases and genetic relationships among species. Both genotyping and phasing typically utilize the alignment between sequencing reads and a reference genome. Thus, prior to genotyping, it is important to determine for each sequencing read, whether it carries the reference allele, generally denoted as "0" or one of the alternative alleles, generally denoted as "1,2,3,...,n" at each of the variant positions it overlaps. This process is commonly referred to as allele detection, formally defined as follows: Let V be a set of all variant positions across the reference genome, let $v \in V$ be a variant position with alleles $a_1, a_2, ..., a_n$, and let B_v be the set of sequencing reads aligned to v. Determine a_{v_b} for each sequencing read $b \in B$, where a_{v_b} denotes the allele carried by b at position v.

Most commonly, short sequencing reads from second-generation sequencing technologies, for example, Illumina, are used for this purpose because long reads obtained us-

ing third-generation sequencing technologies, for example, Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), tend to be more prone to sequencing errors [107], unless techniques like circular-consensus sequencing (CSS) are employed [108]. However, long reads can be much more informative as they can span longer genomic regions and may cover many variant positions and repetitive regions [40, 109].

Over the years, a lot of work has been done to improve basecalling, a process translating raw ONT signal into a DNA sequence. Earlier basecallers employed a two-step process, involving pre-segmentation of raw signals followed by nucleotide label prediction using hidden Markov models (HMMs) [110] or recurrent neural networks (RNNs) [111]. Recent years have seen a surge in development of deep learning-based basecallers, dealing directly with the raw signals, hence avoiding error propagation caused by wrong segmentation [107]. Although state-of-the-art deep learning based approaches and the introduction of R10 flow cells and duplex sequencing have led to significant improvement in the basecalling accuracy [112–119], the error rates of cost-effective and therefore commonly used long-read sequencing techniques remains higher than that of short-read sequencing. ONT's most cost-effective basecaller Guppy, even in the high-accuracy mode achieves basecalling accuracy up to 95% [120], while Illumina HiSeq has basecalling accuracy of around 99.9% [107].

Most commonly used read alignment algorithms, such as BWA [121], do not take sequences of alternative alleles into account for alignment. This results in reference bias [43], and pangenomic approaches have been proposed to overcome this problem [88, 122–124]. Despite these advancements, aligning to a single linear reference genome remains the standard workflow today. Combined with systematic sequencing errors [108, 125], this can make alignments at variant sites unreliable to be used for allele detection, thus commonly resulting in sequencing errors being mistaken for an alternative allele. One approach to deal with the first problem, for example employed by WhatsHap [44, 45, 126], is read realignment. WhatsHap extracts the read sequence from a variant window, 10 bp upstream to 10 bp downstream from the variant position. It then aligns this read sequence to the corresponding reference sequence and to the alternative sequence—produced by interchanging reference with the alternative allele at the variant position. The read is then assigned the allele with lower alignment cost and "unknown" in case of equal scores [45]. The alignment costs are calculated based on edit distance between the sequences. While this technique outperforms the allele detection methods without re-alignment, it does not take systematic sequencing errors into account. Tools like Clair3 [127], DeepVariant [128], and PEPPER [129] perform variant calling and subsequent genotyping of the discovered variants. However, to our knowledge, there are no tools designed specifically for long-read-based genotyping of a set of variants given as input, apart from WhatsHap [44, 45, 126].

To enhance allele detection accuracy by accounting for systematic sequencing errors, we developed a new method, *k*-merald. As the name indicates, *k*-merald operates in *k*-mer space rather than at the single nucleotide level, as *k*-mers help capture the genomic context in which sequencing errors arise. This approach is based on the idea that genomic regions

Table 2.1: Overview of used notations

Notation	Definition
R	the complete reference sequence
B	sequencing reads aligned to R
V	a set of variants for which allele detection is to be performed
w_v	a window of fixed number of w base pairs on each side of $v \in V$
F	R excluding w_v for all $v \in V$
D	B excluding parts of read sequences mapping to a w_v for all $v \in V$
f	a *k*-mer belonging to F
d	a *k*-mer belonging to D
M	a matrix recording the occurrence count for each reference-read *k*-mer pair, (f, d)
P	a matrix recording the probability of occurrence of each reference-read *k*-mer pair, (f, d)

with no variation can be used to learn the characteristics of sequencing errors. The error model derived from these regions is then employed to differentiate between allelic variants and sequencing errors at variant positions. *k*-merald begins by traversing all confident non-variant regions of the genome, recording the sequence and count of read *k*-mers aligning to each reference *k*-mer (reference-read *k*-mer pairs). These pairs include matches, indicating error-free positions, and mismatches, which suggest sequencing errors. The counts of these *k*-mer pairs are used to calculate the probability of observing each reference-read *k*-mer pair across the entire genome. Additionally, *k*-merald employs a novel approach for global sequence alignment in *k*-mer space. For each variant window (excluded during the training phase), the read, reference, and alternative sequences are split into *k*-mers, and the string of *k*-mers are then aligned. Instead of using a fixed penalty for mismatches, *k*-mer mismatches are penalized according to the learned error model, allowing common sequencing errors to be tolerated at a lower cost. The sequencing read is then assigned to the allele with the lowest alignment cost. *k*-merald has been incorporated into WhatsHap and is available as an alternative to the edit distance-based allele detection (`https://github.com/whatshap/whatshap`).

2.2 *k*-merald: Algorithmic Overview

2.2.1 Training the Model

As input, *k*-merald expects a list of candidate variants and the aligned sequencing reads. In the first step, as shown in Figure 2.1A, a sequencing error profile is constructed based on non-variant regions of the genome, that is, regions where the sequencing reads and reference sequence would be identical if sequencing errors were absent. Any changes (for example, insertions, deletions, substitutions) in the read sequences mapping to these regions can give an indication of the nature of sequencing errors inherent to the sequencing technique that generated the data. An overview of the notations used in this chapter is provided in Table 2.1. Let F be the reference sequence excluding all variant windows, where each variant window, w_v, is defined as an interval containing the complete variant v and a flanking re-

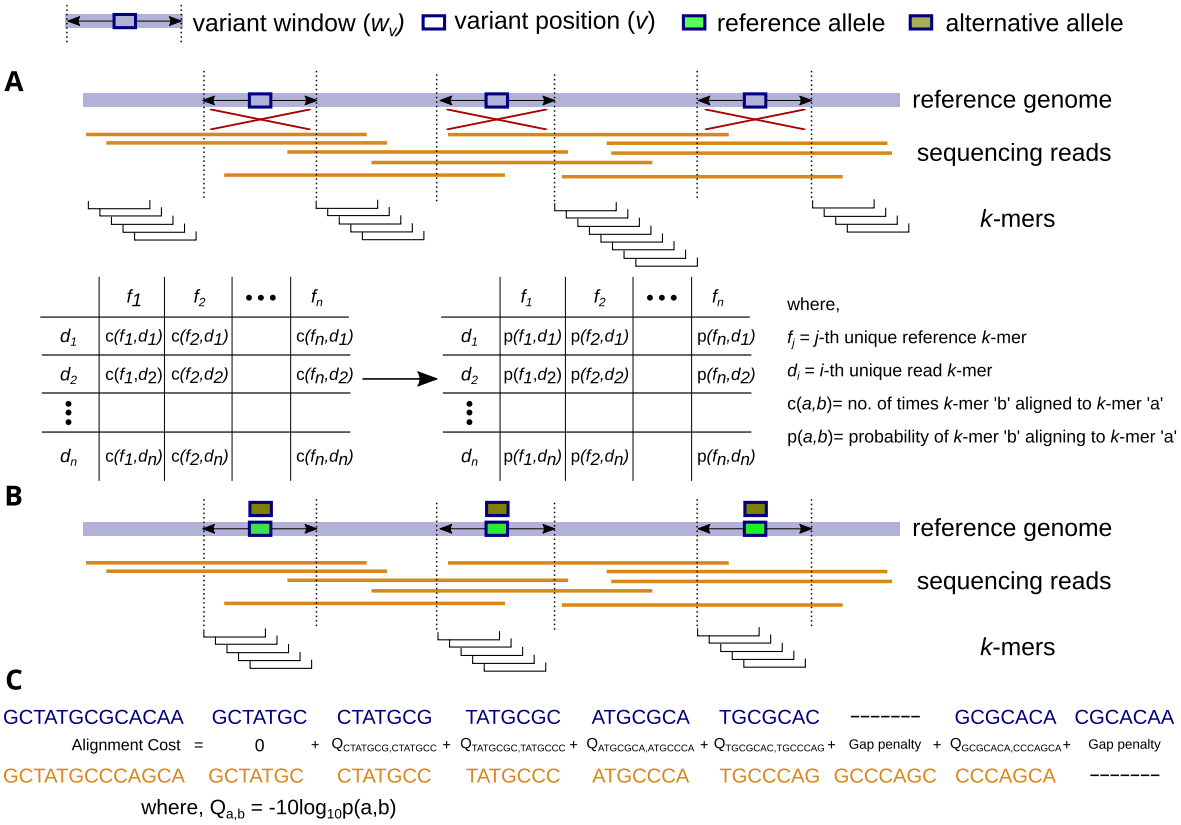


Figure 2.1: *k*-merald Outlook. A. Model training: Counts for all the unique reference-read *k*-mer pairs (f, d) in non-variant regions of the genome are recorded. These counts are then used to construct a matrix storing for each unique reference *k*-mer f, the probability of seeing each read *k*-mer d. **B**. A variant window, w_v, containing the complete variant v and a flanking region of a fixed number of w base pairs on each side is considered. Both reference and read sequences inside w_v are converted into *k*-mers. **C**. Strings of consecutive *k*-mers from each read sequence are aligned individually to the *k*-mer strings obtained from the reference and alternative allelic sequences. A global alignment of the two strings of *k*-mers is done in a similar fashion as global alignment of two base-pair sequences, while using phred-scaled probabilities, stored during model training, as alignment costs.

gion of a fixed number of w base pairs on each side. The training data consists of F and the set of sequencing reads aligned to it, D. Suppose f denotes a *k*-mer belonging to F, while, d denotes a *k*-mer belonging to a sequencing read from D. During model training, described in Algorithm 1, F is traversed from left to right while maintaining, for each f, the count of each mapping d using the mapping positions from the input read alignments. For extracting the reference-read *k*-mer combinations (f, d), the read sequence is considered and not the alignment. For example, if the read *k*-mer `AC-GTCT` is aligned to the reference *k*-mer `ACTGTCT`, the respective (f, d) would be $(\texttt{ACTGTCT}, \texttt{ACGTCT*})$, where $*$ is the nucleotide following `ACGTCT` in the read sequence. These counts of *k*-mer combinations (f, d) are then aggregated across all occurrences of each reference *k*-mer to obtain a unique matrix M, with reference *k*-mers f shown in columns (j) and read *k*-mers d represented in rows (i).

Algorithm 1: Model Training

Input : the complete reference sequence R,
aligned sequencing reads B,
sorted list of variants for which allele detection is to be performed V.
Output: M

```
counter, i ← 0
v ← V[counter]
while i < |R| do
    if i >= v − w & i <= v + w then
        do nothing
    else if i > v + w then
        counter ← counter + 1
        v ← V[counter]
    else
        k_R ← R[i, i + k − 1]
        for b ∈ {b′ ∈ B| alignment of b′ contains R[i]} do
            k_b ← b[j, j + k − 1] | b[j] aligns to R[i]
            M[(k_b, k_R)] ← M[(k_b, k_R)] + 1
        end
    end
    i ← i + 1
end
return M
```

An entry M_{ij}, thus shows the number of times the read *k*-mer d_i aligned to the reference *k*-mer f_j across the whole length of the reference sequence F. Although there are 4^k possible sequence combinations for a *k*-mer of length *k*, many of these combinations are not observed. The (f, d) *k*-mer combinations that are not observed across the whole length of F are each given a pseudocount value ϵ. Instead of representing presence and absence by "0" and "1" respectively, a pseudocount value ϵ implies that these *k*-mer combinations can theoretically exist, but have a low probability of occurrence based on our training data. For each reference *k*-mer f, we define K_f as the set of all *k*-mers d aligned to f, that is, the pair (f, d) has an entry larger or equal to 1 in our matrix M. The sum of individual counts over all of these pairs is denoted by t_f. The matrix of counts M, is then converted into a matrix P, storing the probability of observing each possible reference-read *k*-mer pair (f, d). So, P_{ij} represents the probability of observing a *k*-mer combination (f_j, d_i) and is calculated as follows:

$$P_{ij} = \frac{M_{ij}}{t_{f_j} + (4^k - |K_{f_j}|) \cdot \epsilon} \tag{2.1}$$

In our implementation, the input data required for this training phase is provided as a VCF file with variant positions, a reference sequence in a FASTA file and a BAM or SAM file containing sequencing reads aligned to the reference sequence. This model training step can be performed using the "`learn`" module in WhatsHap.

Algorithm 2: *k*-mer Alignment

Input : $S_1 \leftarrow$ list of *k*-mers from the target sequence,
$S_2 \leftarrow$ list of *k*-mers from the query sequence,
$C_{\text{gap}} \leftarrow$ gap penalty,
$Q \leftarrow$ Matrix with phred-scaled probabilities.

Output: Optimal cost for aligning S_1 to S_2

```
for i ← 0 to length(S1) do
    DP[i, 0] ← C_gap * i
end
for j ← 0 to length(S2) do
    DP[0, j] ← C_gap * j
end
for i ← 1 to length(S1) do
    for j ← 1 to length(S2) do
        C_match ← DP[i − 1][j − 1] + Q_{S1[i−1], S2[j−1]}
        C_delete ← DP[i − 1][j] + C_gap
        C_insert ← DP[i][j − 1] + C_gap
        DP[i][j] ← min(C_match, C_delete, C_insert)
    end
end
return DP[length(S1), length(S2)]
```

2.2.2 Alignment Algorithm

In this step, read re-alignment to the alternative and reference sequence is performed. *k*-merald uses the probability matrix P, which represents the model of sequencing errors generated in the previous step, to define an alignment cost and, based on this, determines the minimum cost allele defined as follows: Let b_v be the read sequence segment aligned to a variant window w_v, and let $Q = [q_1, ..., q_n]$ be the set of all possible allele sequences belonging to w_v, that is, q_1 corresponds to w_v sequence with reference allele at v and $q_2, \ldots, q_n$ represent the sequences with alternative alleles at v. If $cost(x, y)$ denotes the alignment cost for two sequences x and y, then

$$a_{v_b} = \arg \min_{i \in \{1,\ldots,n\}} cost(b_v, q_i)$$

where, a_{v_b} denotes the allele carried by b at position v. Therefore, for a given variant position, *k*-merald seeks to determine whether an observed sequencing read is more likely to have originated from the reference allele or from one of the alternative alleles. In this phase, *k*-merald only deals with variant windows, that is, the regions that were not considered in the model training phase described in Section 2.2.1. The read sequences from each w_v are aligned to both the reference and alternative sequence of the respective w_v, as shown in Figure 2.1B. The reference sequence for each w_v is extracted directly from the reference genome, while the alternative sequence is obtained by replacing the reference allele with the alternative allele at the variant position. For alignment, we developed a modified ver-

sion of the Needleman-Wunsch algorithm [41]. This modified algorithm, described formally in Algorithm 2, performs *k*-mer based comparisons (Figure 2.1C) instead of the conventional single-character based sequence comparison. Each sequence is first converted into a string of consecutive *k*-mers and the resulting strings are then aligned by comparing respective *k*-mers. The algorithm uses "phred-scaled" probability scores (-10 $\cdot \log(\text{probability})$) for alignment cost calculation, where probability values are obtained from the matrix P learned from the training phase (Section 2.2.1). This cost model is used to penalize mismatches when the reference *k*-mer and the read *k*-mer are not identical. The mismatching *k*-mer pairs frequently observed across the non-variant positions, due to systematic sequencing errors, hence having a high probability in matrix P, get a lower penalty as compared to those seen occasionally due to sporadic sequencing errors. For gaps, the probability value can be specified by the user as a parameter.

In summary, by design, this modified global alignment algorithm ensures that a read carrying a sequencing error aligns to the reference with a cost lower than to the alternative allele, thus minimizing the risk of a sequencing error being mistaken for a variant allele. The read is assigned the allele resulting in lowest alignment cost. However, equal alignment costs indicate that the algorithm was unable to detect the correct allele based on the alignment. In case of multi-allelic variants, the alignment is performed using each alternative sequence. As mentioned above, *k*-merald has been implemented inside WhatsHap and can be used as an alternative approach for allele detection in (i) ***haplotagging*** [45], the process to label each read with a haplotype of origin, (ii) ***genotyping*** [109] and (iii) ***phasing*** [44].

2.3 Results

For details regarding the data used for the generation of results presented in the following sections please refer to Appendix A.

2.3.1 Sequencing Error Profiles

We first visualized the sequencing error profiles for ONT, PacBio CLR and PacBio HiFi, respectively. These profiles were generated using sequencing reads from sample HG002 aligned to human reference genome GRCh38. For comparison, we generated simulated long-read data with uniform error distribution with an error rate of 0.05, 0.1, and 0.15, each with an average read length of 20 kbp and 35$\times$ mean coverage across available positions. The rate of mutations was set to 0.1$\%$, of which 10$\%$ are indels. The aligned simulated reads and simulated variants were used for generation of the error profiles, as described in Algorithm 1. Figure 2.2 shows the error profiles generated by setting $k = 7$ and $w = 25$. The error rate for each reference *k*-mer represents the sum of probabilities of observing each *k*-mer pair (f, d) such that $d \neq f$. Figure 2.2 shows that in contrast to the error rate pattern observed for data with uniform base-line error rate, the error rate distribution differs across the sequencing technologies and is non-uniform for each of them. A closer look at the 25 most erroneous

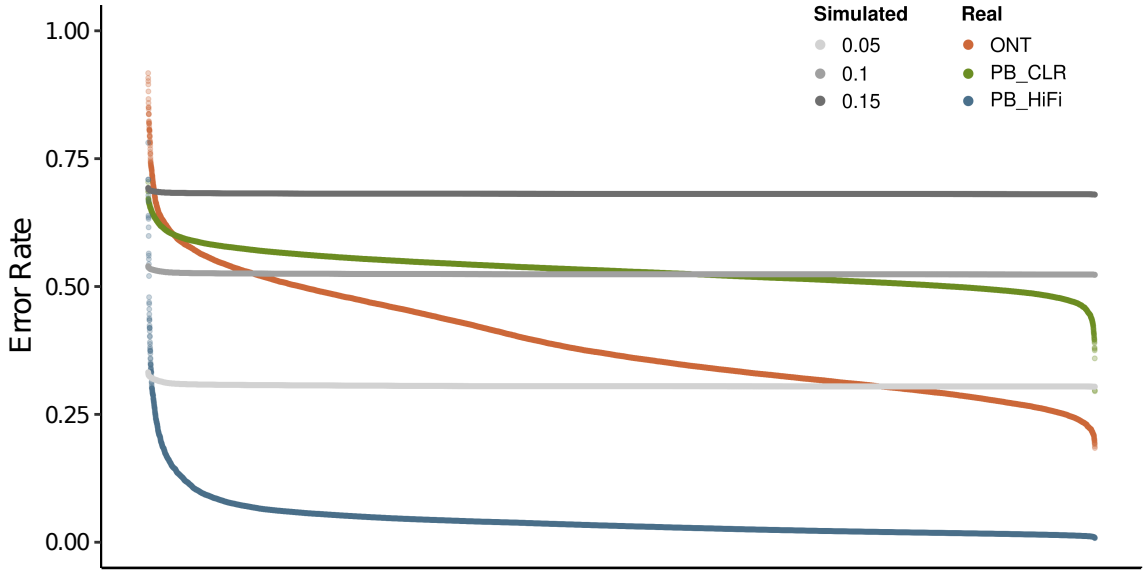


Figure 2.2: Distribution of 7-mer error rates observed for simulated and real long-read data sets. The simulated short reads have a uniform base-line error rate distribution with an error rate of 0.05, 0.1 and 0.15, each using a read length of 20 kbp and 35× mean coverage across available positions. The rate of mutations was set to 0.1%, of which 10% are indels. The real data set includes sequencing reads from ONT, PacBio CLR and PacBio HiFi for sample HG002. The horizontal axis represents the unique *k*-mers belonging to the GRCh38 reference genome (removed for ease of visualization).

k-mers for ONT, PacBio CLR and PacBio HiFi, each, reveals that the nature of erroneous *k*-mers also differs across the sequencing technologies (Figure 2.3). The erroneous *k*-mers from PacBio CLR seem to be more GC-rich while ONT erroneous *k*-mers appear to be AT-rich. The fact that these error distributions are not uniform and are sequence-wise distinct from one another, supports our hypothesis that instead of using a generalized method across all platforms, considering technology-specific error profiles can help improve allele detection accuracy.

Genome in a Bottle (GIAB) variant callsets, used for generation of the results presented in this chapter (Appendix A), come with a designation of high confidence regions in which the callsets can be considered complete. However, for the remainder of the genome, they are less complete. To assess the impact of missing variant positions on the error profiles, we evaluated the genotyping performance across error models learned using multiple variant callsets. Each of these callsets contained only a percentage of variants, ranging from 1% to 95%, from the full GIAB benchmark callset [130]. We observed that the genotyping error rates remained almost unaffected even after excluding a large fraction of variant positions (Figure 2.4), hence proving *k*-merald's robustness.

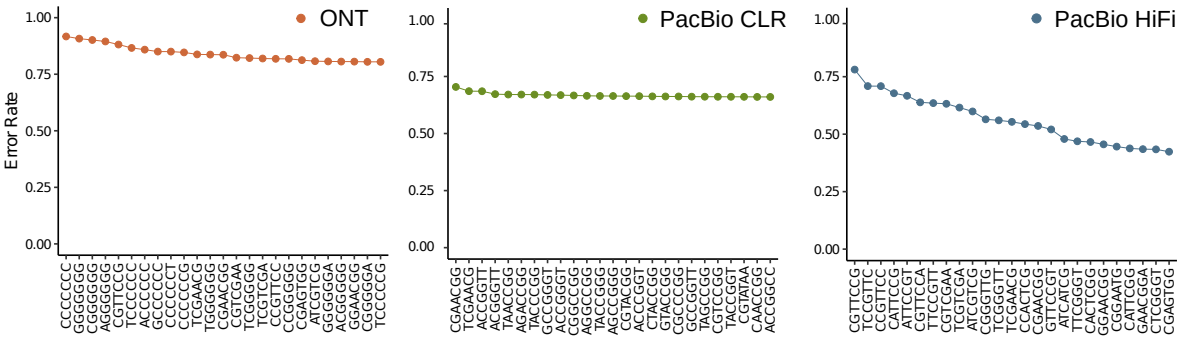


Figure 2.3: Error rate distribution across sequencing technologies. Error rates for the 25 top most erroneous 7-mers, belonging to the GRCh38 reference genome, for sequencing reads from ONT, PacBio CLR and PacBio HiFi individually.

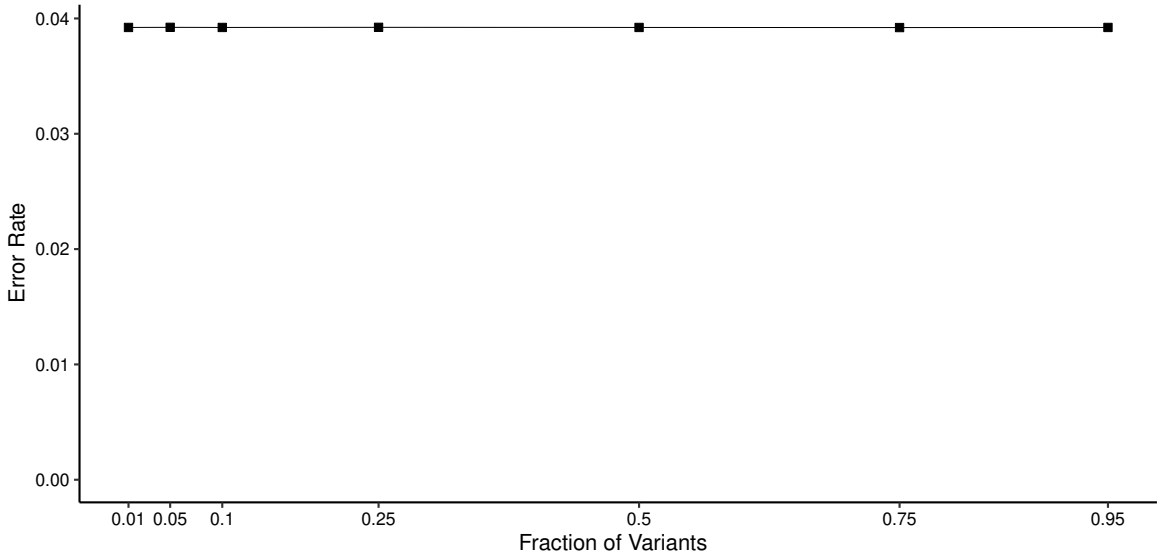


Figure 2.4: Robustness to missing variant positions. Assessment of *k*-merald's robustness by comparing the genotyping performance across error models learned using multiple variant callsets. Each of these callsets contained only a percentage of variants, ranging from 1% to 95%, from the original GIAB v4.2.1 benchmark callset. The horizontal axis represents the fraction of variants used for the training phase, while the vertical axis represents the corresponding genotyping error rate.

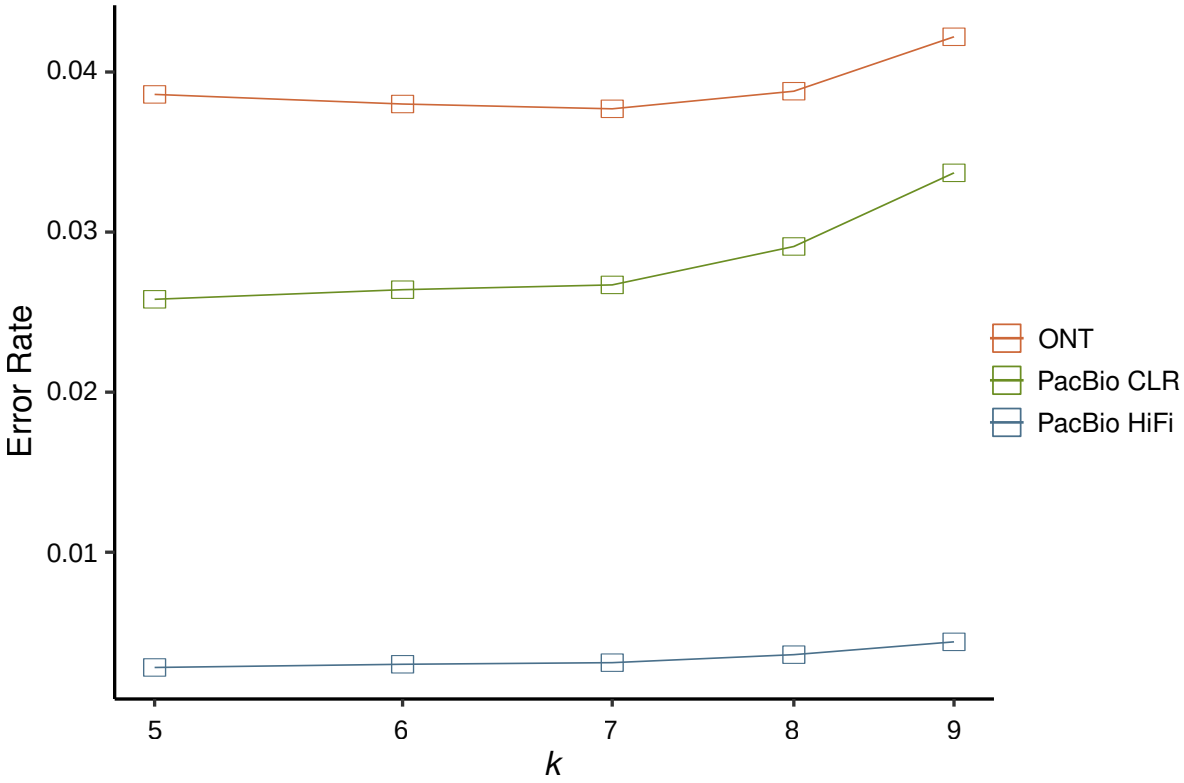


Figure 2.5: Robustness to the value of *k*. Error rates observed across different values of *k* used by *k*-merald, across ONT, PacBio CLR and PacBio HiFi data for chromosome 22.

2.3.2 Comparison to WhatsHap's edit distance-based genotyping

Correct allele detection from individual reads plays a pivotal role in genotyping. So, in order to test *k*-merald's performance, we performed WhatsHap genotyping (`whatshap genotype`) in the following two settings:

- using WhatsHap's original implementation using edit distance-based allele detection
- using *k*-merald for allele detection

and compared the genotyping performance across them. We based our evaluation on GIAB samples, HG001 (NA12878) and HG002 (NA24385) and the GIAB v4.2.1 high confidence benchmark callsets [130]. The genotyping was performed using various coverages of ONT Ultra-long, PacBio CLR and PacBio HiFi sequencing reads. We used a *k*-mer value of $k=7$, variant window $w=25$, gap probability $=10^{-4}$, that is, a cost value of "40" and $\epsilon=0.15$, for the genotyping results presented in this chapter. Before selecting a specific value, we performed a comparison of genotyping error rates across multiple values of *k* and observed that *k*-merald's performance stayed quite consistent across them, as shown in Figure 2.5. To evaluate genotyping performance, we calculated genotype concordance, that is, the percentage of variants genotyped correctly. Additionally, we used RTG Tools "vcfeval" [131] to calculate precision, sensitivity and F_1 score for the predicted genotypes. Finally, we used GIAB

Table 2.2: Genotyping performance for HG002

	GT-concordance (%)	Precision (%)	Sensitivity (%)	F_1 score(%)
ONT-UL				
WhatsHap	95.22	93.57	94.34	93.95
k-merald	96.08	97.46	95.19	96.31
PacBio CLR				
WhatsHap	96.32	97.65	93.97	95.78
k-merald	97.24	97.89	94.90	96.37
PacBio HiFi				
WhatsHap	99.70	99.75	98.77	99.26
k-merald	99.67	99.78	98.74	99.26

v3.0 stratifications (Appendix A) to compare the genotyping performance in difficult-to-map and low-complexity regions of the genome. Following sections stratify the genotyping performance from multiple aspects.

Genotyping performance for ONT sequencing reads

We first evaluated the genotyping performance using ONT sequencing reads for samples HG001 and HG002. Considering SNPs and indels together, we observed that genotyping using *k*-merald for allele detection shows an improved performance in comparison to WhatsHap's genotyping results based on the conventional edit distance-based allele detection approach. For 54× HG002 ONT sequencing reads, the genotype concordance improved from 95.22% to 96.08%, indicating an 17.99% decrease in error rate (Table 2.2, Figure 2.6). Precision, sensitivity and F_1 score values also depict this improvement (Figure 2.6A). To assess the robustness across different error profiles, we also evaluated the genotyping performance for sample HG001, using the error profiles trained using ONT sequencing data for HG002. A similar trend was observed for the 34× HG001 ONT sequencing reads, with genotype concordance improving from 92.78% to 94.18% indicating a 19.39% decrease in error rate (Figure 2.6A). This consistent improvement in genotyping performance seen while using different samples for training and testing confirms that the characteristics of error profiles captured by *k*-merald are not sample-specific. Thus, an error profile generated using one sample can be readily used for genotyping multiple samples with sequencing data generated from the same source.

Genotyping performance individually for SNPs and Indels

Furthermore, we evaluated the genotyping performance individually for SNPs and indels. For HG001, we observed 50.00% decrease in error rate for SNPs and 8.47% for indels. For HG002, the percentage decrease was 55.56% and 18.52%, for SNPs and indels, respectively (Figure 2.6B).

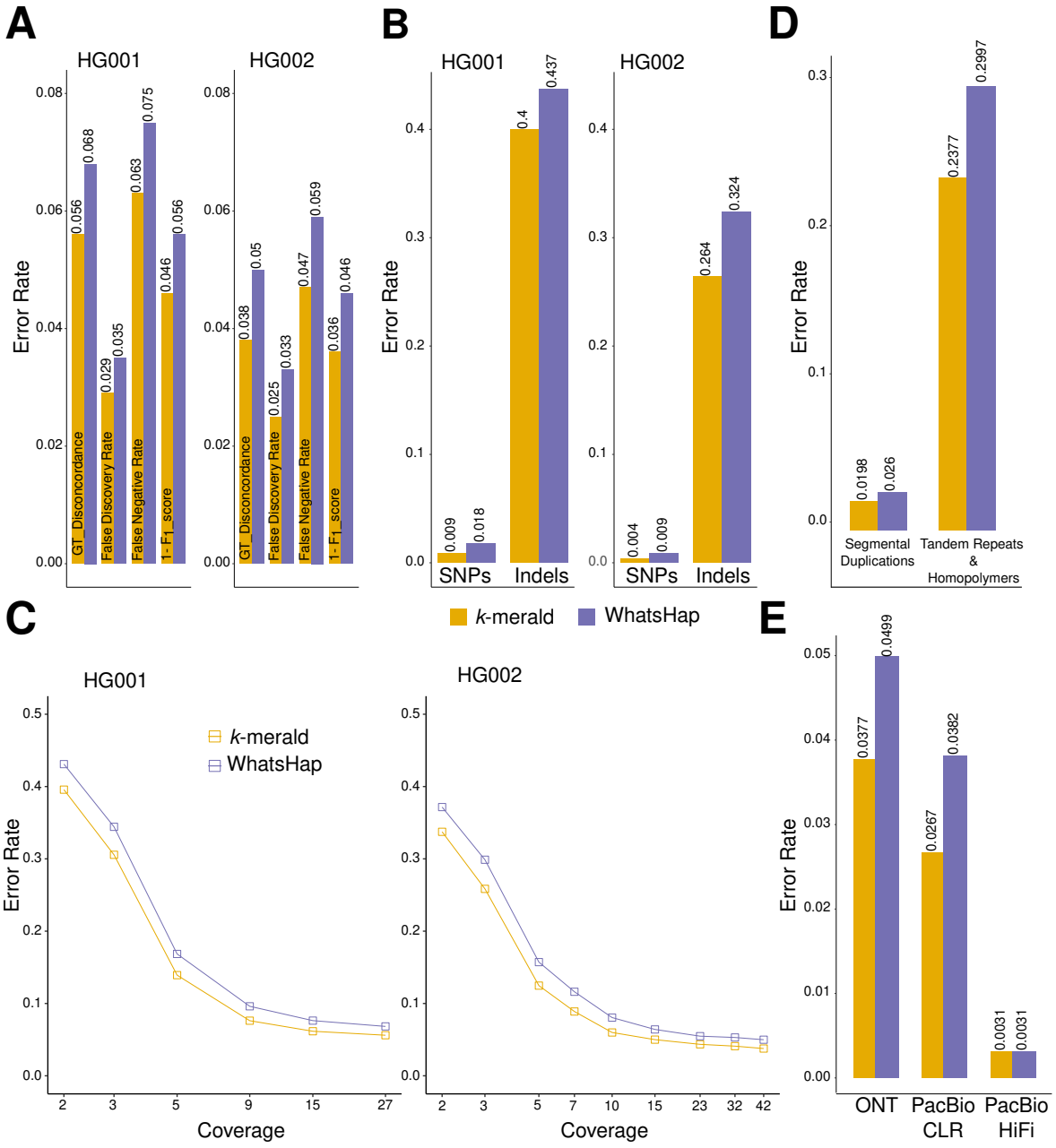


Figure 2.6: Genotyping performance comparison. **A**. Genotyping performance comparison between WhatsHap with conventional edit distance-based allele detection and *k*-merald, using ONT reads for sample HG001 and HG002. **B**. Genotyping performance comparison for SNPs and indels, individually, using ONT sequencing reads. **C**. Genotyping performance comparison across multiple coverages of ONT sequencing reads. **D**. Genotyping performance comparison across multiple genome stratifications using ONT sequencing reads for sample HG002. **E**. Genotyping performance comparison, individually for ONT, PacBio CLR, and PacBio HiFi data, for sample HG002.

Genotyping performance across multiple coverages

We reasoned that the negative impact of sequencing errors on allele detection might become even more prominent at low coverage, and therefore evaluated the genotype performance across multiple coverages of sequencing reads. For HG002, we downsampled the ONT data to coverages ranging from 3× to 54×. For HG001, we downsampled the ONT data to coverages ranging from 3× to 34×. For both these samples, we observed, in line with our hypothesis, that although *k*-merald outperforms the conventional allele detection algorithm at all coverages, the absolute difference becomes more pronounced at lower coverages (Figure 2.6C).

Genotyping performance individually for different genome stratifications

Additionally, we compared the genotyping performance, for sample HG002, in low mappability SDs as well as low complexity regions like tandem repeats (dinucleotide, trinucleotide and quadnucleotide STRs and simple repeats) and homoploymers (perfect homopolymers >6bp and imperfect homopolymers >10bp). We observed that across all these regions, *k*-merald gives better genotyping performance than the conventional edit distance-based genotyping with 21% decrease in error rate for tandem repeats and homopolymers and 24% for segmental duplications (Figure 2.6D).

Genotyping performance across multiple platforms

Furthermore, to evaluate performance across different sequencing platforms, we evaluated the results obtained by using PacBio CLR and PacBio HiFi sequencing reads. For 20× HG002 PacBio CLR sequencing reads, the genotype concordance improved from 96.32% to 97.24% indicating a 25.00% decrease in error rate (Table 2.2 and Figure 2.6E). For 35× HG002 PacBio HiFi sequencing reads, we observed very similar genotyping performance from both approaches (Table 2.2, Figure 2.6E). This supports the hypothesis that our method provides a particular advantage for more error-prone sequencing reads.

Comparison of genotype quality

Lastly, we hypothesized that using our sequencing error profiles would also improve the process of estimating genotype quality values, particularly for indels. That is, our method is better able to assess the reliability of genotypes and to express it as genotype quality provided along with the genotypes, which is potentially beneficial for downstream applications. To evaluate this, we compared the genotype quality between *k*-merald and edit distance-based WhatsHap genotypes for GIAB v4.2.1 whole genome high confidence indels, genotyped using 54× ONT data for sample HG002. We observed that the correct *k*-merald based genotypes tend to be of higher genotype quality as compared to the correct genotypes obtained using WhatsHap's genotyping using edit distance-based allele detection. In total,

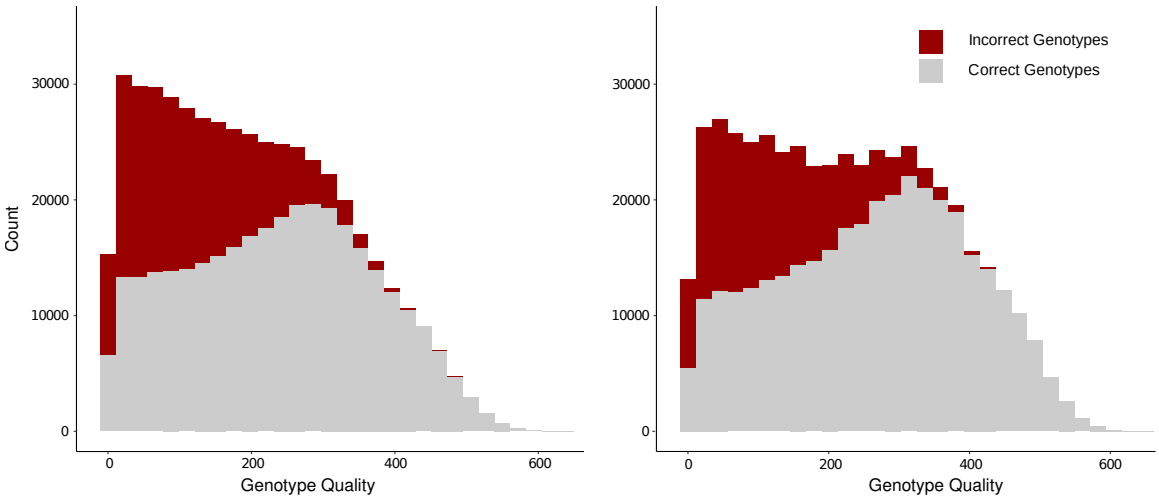


Figure 2.7: Genotype quality improvement. A comparison of whole genome indels genotype quality between WhatsHap using *k*-merald (left) and WhatsHap using edit distance-based allele detection (right) for sample HG002 using ONT sequencing data.

67% of the correct *k*-merald based genotypes exhibited a genotype quality of at least 200, while this percentage was 60% for edit distance-based genotypes. For all genotypes with a quality of at least 200, the percentage of correct genotypes was 89% for *k*-merald while 85% for WhatsHap's original implementation (Figure 2.7).

Runtime comparison

For 54× ONT reads, generating genome-wide error profile took about 145 CPU hours collectively. Whole genome genotyping collectively took about 29 single-core CPU hours using `whatshap genotype` with conventional edit distance-based allele detection, while about 139 single-core CPU hours using `whatshap genotype` with *k*-merald. We attribute the increased run time to the more involved bookkeeping for working with *k*-mers, as shown in Algorithm 2 , compared to the single-nucleotide sequence alignment. However, it should be noted that both training and alignment steps can be performed in parallel in a chromosome-wise manner. Given the runtime of read alignment that happens before genotyping, we do not consider this increased runtime to be the main bottleneck in processing a long-read data set.

2.3.3 Comparison with PEPPER

As mentioned before, to our knowledge, there is no tool designed specifically for long-read-based genotyping of a variant callset other than WhatsHap, so a one-to-one performance comparison with another tool was not possible. However, we aimed to compare our approach to the state-of-the-art tool PEPPER [129], which detects candidate variants, genotypes, and phases them in an integrated workflow. Comparing a genotyper's performance to such an integrated variant caller is not a straight-forward process. To avoid a skewed com-

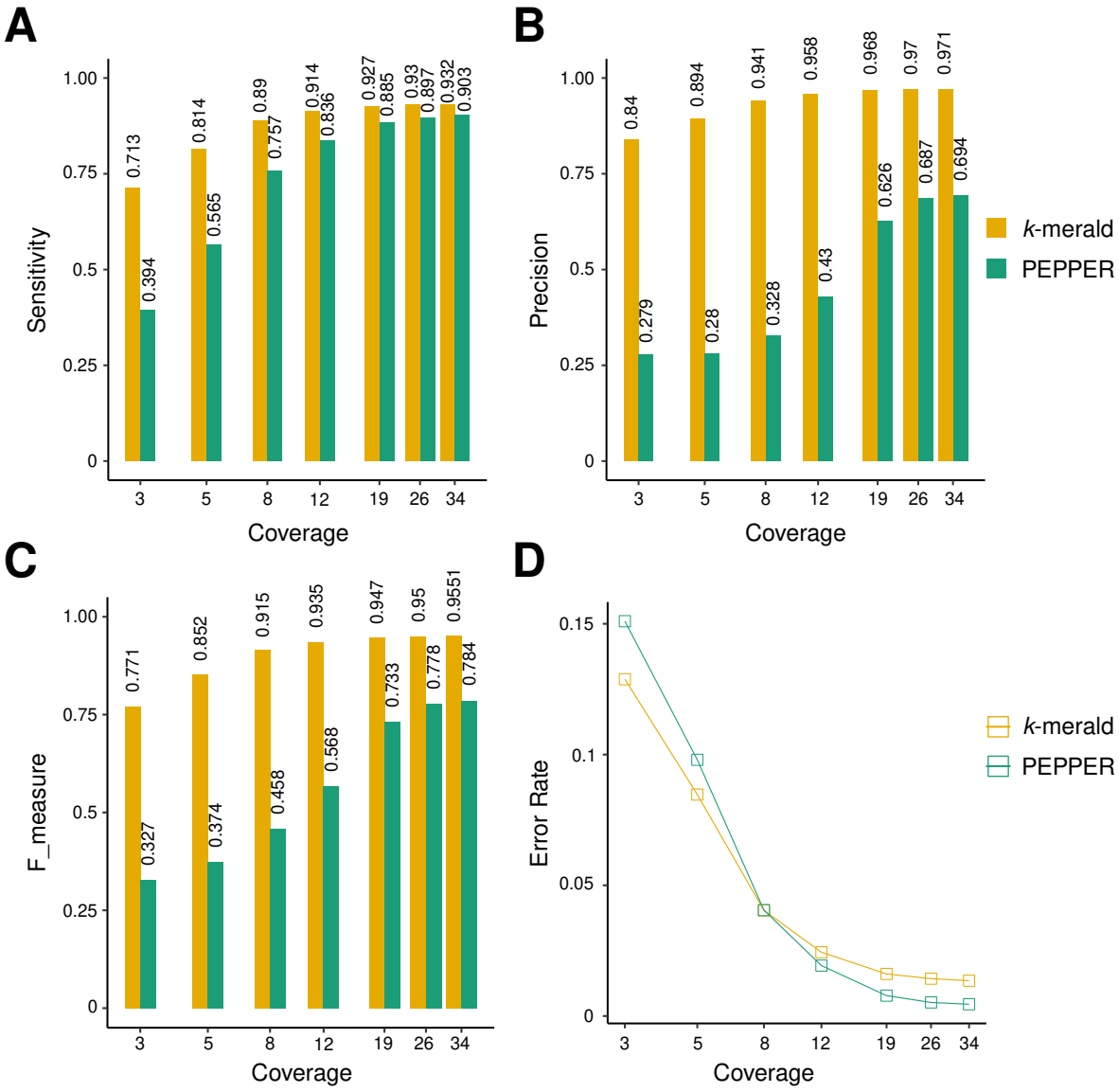


Figure 2.8: Comparison of genotyping performance between *k*-merald and PEPPER. Comparison of genotyping performance between *k*-merald and PEPPER for sample HG001, across multiple coverages of ONT sequencing reads, using high confidence GIAB v4.2.1 genotypes as ground truth. For *k*-merald-based genotyping, we used error profiles generated for HG002. **A**. Comparison of genotyping sensitivity. **B**. Comparison of genotyping precision. **C**. F_measure comparison. **D**. Error rate comparison between *k*-merald and PEPPER. The comparison was restricted to variants common between *k*-merald and PEPPER callsets and high confidence GIAB v4.2.1 genotypes were used as ground truth.

parison, we performed this comparison in two ways. Firstly, we computed precision, recall, and F_1 score for all the variants called/genotyped by each method in their respective default mode. That is, *k*-merald-based genotyping is performed with the set of all variants given as input, while PEPPER runs both discovery and genotyping. We performed this comparison using multiple coverages of ONT reads for sample HG001, while using the error profiles for HG002. For all these measures, we observed that *k*-merald based genotyping showed better results as compared to PEPPER at all coverages (Figure 2.8 A,B,C). However, it should be noted that PEPPER had to perform the additional step of variant discovery before genotyping. Therefore, this evaluation method could potentially favor the genotyper. To address this, we additionally computed genotype concordance only for the variants common between GIAB v4.2.1 callset and the PEPPER callset. Even though this method of comparison favors PEPPER, as we restrict our evaluation only to the variants that could be called by the variant caller, we observed that *k*-merald-based genotypes still had lower error rate as compared to PEPPER at low coverage (Figure 2.8D).

2.4 Discussion

Correct detection of alleles carried by sequencing reads is vital for variant genotyping and haplotype phasing [132]. In comparison to short reads, long reads span larger regions, hence providing more information, especially in highly repetitive regions of the genome. However, sequencing errors generated by long-read sequencing technologies pose a challenge for allele detection. The sequencing error profiles vary across multiple sequencing technologies such as ONT, PacBio CLR, and PacBio HiFi. That includes different error distributions as well as different characteristics of sequencing errors (Figure 2.2, 2.3). The conventional allele detection methods are mostly based on edit distance between sequences, which penalizes all sequence mismatches equally. We hypothesized that instead of fixed costs, using technology-specific sequencing error profiles for determining alignment costs can provide more insights to distinguish a variant allele from a sequencing error, hence improving the allele detection accuracy. To achieve this, we developed *k*-merald, an allele detection method that generates technology specific *k*-mer-based error profiles by traversing aligned sequencing reads in the non-variant regions of the genome. Furthermore, *k*-merald employs a *k*-mer based alternative to global sequence alignment which instead of aligning the sequences of base pairs, aligns strings of consecutive *k*-mers generated from the respective sequences, while using the generated error profiles for alignment cost calculation.

We observed that WhatsHap genotyping using *k*-merald showed better genotyping performance as compared to the original WhatsHap implementation, which detects alleles using edit distance-based sequence alignment. We observed 18% and 25% decrease in genotyping error rate for $54\times$ ONT and $20\times$ PacBio CLR sequencing reads, respectively. The genotyping performance, however, was similar for PacBio HiFi sequencing data, potentially due to its lower error rate compared to ONT and PacBio CLR. While evaluating the genotyping perfor-

mance individually, we observed a 56% decrease in error rate for SNPs while 18% for indels, for sample HG002. A comparison of genotyping performance across multiple coverages of ONT data revealed that the improvement in genotyping performance shown by *k*-merald becomes even more prominent at low coverages.

At present, ONT is the most cost-effective long-read sequencing platform in terms of cost per sequenced base pair. But this comes at the disadvantage of increased and more systematic sequencing errors. *k*-merald attempts to solve this problem and provides substantial improvements in allele detection in order to push genotyping performance to its limits. Of note, the use of error models trained for a given sequencing data set provides a way to take technology-specific differences into account when computing genotype likelihoods, hence allowing us to quantify uncertainty in a more informed way. This is reflected in our results showing that variants genotyped with high genotype quality above 200 are more strongly enriched for correct genotypes when using *k*-merald. Our training procedure exploits the similarity of a sequenced sample and the reference genome by using variant-free regions for training. In this way, our model can be readily retrained even on a single data set, which potentially allows it to adapt to subtle differences such as version of the sequencing chemistry and other batch effects. Because the learning procedure is technology-agnostic, we anticipate that our method can readily be applied to future long-read data types.

Chapter 3

Genotyping and validation of inversions and copy number variations using Strand-seq data

This chapter introduces ArbiGent, a tool for genotyping inversions and copy number changes in arbitrary regions of the genome, utilizing Strand-seq data. I co-developed this tool with Wolfram Höps, building upon the statistical framework established by Sascha Meiers and Maryam Ghareghani [37]. ArbiGent was first introduced as a Strand-seq based inversion genotyper in the Human Genome Structural Variation Consortium (HGSVC) study published in Science in 2021 [12]. Subsequently, after Wolfram and I further refined its performance, it was used as the primary tool for inversion genotyping and validation in an HGSVC companion study published in Cell in 2022 [13], where I share co-first authorship with David Porubsky and Wolfram Höps. This chapter mainly focuses on this study, as it encompasses the majority of the work related to refining and validating ArbiGent's performance, as well as utilizing its full potential to develop a comprehensive, multi-platform-based catalog of inversions in the human genome. The related work contributed by co-authors is also discussed in this chapter and clearly acknowledged. In addition to the two studies mentioned above, ArbiGent has also served as the primary tool for inversion genotyping and callset unification in both the study published in Genome Biology in 2023 [14], and the study provisionally accepted for publication in Nature in 2025 [15]. This chapter re-uses relevant text, tables and figures from these publications, clearly acknowledging co-author contributions. For publication details and author contributions, please refer to Sections C.2, C.3, C.4 and C.5.

3.1 Introduction

As discussed in detail in Section 1.5.2, inversions remain an under-explored class of genetic variation, largely due to the presence of flanking SDs often exceeding sequencing read

lengths, making their discovery challenging. One of the early efforts to characterize the full spectrum of human genetic variation was carried out by Sudmant et al. in 2015 [10]. They constructed an SV catalog using 2,504 human genomes from the 1KGP, including 786 inversions. However, since Illumina reads were the primary source of SV discovery in this study, with limited support from PacBio reads, all reported inversions were shorter than 100 kbp. In 2019, Chaisson et al. developed a multi-platform SV discovery approach, under the umbrella of the HGSVC, offering a more refined structural variation landscape [11]. They reported 308 inversions, 74% of which were either primarily discovered or additionally supported by Strand-seq. The use of Strand-seq also enabled the detection of inversions longer than 100 kbp, as it identifies inversions exclusively through strand switches along the chromosomes [11].

As mentioned in Section 1.5.2, the higher efficiency of Strand-seq in inversion discovery, compared to other technologies, is mainly due to its ability to detect inversions in the genome regardless of the length of the flanking repeats [11, 133]. However, there are only a few computational tools specifically designed for Strand-seq data, and even fewer that are tailored for inversions. Existing computational tools that aid in identifying structural variants using Strand-seq data include breakpointR [134]. This tool has contributed to Strand-seq based inversion discovery in several studies [12], [13], [14]. breakpointR leverages template strand switches as markers to estimate SV breakpoints. However, it functions solely as a breakpoint estimation tool and does not provide information about the structural variation class of the identified region. In other words, breakpointR can provide an estimated location of an SV but cannot determine the exact SV type. Another Strand-seq based tool, primarily developed as a somatic SV caller, is MosaiCatcher [37]. MosaiCatcher employs a comprehensive and integrated workflow designed for detecting large SVs (>100 kbp) in somatic cells. Although this workflow performs quite well for calling somatic variations and can aid in discovering subclonal structural variation in different diseases, for example, cancer, it is unable to detect smaller events. Moreover, it performs SV detection in genomic bins of fixed size and does not provide genotype likelihoods for user-defined arbitrary regions.

To address these limitations, we developed ArbiGent—**Arbi**trary segment **Geno**typer—that utilizes the statistical framework of MosaiCatcher [37] while extending it to estimate SV genotype likelihoods for arbitrary genomic segments provided as input [12, 13]. ArbiGent determines an individual's genotype for a specific locus by integrating single-cell information derived from MosaiCatcher's statistical framework. Another key improvement is the implementation of mappability-based read-count normalization strategy, which improves sensitivity in difficult-to-map regions. Additionally, ArbiGent facilitates the unification of inversion calls generated from multiple data sources across all samples included in the genotyping cohort. ArbiGent was first utilized for genotyping and refining the inversion callset presented in the 2021 study by Ebert et al. [12]—a major follow-up effort by the HGSVC after Chaisson et al. [11]—aimed at enhancing the understanding of structural variation in human genomes. This study incorporated fully phased assemblies for 32 human samples as

an additional orthogonal support for SV discovery and classification. The reported inversion callset, comprising 316 inversions, was constructed using support from Strand-seq, phased assemblies, and Bionano optical mapping [12]. While this resource substantially advanced the known inversion landscape of the human genome, it became clear that extensive curation, refinement, and a dedicated in-depth analysis of the callset were beyond the scope of that study. This realization led to the formation of an inversions-working-group under the umbrella of the HGSVC. One of the main goals of this group was to generate a curated and extensive catalog of inversion polymorphisms in the human genome. This chapter covers some of the efforts dedicated to achieving this goal, including multi-platform inversion discovery, validation and genotyping of the merged callset using ArbiGent and manual refinement of inversion breakpoints through fully phased assemblies. Another major aim of this study was to investigate inversion polymorphisms at a population scale, with a particular focus on the relatively understudied phenomenon of inversion recurrence and its potential role in disease associations. The work directed towards achieving that objective is discussed in Chapter 4.

3.2 ArbiGent

This section reuses material from [13] and introduces ArbiGent, co-developed by me and Wolfram Höps. The mappability-based normalization approach described in Section 3.2.1 was developed by me while the workflow for producing sample genotype likelihoods from single-cell estimates described in Section 3.2.3 was developed by Wolfram Höps.

3.2.1 Leveraging uniqueness of a region to normalize read counts

While Strand-seq outperforms other contemporary data types in addressing the challenging task of inversion discovery and genotyping, read alignments in difficult-to-map genomic regions still contain some ambiguity. Due to the highly repetitive and complex nature of these regions, short reads cannot be uniquely mapped, leading to low read counts and, consequently, incorrect low-copy-number predictions. Since inversions predominantly occur in highly repetitive and SD-rich regions, the Strand-seq read counts in these regions cannot be taken at face value when determining the genotype. To address this issue, we developed an approach that generates a "mappability track" for the reference genome, taking into account the characteristics of Strand-seq reads. This track quantifies uniqueness, that is, the extent to which the reads can be accurately aligned across the entire reference sequence. The mappability of a region is then used to normalize the Strand-seq read counts before downstream analysis. This normalization approach, also visually represented in Figure 3.1, operates as follows:

1. Simulated 100 bp paired-end reads—mimicking the characteristics of Strand-seq reads—are generated for each position across the reference genome using its nucleotide sequence.

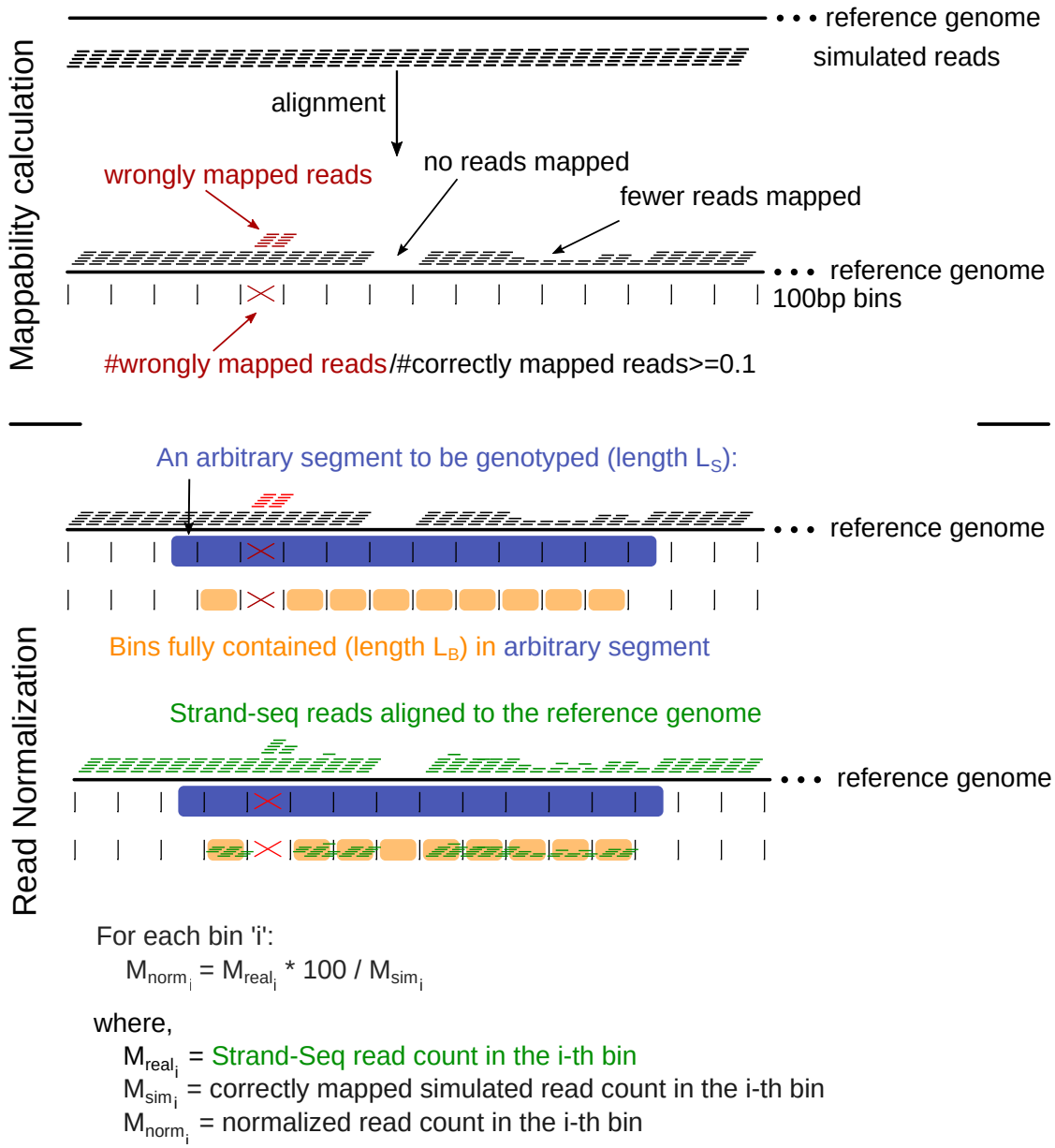


Figure 3.1: Mappability based Strand-seq read count normalization. A schematic representation of a read-simulation-based approach devised to determine the mappability of a genomic region, later used for normalizing Strand-seq read counts before genotyping. **Mappability Calculation:** Simulated 100 bp paired-end reads are generated for each position across the reference genome. These read pairs are then aligned to the reference genome using the same alignment settings as used for the real Strand-seq data. Resulting alignments are evaluated to quantify the mappability of each 100 bp region, defined as the fraction of correctly mapped reads. Bins where the number of spurious alignments exceeds the correctly mapped ones by a factor of 0.1 or greater are discarded. **Normalization:** For each arbitrary segment to be genotyped, the Strand-seq reads coming from each 100 bp bin contained in the segment are normalized individually using the mappability-based normalization factor and aggregated.

2. The simulated reads are then aligned to the genome of origin using the same settings applied to real Strand-seq data to be used for genotyping, in order to minimize alignment bias.

3. In sequence alignment terms, a region's mappability is defined as the proportion of reads that originate from that region and correctly align to their source. Keeping that in view, the alignments of simulated reads are analyzed in a bin-wise manner, recording the number of correct and erroneous alignments within each 100 bp bin. Bins where erroneous alignments exceed correct ones by a factor of 0.1 or more are flagged as "invalid" and excluded from further analysis.

4. The normalization factor for the *i*-th 100 bp bin, NF_i, is calculated as follows:

$$NF_i = \frac{100}{M_{sim_i}} \tag{3.1}$$

where, M_{sim_i} is the number of correctly mapped simulated reads belonging to the *i*-th bin. To normalize the read counts for a specific segment, the normalization factor is applied individually to the read counts of each 100 bp bin within that segment, that is,

$$M_{norm_i} = M_{real_i} \cdot NF_i \tag{3.2}$$

where, M_{real_i} is the number of observed real reads and M_{norm_i} is the normalized read count for the *i*-th bin. The normalized read counts for each bin contained in the segment are aggregated to get the normalized read count for the whole segment, M_{norm}.

$$M_{norm} = \sum_{i=1}^{n} M_{norm_i} \tag{3.3}$$

where, *n* is the total number of bins contained in the respective segment.

Notably, for read count normalization, only those bins that are fully contained within the respective segment are considered. For example, if a segment spans from 73 bp to 745 bp, the first bin would range from 100–200 bp, and the last from 600–700 bp. As a result, some reads at the segment boundaries may be excluded. However, due to the small bin size, the downstream impact is expected to be minimal.

3.2.2 MosaiCatcher

The normalized strand-specific read counts are fed into MosaiCatcher's [37] statistical framework to compute likelihoods for various possible SV genotypes for each individual cell of a sample. While a full description of MosaiCatcher's workflow is beyond the scope of this thesis, a brief summary of the main algorithmic steps is stated as follows:

The workflow starts with recording the number of Watson (W) and Crick (C) Strand-seq reads per 100 kbp bin for each cell. Following read counting (and normalization when

working in ArbiGent mode), a joint segmentation using read-depth data across all single cells of a sample is performed to detect potential SV boundaries. StrandphaseR [46] then utilizes heterozygous SNP positions to construct consensus haplotypes, which are then used to haplotag individual reads in each cell in order to facilitate haplotype-resolved SV calling. As described in Section 1.5.2 and Figure 1.8B–E, SCEs or SVs can lead to varying strand states across different segments on the same chromosome. For a segment-wise classification of the SV (or reference) state, MosaiCatcher employs a Bayesian model which models the expected read coverage using a negative binomial (NB) distribution individually for each (W, C) strand. The joint NB distributions are used to determine the likelihoods of different strand states for each segment. These strand states, when compared to the ground state—strand state observed across majority of the chromosome—can be translated to the SV type depicted by each segment (Section 1.5.2, Figure 1.8). Using this strategy, the SV genotype log-likelihoods for a particular segment individually for each Strand-seq cell are generated.

MosaiCatcher calculates likelihoods for different haplotype configurations considering both orientation and copy number changes while allowing a maximum copy number value of four. To consolidate orientation and copy number information into a single configuration, instead of using the standard "1,0" genotype notation, which represents only one variant type, MosaiCatcher employs a four-digit notation, "ABCD". In this notation, "A" and "B" denote the observed copy number for the first haplotype in direct and reverse orientations, respectively, while "C" and "D" represent the same for the second haplotype. For example, "1001" indicates that the second haplotype is inverted, whereas "0210" represents an inverted duplication on the first haplotype.

3.2.3 Genotyping and filtering

To determine the genotype of a sample for a specific region of interest, we developed a workflow that aggregates the genotype log-likelihoods generated by MosaiCatcher, across all cells within the sample, producing a combined genotype likelihood estimate. For high quality genotype prediction, only regions with at least 500 bp of uniquely-mappable sequence—defined as having at least 75% of simulated reads aligning correctly (as described in Section 3.2.1)—are considered. Furthermore, genotype calls are classified as "high confidence" only if they exhibit a likelihood ratio greater than 10^3 compared to the reference state ("1010").

When used for genotyping a cohort of samples, ArbiGent also generates a label for each of the genotyped locus based on the sample genotypes. Some of these labels help with population-based labeling of false positives or potentially complex loci, including:

- **False Positive**: no SV detected for any of the samples, that is, "reference" genotype observed across the whole cohort.
- **Always Complex**: no "simple" (heterozygous or homozygous inversion genotypes) and no reference calls observed in any of the samples.

- **Misorient**: putative misorient in the reference genome indicated by the locus being genotyped as a homozygous inversion across all samples.
- **Inverted Duplication**: inverted duplication observed in at least one sample.
- **Low Confidence**: less than 500 bp region with good mappability ($\geq$75%), hence, making the read alignments in this region unreliable for confident genotype predictions.
- **Mendel Fail**: the (parent-child) trio genotypes failing the Mendelian consistency test.

3.3 Performance Evaluation and Applications of ArbiGent

This section reuses material from [13] and describes the collaborative work conducted in this study to generate a multi-platform based inversion callset. The inversion discovery described in Section 3.3.1 was performed by Ashley Sanders, David Porubsky, Peter Audano and Feyza Yilmaz. ArbiGent genotyping and refinement of the inversion calls described in Section 3.3.2 and evaluation of genotypes described in Section 3.3.3 was conducted by me and Wolfram Höps with input from co-authors.

3.3.1 Multi-platform-based inversion discovery

The first step of this integrated workflow was the discovery of inverted regions across the human genome using a cohort of 44 samples from the 1KGP [5], representing diverse ancestries: African (13), American (8), East Asian (9), European (8) and South Asian (6). All the analyses in this study were performed using GRCh38 as the reference genome. Since inversion detection is inherently challenging, as detailed in Section 1.5.2, we employed three complementary approaches for this task:

1. Strand-seq-based inversion discovery using breakpointR [134] to detect strand switches, followed by manual curation and refinement through visual inspection of Strand-seq data.
2. Haplotype-resolved-assembly-based inversion discovery using the phased assembly variant caller, PAV [135].
3. Bionano optical mapping [136] based inversion discovery.

Following the initial discovery and refinement at the individual callset level, SV-Pop [12, 137] was used to merge the identified loci to generate a non-redundant inversion callset. This process resulted in a provisional callset of 615 inverted loci.

3.3.2 Genotyping and filtering

After generating a consolidated, multi-platform supported set of candidate loci, ArbiGent was employed to genotype and validate the inversion calls. The 615 loci were genotyped

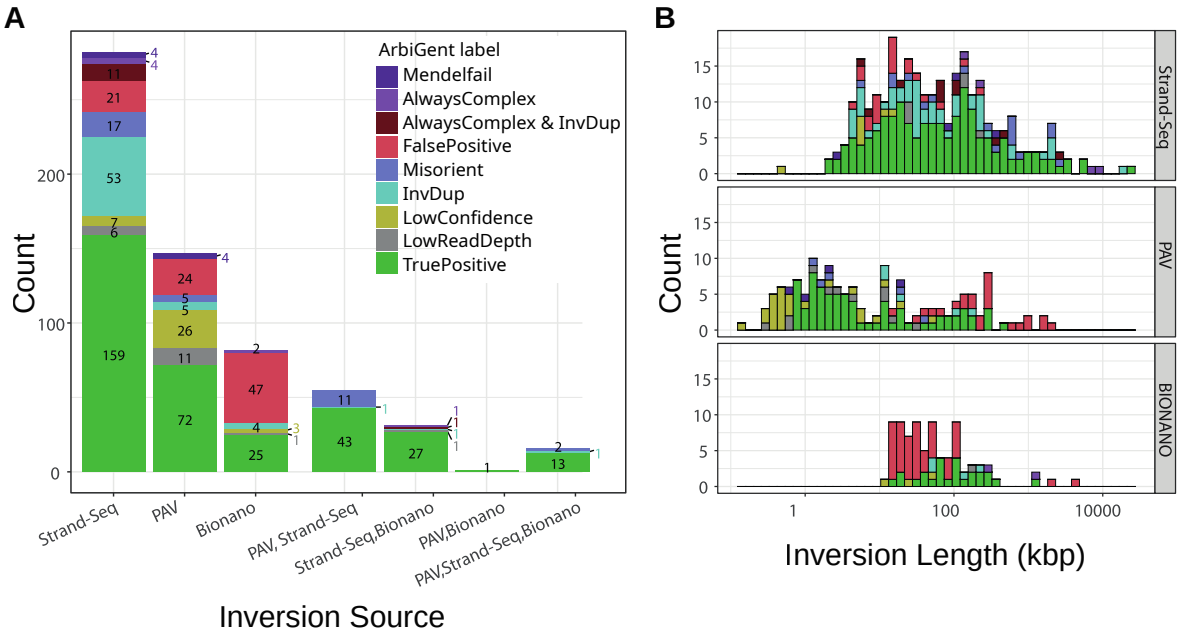


Figure 3.2: Inversion genotyping and characterization. A. ArbiGent genotypes stratified by source technology, showing Strand-seq contributing most of the loci, followed by PAV and Bionano. **B.** Stratification of ArbiGent genotypes based on the length of the event across multiple source platforms. Figure created by Wolfram Höps [13].

across all 44 samples using Strand-seq data and the population-based labels assigned by ArbiGent, as described in Section 3.2.3, were used to filter out false positives. Loci labeled as "False Positive", "Always Complex" or "Inverted Duplication" by ArbiGent were excluded from further analysis. Additionally, inversion calls with $\geq$90% reciprocal overlap were merged, reducing the total number of inversions to 418. Stratification based on inversion source and length reflected expected technological biases: Strand-seq emerged as the dominant method for identifying inversions, particularly longer ones. PAV was more sensitive at detecting shorter inversions while Bionano appeared to be the least sensitive technique overall (Figure 3.2). Notably, inversions supported by multiple technologies were predominately characterized as true events, emphasizing the importance of using orthogonal approaches for both discovery and validation of inversion calls (Figure 3.2A).

To refine inversion breakpoints, we turned to haplotype resolved assemblies and analyzed the dotplot alignments of phased assemblies from Ebert et al. [12] against the GRCh38 reference genome. Only inversions where both breakpoints were spanned by one contig (183/418 inversions, 44%) were analyzed using this strategy to annotate breakpoints with greater precision. The resulting callset, now with refined breakpoints, was re-genotyped using ArbiGent. Since phased assemblies enhance PAV's sensitivity for detecting small inversions (typically $<$5 kbp), we relied on PAV's genotypes for these smaller events in cases where ArbiGent could not genotype them with high confidence.

After another round of ArbiGent-label-based filtering, similar to the initial iteration, we obtained a non-redundant, callset of 399 inversions, characterized as follows:

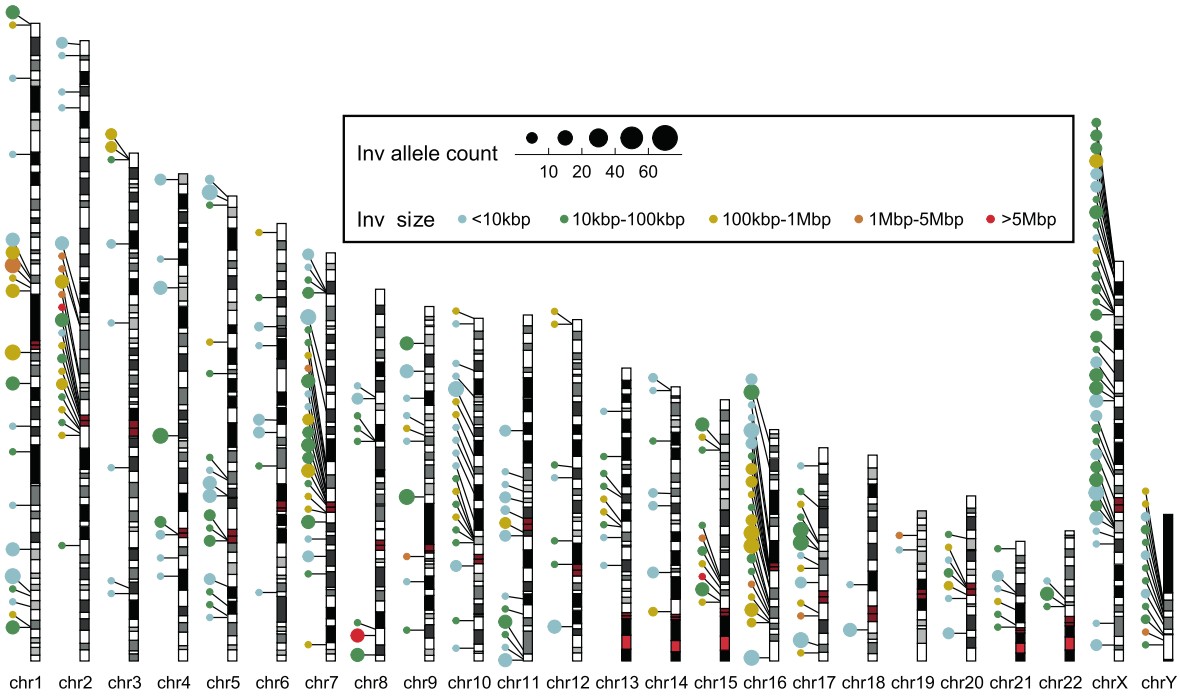


Figure 3.3: Genomic landscape of balanced inversions. Landscape of 292 balanced inversions included in the final callset representing length and allele frequency distribution across chromosomes and highlighting inversion hotspots. Figure created by David Porubsky [13].

- 292 were classified as "balanced" inversions, meaning at least one sample in the genotyped panel showed a homozygous or heterozygous inversion without accompanying copy number changes.
- 40 were identified as inverted duplications.
- 29 were labeled as structurally complex loci—that is, regions that were difficult to genotype or showed copy number variation.
- 38 were flagged as potential misorients in the GRCh38 assembly.

An overview of the genomic landscape of the balanced inversions, shown in Figure 3.3, highlights distinct "inversion hotspots". The most prominent hotspots were observed on chromosomes 2, 7, 10, 16, and X, predominantly located near SD-rich centromeric satellite regions. Additionally, chromosomes 1, 2, 7, 10, 15, 16, and 17 showed a notable enrichment of large inversions (>100 kbp), which were frequently accompanied by SDs. These findings align with expectations, as SD-rich regions create an ideal environment for NAHR, which is believed to be the primary mechanism driving inversion formation.

Chaisson et al. 2019

ArbiGent	1/1	0/1	0/0
1/1	25	6	1
0/1	1	40	0
0/0	0	13	48

Figure 3.4: Genotype concordance evaluation. Comparison of ArbiGent genotypes with 134 overlapping inversions ($\geq$5 kbp) reported by Chaisson et al. [11].

3.3.3 Performance Evaluation

Comparison with other callsets

We assessed the genotyping performance by computing the genotype concordance between ArbiGent genotypes and the genotypes reported by Chaisson et al. [11]. We selected the sample HG00512 for this comparison and observed a genotype concordance of 84% (113/134, $\geq$5 kbp loci overlapping between the two callsets). Most discrepancies (19 out of 21 cases) were caused by ArbiGent classifying the locus as reference or homozygous inversion, while Chaisson et al. reporting a heterozygous inversion (Figure 3.4). In order to improve the accuracy and resolution of genomic analysis, particularly for identifying and phasing SVs, sequencing data from multiple single-cell libraries from the same donor is typically aggregated together into "composite" files [133]. Hanlon et al. [138] reported that alignment errors and assembly collapses may lead to misinterpretations of heterozygous inversions in the "WW" composite files used by Chaisson et al. [138]. This observation indicates that the 19 discrepant inversions might in fact be wrongly genotyped by Chaisson et al. [11] and the genotype determined by ArbiGent is the correct one as it considers all ("WW","CC","WC","CW") libraries for genotype prediction. Considering this, the observed genotype concordance supports ArbiGent's reliability, indicating no apparent biases in its genotyping performance.

Mendelian Consistency

The labels generated by ArbiGent, taking population-based genotypes into account, not only provide insights into the structural variation exhibited by a locus but also serve as a quality control measure for evaluating the genotyping performance. For instance, the "Mendelfail" label is generated based on the Mendelian consistency test for parent-child trios included in the genotyped panel. ArbiGent also classifies the individual genotypes as "simple/balanced" (homozygous/heterozygous inversion or reference genotype) and "complex" (copy number variation). Only loci where all trio members exhibit a simple genotype are tested for Mendelian inheritance. Our sample panel included three trios. Among the 399 inversions reported in our integrated callset (Section 3.3.2), 260 showed exclusively "simple" geno-

types across all of the trios. Of these, 95% (247/260) were Mendelian consistent. When loci with at least one low-confidence genotype were excluded, this percentage increased to 99.5% (200/201).

Hardy-Weinberg Equilibrium

Several inversion sites in our callset exhibited more than two distinct allelic conformations across samples. This complexity made direct Hardy-Weinberg equilibrium (HWE) testing of the raw genotypes challenging. To address this, we implemented a preprocessing strategy to convert multi-allelic sites into "pseudo bi-allelic" ones. This process involves the following steps:

1. For each locus, a list of unique alleles along with their occurrence counts across all samples is computed.
2. The most frequent allele is designated as the "major allele" while all other alleles are grouped under the "minor allele" category.
3. All the genotypes are then transformed by encoding the major allele as "1" and those belonging to the minor allele category as "0".
4. The heterozygosity for each locus is thus determined based on the major allele, that is, only genotypes carrying the major allele and one of the alleles from the minor allele category are considered as heterozygous.

After transforming the genotypes using this approach, VCFtools (--hardy) [139] was used to perform HWE test on 275 autosomal inversion sites with no missing genotypes. We applied the Benjamini-Hochberg correction [140] to correct the *p*-values for multiple testing. In total, 224 out of 275 loci (81.45%) passed the HWE test, confirming the high quality of ArbiGent's genotypes. When stratifying the test results based on ArbiGent labels (Figure 3.5), we observed that most loci failing the HWE test fell into the complex category. In contrast, most simple (balanced) inversions passed the test while closely following the theoretical 2pq curve (representing the expected frequency of heterozygous individuals).

Robustness to the number of Strand-seq cells available for genotyping

Since the number of "good quality" Strand-seq cells available for a given sample can vary (as discussed in Section 1.5.2), we conducted a benchmarking experiment to assess how ArbiGent's genotyping performance is influenced by the number of cells used for genotyping. For this downsampling experiment, we selected sample HG00733, which contained 115 good quality Strand-seq cells (based on visual inspection of Strand-seq libraries). For genotyping, 249 autosomal inversions from our inversion callset (Section 3.3.2) that were classified as simple (balanced) inversions in HG00733, were selected.

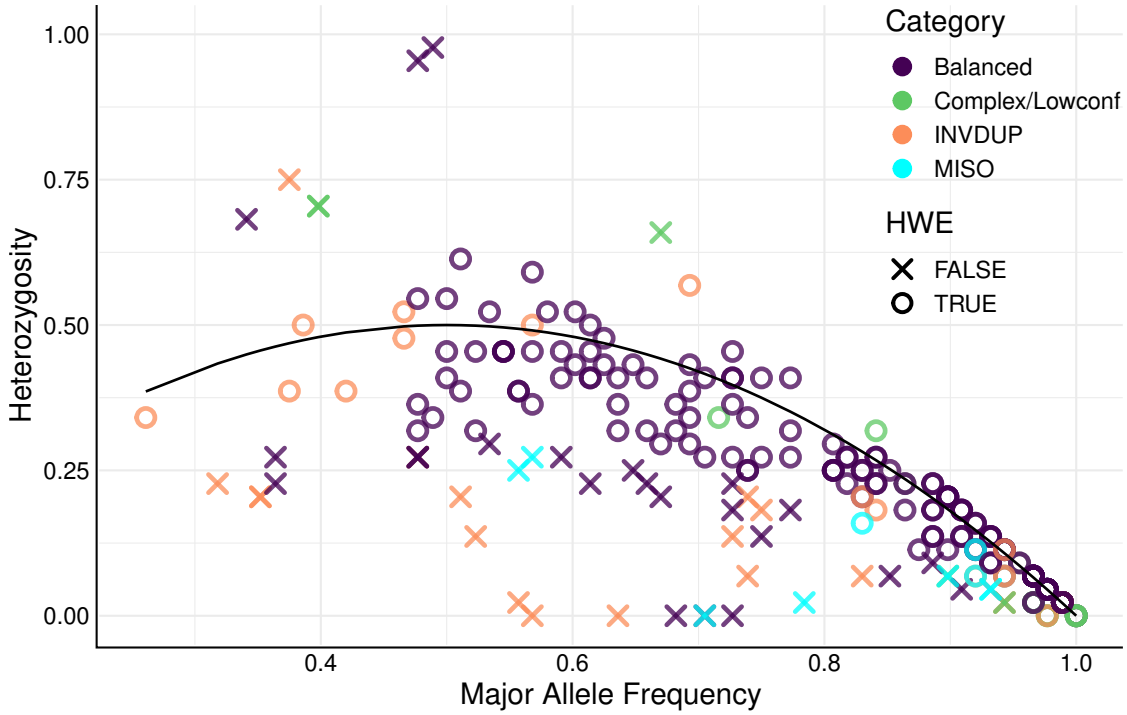
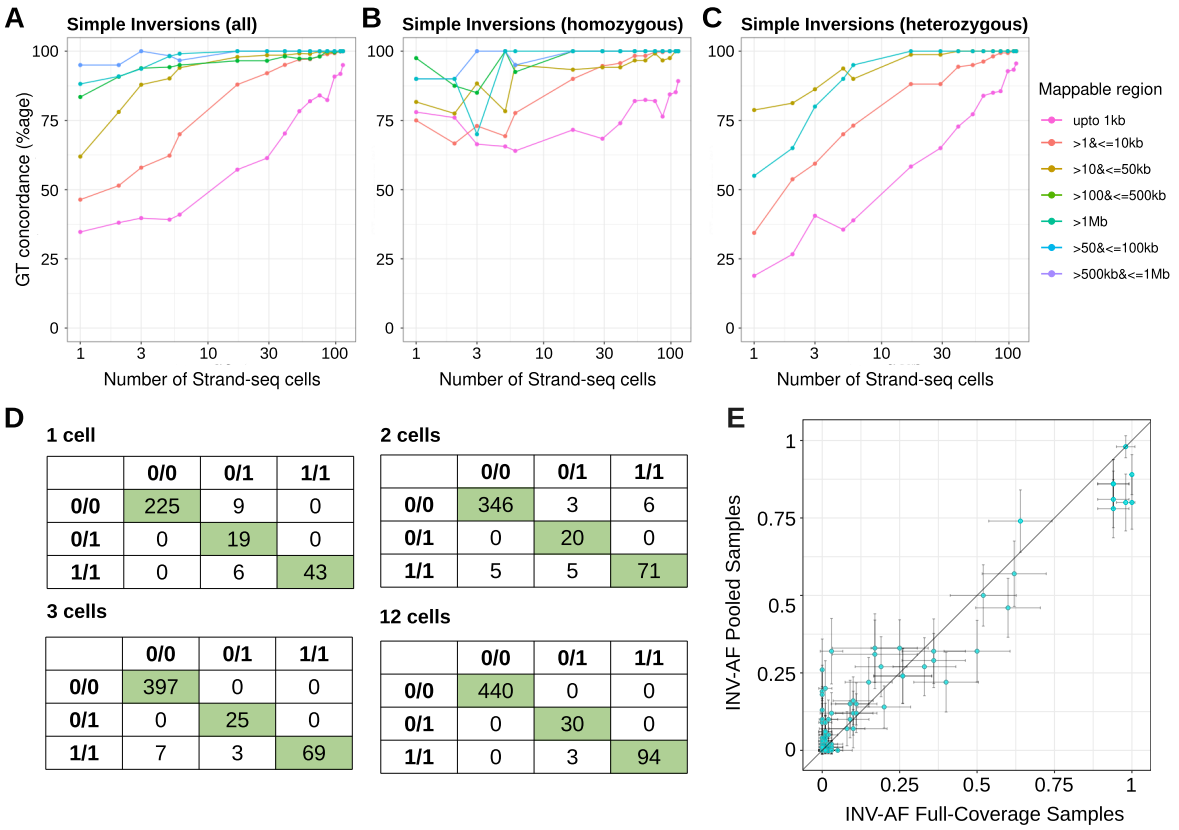


Figure 3.5: Hardy-Weinberg equilibrium (HWE) evaluation. The horizontal axis represents the major allele frequency, while the vertical axis represents the heterozygosity, that is, the fraction of heterozygous genotypes (where a genotype is heterozygous if it carries the major allele and one of the alleles included in the minor allele category as described in Section 3.3.3). The theoretical *2pq* curve is shown in black. The plot represents 275 autosomal inversions belonging to the following categories: (i) Balanced: loci with at least one sample showing a "simple" inversion, (ii) Complex/Lowconf: Complex loci or loci with low confidence genotypes, (iii) INVDUP: Inverted duplications and (iv) MISO: putative assembly misorientations.

We genotyped these loci in multiple sets of randomly downsampled single cells, with downsampling percentages ranging from 1% to 99% of the full set. For each downsampling level, we generated ten independent random sets. The average genotype concordance at each downsampling percentage was then computed using the genotypes based on the full set of 115 cells as the ground truth. Figure 3.6A–C illustrate the stratification of genotyping performance by the length of uniquely mappable region ($\geq$75% mappability), shown collectively for all inversions and separately for homozygous and heterozygous inversions. The results clearly indicated that for inversions with more than 10 kbp of uniquely-mappable region, ArbiGent attained a high genotype concordance even with fewer cells. As expected, this trend was more pronounced for homozygous inversions than for heterozygous ones. To gain further insights, we examined confusion matrices individually for sets containing 1, 2, 3 and 12 cells as shown in Figure 3.6D. Based on the trends observed in Figure 3.6A–C, we only focused on inversions with $\geq$50 kbp of uniquely-mappable sequence. These confusion matrices provide a quantified support to the observation that for large inversions belonging to uniquely-mappable regions, ArbiGent's genotype predictions from sparse data remain



1 cell

	0/0	0/1	1/1
0/0	225	9	0
0/1	0	19	0
1/1	0	6	43

2 cells

	0/0	0/1	1/1
0/0	346	3	6
0/1	0	20	0
1/1	5	5	71

3 cells

	0/0	0/1	1/1
0/0	397	0	0
0/1	0	25	0
1/1	7	3	69

12 cells

	0/0	0/1	1/1
0/0	440	0	0
0/1	0	30	0
1/1	0	3	94

Figure 3.6: ArbiGent's genotyping performance across varying number of Strand-seq cells. A, B, C. Genotype concordance (vertical axis) observed across varying number of single-cells (horizontal axis) for sample HG00733, considering all, homozygous and heterozygous inversions, respectively. **D.** Confusion matrices with rows representing ground truth genotypes (based on full coverage data) and columns representing predicted genotypes obtained from randomly subsampled single-cell libraries (in sets of n = 1, 2, 3 and 12 cells, respectively). All inversions with >50 kbp of uniquely-mappable sequence (>75% mappability) are assessed. **E.** Comparison of inversion allele frequencies (AFs) observed in the full cohort of 44 samples (horizontal axis) with AFs observed by genotyping the same inversion sites in the single-cell pools containing 66 samples (vertical axis). Depicted 95% confidence intervals are the ±2 standard errors (SE), computed as $SE = \sqrt{\frac{p(1-p)}{n}}$, with p being the allele frequency and n being the total number of alleles observed.

highly consistent with those derived from high-coverage Strand-seq libraries. These findings also suggest that shorter inversions or those located in difficult-to-map regions may benefit from aggregating genotyping information across multiple single cells to enhance accuracy.

Strand-seq pooling experiment

As with other SV classes, sequencing an increasing number of samples is essential for conducting a comprehensive, population-wide analysis of inversion polymorphisms. However, sequencing costs pose a significant challenge in achieving this goal. To mitigate this issue with respect to Strand-seq sequencing, the HGSVC plans to gradually transition to a new

Strand-seq approach known as “cell pooling”. In this strategy, rather than sequencing cell lines from individual samples on separate plates, cells from multiple human samples of diverse origin are pooled into a single sequencing plate. This approach reduces the number of sequenced cells per sample but enables the inclusion of a greater number of samples at the same sequencing cost.

Encouraged by ArbiGent’s robust performance in the downsampling experiment described above, we conducted the first proof-of-concept experiment to evaluate the practical usefulness of pooled Strand-seq data for inversion genotyping. We analyzed 120 samples from the 1KGP, distributed across three independent pools, each consisting of cells from 40 samples of diverse ancestry, with an emphasis on samples of African descent. Each pool was sequenced following the standard Strand-seq protocol (as described in Section 1.5.2 and shown in Figure 1.8A). After quality control, 132 cells (across all pools) were selected for further analysis. To determine the sample of origin for each sequenced cell, we compared the SNP genotypes of each cell against the full set of high-coverage 1KGP samples (n = 3,202) [141]. This approach allowed us to confidently assign 60 samples. We then genotyped 74 inversions containing $>$50 kbp of uniquely-mappable sequence from our inversion callset in these 60 samples using ArbiGent. The estimated inversion allele frequencies derived from the pooled data closely matched those obtained from the original diversity panel as shown in Figure 3.6E. This experiment provided additional validation for ArbiGent’s performance while also demonstrating the potential effectiveness of the Strand-seq pooling strategy.

3.3.4 Application in other studies

In addition to the *Science* [12] and *Cell* [13] studies mentioned above, ArbiGent has been utilized in multiple HGSVC projects, serving as the primary tool for inversion genotyping, validation and population-wide characterization. These include a *Genome Biology* [14] study published in 2023, focused on understanding the impact of having a complete human reference on characterizing inversion polymorphisms. The inversion callset generated in this study, providing a landscape of inversion polymorphisms in the T2T-CHM13 reference genome, across 41 human samples, was genotyped and refined using ArbiGent. ArbiGent was also employed in the recent HGSVC study [15], using a cohort of 65 human samples and improved data sources to generate new inversion callsets for both GRCh38 and T2T-CHM13 reference genomes, while also re-genotyping previous callsets [13, 14] in the latest cohort. This study has been provisionally accepted for publication in *Nature*. These contributions emphasize the reliability of ArbiGent as a genotyping tool and demonstrate its utility in building the catalog of inversion polymorphisms.

3.4 Discussion

This chapter primarily focuses on ArbiGent, a Strand-seq based genotyper designed for detecting inversions and copy number variations. One of the key advancements of ArbiGent

over existing Strand-seq based SV identification approaches is its ability to accurately classify variation in difficult-to-map regions of the genome. These regions, enriched in SDs, are hotspots for inversions and copy number variations but pose significant challenges for genotyping due to unreliable alignments. To address this issue, we developed an approach that quantifies read mappability for each genomic region and incorporates this information to normalize observed sequencing read counts, correcting for alignment artifacts.

ArbiGent also improves upon previous approaches by utilizing all available Strand-seq libraries. In contrast to earlier SV detection methods [11], which relied only on libraries in "WW" and "CC" strand-states due to limitations associated with the use of composite files, ArbiGent additionally incorporates "WC" and "CW" libraries. This helps correct erroneous heterozygous inversion calls reported in prior Strand-seq-based callsets (Figure 3.4). Furthermore, ArbiGent provides genotype likelihood estimations for arbitrary genomic segments of choice, making it a valuable tool for validating and integrating SV loci discovered across different platforms.

Validation experiments including Mendelian consistency and Hardy-Weinberg equilibrium assessments (Section 3.3.3), confirmed the high quality of ArbiGent genotypes. To evaluate ArbiGent's robustness across varying read coverages, we designed a downsampling experiment that demonstrated its ability to perform well even at low coverages, particularly in uniquely mappable regions (Section 3.3.3, Figure 3.6A–D). Encouraged by these findings, we further assessed its genotyping performance using pooled Strand-seq data and observed that inversion allele frequencies estimated from pooled samples closely matched those obtained from full-coverage data (Section 3.3.3). This opens up future prospects for ArbiGent to be used for genotyping and validating inversions, particularly in the context of upcoming HGSVC projects that plan to employ Strand-seq pooling as a cost-effective strategy for sequencing larger cohorts of human samples.

Despite its advancements, ArbiGent has certain limitations. Being a genotyper, it does not correct for any breakpoint inaccuracies, meaning that its genotyping accuracy hinges on the breakpoint precision of the input loci. Additionally, due to the inherent sparsity of Strand-seq data, ArbiGent is less reliable for genotyping very short loci ($\leq$5 kbp). Another Strand-seq based inversion genotyper, InvertypeR, was developed in parallel by Hanlon et al. [138]. While a direct performance comparison between ArbiGent and InvertypeR was not possible at the time of development, such an analysis would be interesting for future iterations.

This chapter also details the efforts of the HGSVC's inversions-working-group to construct a highly curated inversion callset using 44 human samples from the 1KGP [14]. A multi-platform SV discovery approach was employed to leverage the strengths of different data types, including Strand-seq, fully-phased genome assemblies and Bionano optical maps. These individual callsets were integrated into a consolidated callset, which was subsequently validated and genotyped using ArbiGent. Following extensive quality checks, curation and refinement, we generated a high-confidence callset comprising 399 inversions, further clas-

sified based on their distinct variation patterns. Our analysis revealed that an average of 11.6 Mbp (0.39%) of a haploid genome is inverted—twice the genomic length affected by indels [12] and four times that affected by SNPs [5]. Additionally, we observed SD-rich regions on chromosomes 1, 2, 7, 10, 15, 16, and 17 to be particularly enriched for long inversions, supporting the role of NAHR in mediating inversion formation. These findings motivated several downstream analyses including development of targeted approaches to investigate the under-explored phenomenon of inversion recurrence, which will be discussed in detail in Chapter 4.

Chapter 4

Detection and analysis of recurrent inversion polymorphisms in human genomes

This chapter focuses on the inversion recurrence analysis work presented in the study published by the HGSVC's inversions-working-group [13], in which I am a co-first author. This study has already been introduced in Chapter 3, which primarily focused on the efforts to develop a comprehensive inversion callset. This chapter expands on that by detailing downstream analyses conducted using that callset, with a particular emphasis on inversion recurrence. Text, figures and tables from this study have been reused in this chapter, with contributions from other co-authors explicitly mentioned. The primary focus of this chapter is on the methodological framework and findings of the "toggling-indicating SNPs-based approach", which I developed to identify recurrent inversions in the human genome. An orthogonal haplotype-based approach was also developed in this study by PingHsun Hsieh and Matthias Steinrücken. Relevant contributions from other co-authors are also included in this chapter, with individual roles explicitly acknowledged. For publication details and author contributions, please refer to Section C.3.

4.1 Introduction

Inversions have been documented to recurrently toggle in orientation over the course of primate evolution, with non-allelic homologous recombination between flanking inverted repeats recognized as the primary driving mechanism behind this phenomenon [31, 32, 84]. As briefly mentioned in Chapter 3 (Section 3.4), our inversion callset based on 44 human samples revealed that inversions affect twice as much sequence as indels, and four times as much as SNPs [13]. A comparative assessment of the callset growth rate across SV classes demonstrated that the number of insertions and deletions progressively increase with the addition of new genomes with a sharp inflection point observed upon the inclusion of individuals of African descent exhibiting greater genetic diversity [12] (Figure 4.1 orange

lines). However, the growth rate of novel inversions plateaus rapidly, exhibiting about 2-fold reduction in the discovery rate (Figure 4.1). Concurrently, we observed an excess of common alleles (with $>5\%$ minor allele frequency) among inversions relative to other SV classes, quantified as 67% for inversions vs. 48% for indels, ($p = 2.6 \times 10^{-11}$, two-tailed Fisher's exact test). We hypothesized that this excess arises from independent NAHR-mediated recurrent mutations, aligning with the observation that SD-rich regions of the human genome serve as hotspots for inversions, particularly large events [13] (Figure 3.3, Section 3.3.2).

Mechanistically, an NAHR-mediated event occurring independently at the same locus might not always involve the same breakpoints. Consequently, in theory, recurrent inversions should be distinguishable based on shifts in breakpoint positions within haplotypes. Furthermore, specific variation patterns or "scars" [13] adjacent to the inverted loci might help identify if the inversion occurred in different haplotype backgrounds. However, despite advancements in sequencing technologies, sequence resolution within the SD-rich architecture flanking inversions remains imperfect. Owing to the complexities inherent in the characterization of inversions (as outlined in Section 1.5.2), inversion recurrence remains an under-explored topic, marked by a scarcity of readily usable computational methods.

We developed two distinct yet complementary approaches to detect inversion recurrence [13]. Firstly, we designed the toggling-indicating SNPs (tiSNPs)-based method to evaluate the recurrence status of an inversion locus. This approach systematically screens the locus for SNPs inconsistent with a single inversion origin, leveraging aggregated locus-wide evidence to infer the recurrence status. For this study [13], we applied the tiSNPs-based approach using haplotype-resolved Strand-seq sequencing reads. Secondly, we developed an orthogonal haplotype-based approach working with phased haplotypes (integrated using Strand-seq and PacBio sequencing data). This method employs phylogenetic analyses, such as haplotype-based principal component analysis (PCA) and ancestral recombination graph reconstruction, to infer both the evidence and rate of inversion recurrence. These two approaches are inherently complementary. The tiSNPs-based method evaluates individual within-inversion SNPs independently, making it largely robust to recombination effects and noise in the sequencing data. In contrast, the haplotype-based approach leverages linkage and genetic variation patterns to estimate both the number of recurrent events and the inversion rate per generation.

4.2 Methodological framework

The toggling-indicating SNPs-based approach utilizes haplotype-resolved sequencing data to detect inversion recurrence by analyzing the occurrence and orientation of biallelic SNPs within the inverted locus. The conceptual foundation for this approach (also depicted in Figure 4.2) is as follows: Consider a biallelic SNP within the inverted locus. A single origin of both the SNP and the inversion implies a temporal relationship between the two events. Assuming the SNP occurred first, at most three SNP-inversion haplotype configurations can

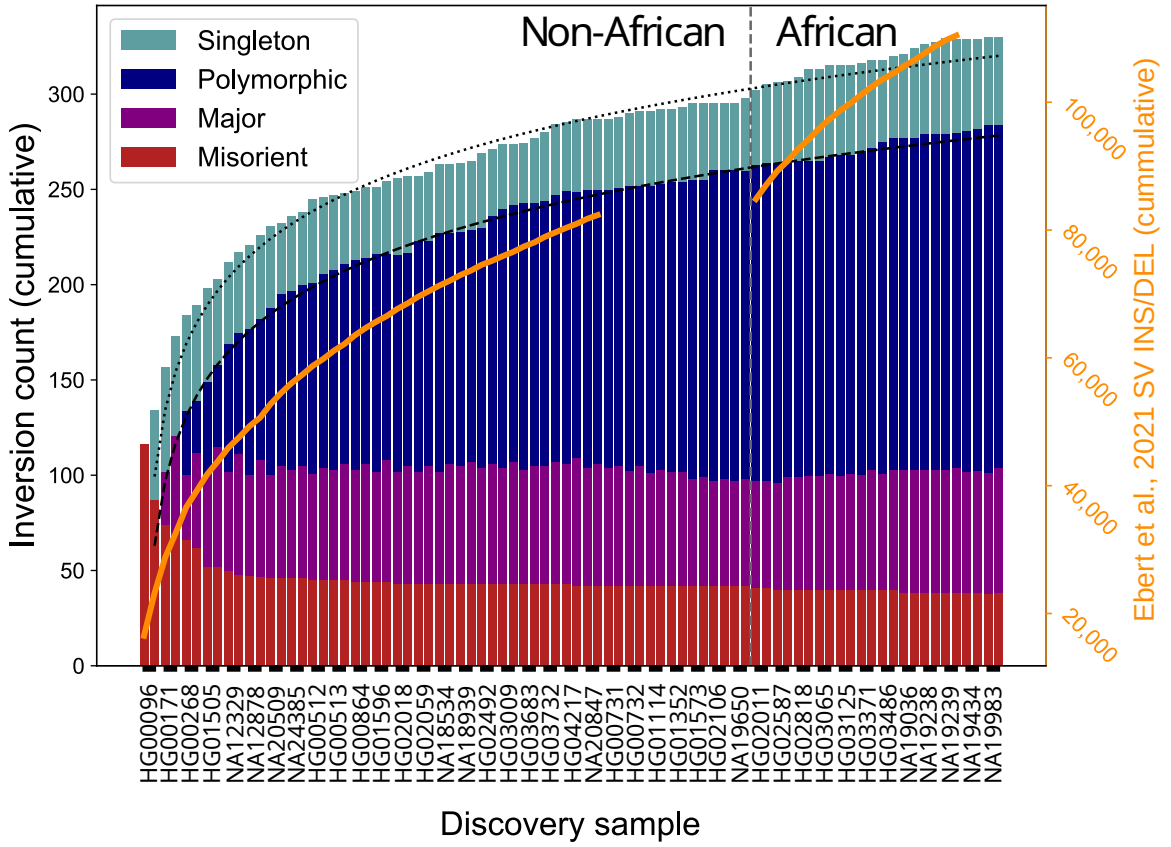


Figure 4.1: Growth rate of inversions vs indels. Rate of balanced inversions discovered with each added genome differs from insertions and deletions (orange lines, right axis). Dotted lines represent logarithmic growth model fits. Singleton: 1 allele; polymorphic: AF $<$ 50%; major: AF $\geq$ 50% and $<$ 100%; putative misorient: AF = 100%. Figure created by Peter Audano [13].

be observed on the population level (as shown in Figure 4.2, left). Theoretically, in this scenario, a haplotype carrying the inversion allele but not the SNP allele cannot exist. More generally, regardless of which event happened first, if both occurred only once, one haplotype configuration must always be missing. However, if a SNP displays all four possible SNP-inversion haplotype configurations, it suggests that either the SNP or the inversion has recurred. A large number of such SNPs, which we refer to as tiSNPs, distributed across the entire inverted locus serves as strong evidence that the inversion recurred.

We used haplotype-resolved Strand-seq data to develop and validate this approach, so its workflow in the context of Strand-seq sequencing reads is described in the following sections. However, the method is broadly applicable to any haplotype-resolved sequencing data. Given the SNP information (position and alleles) and Strand-seq read alignments, the approach proceeds as follows:

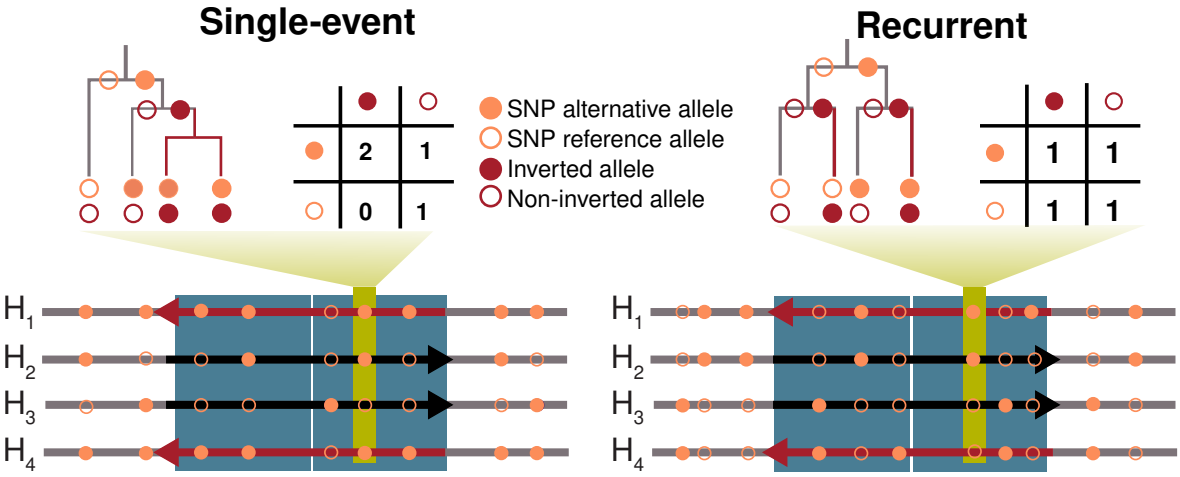


Figure 4.2: Theoretical framework behind the tiSNPs-based approach. Schematic overview of the conceptual framework underlying the tiSNPs-based approach. If an inversion occurred only once in evolutionary history (left), there are two possible temporal sequences in which a SNP (located within the inversion) and the inversion event itself could have arisen. In either scenario, this results in at most three possible SNP–inversion haplotype configurations. In contrast, if all four haplotype configurations are observed—meaning each SNP allele appears in both the inverted and non-inverted state—this indicates that the inversion must have occurred independently more than once (right). H_1 to H_4 denote phased haplotypes used to determine allele counts across the configurations shown in the four-cell matrices above. Red backward arrows within the haplotypes represent the inverted allele, while black forward arrows indicate the non-inverted allele.

Allele-specific read counting

As the inverted locus is traversed from left to right, the number of Strand-seq reads in Watson (W) and Crick (C) orientations is recorded for each biallelic SNP. To ensure data quality, reads are filtered by removing secondary alignments, duplicates and those with low mapping quality ($\leq$10). These read counts are tracked individually for each Strand-seq cell per sample. In this step, only common biallelic SNPs with $\geq$5% allele frequency (AF) are considered, as rarer SNPs do not provide sufficient statistical power to detect evidence of recurrence (Figure 4.41).

Assigning orientation to reads

Using the normal cell state, that is, the strand structure of the cell in non-variant regions, the "W" and "C" strand notations are translated into "non-inverted" (forward) and "inverted" (reverse) designations. For instance, for a locus with a normal cell state of "CC", only "C" reads are expected in the absence of variation; any aligned "W" reads would suggest an inversion, as shown in Figure 4.3. Accordingly, all "C" reads aligned to this locus are labeled as non-inverted while all "W" reads are labeled as inverted. Since for cells with a "WC/CW" strand configuration this characterization is non-trivial, we exclusively consider "WW" and "CC" cells from each sample (Figure 4.4 2).

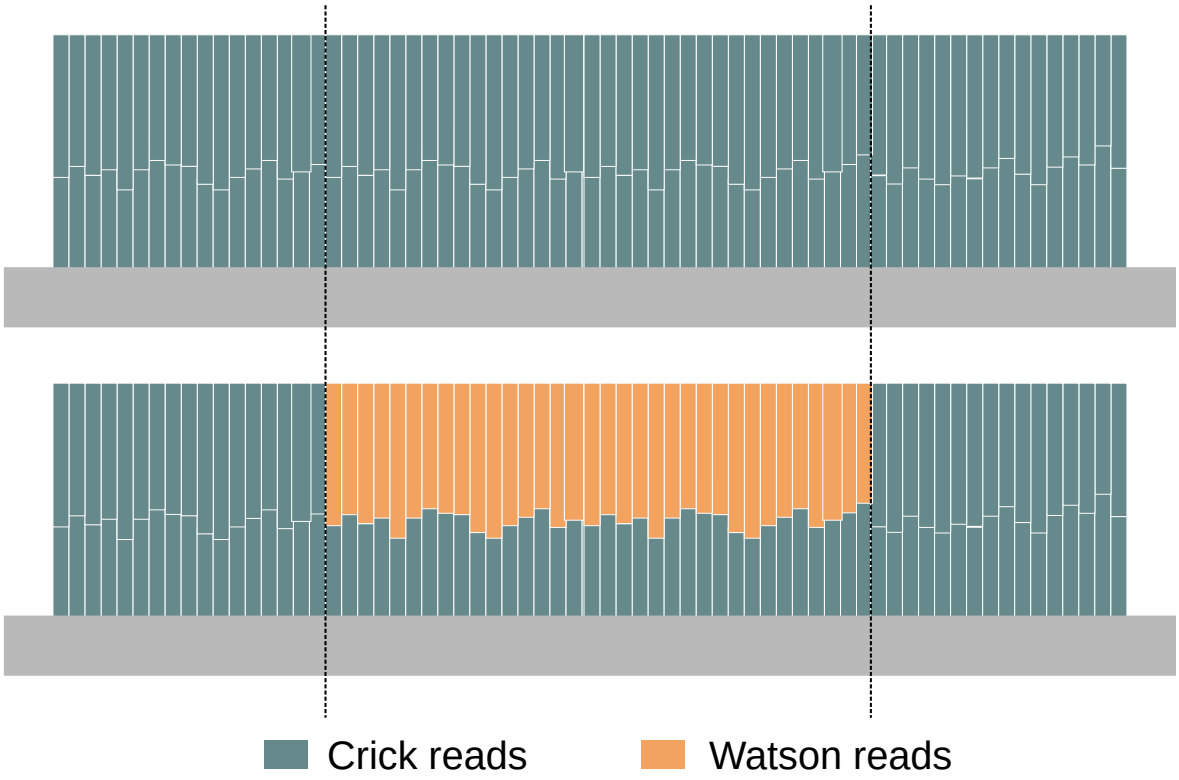


Figure 4.3: Assigning orientation to Strand-seq reads inside the inverted locus. The region between the dotted lines represents an inversion locus. Two CC cells (strand-state determined based on the reads aligned to the region outside the inversion), are shown. The cell shown on the bottom carries the inversion indicated by the Watson reads aligning to this locus. Therefore, all the Watson reads aligned at this locus would be labeled as "reverse" reads.

From single cell to population level

At this stage, we have a record detailing, for each biallelic SNP within the inverted region, the occurrence count of each allele in both the forward and reverse states. However, as mentioned above, these counts are initially recorded at the single-cell level. To derive a sample-level consensus, the occurrence counts are aggregated across all cells within a sample. Since inversion recurrence is a population-wide phenomenon, these aggregated counts are further concatenated across all samples. By the end of this step, we obtain a region-wide overview of the possible SNP-inversion haplotype configurations (Figure 4.4 3).

Detection of tiSNPs

The final step is to identify tiSNPs, that is, the biallelic SNPs where each allele is observed in both inverted and non-inverted orientations across samples. Although detecting all four configurations at least once is theoretically enough to classify a SNP as a tiSNP, gene conversion or sequencing errors ("background" reads in the wrong orientation in Strand-seq data [92]) could generate false signals of recurrence. To account for this, we set a threshold

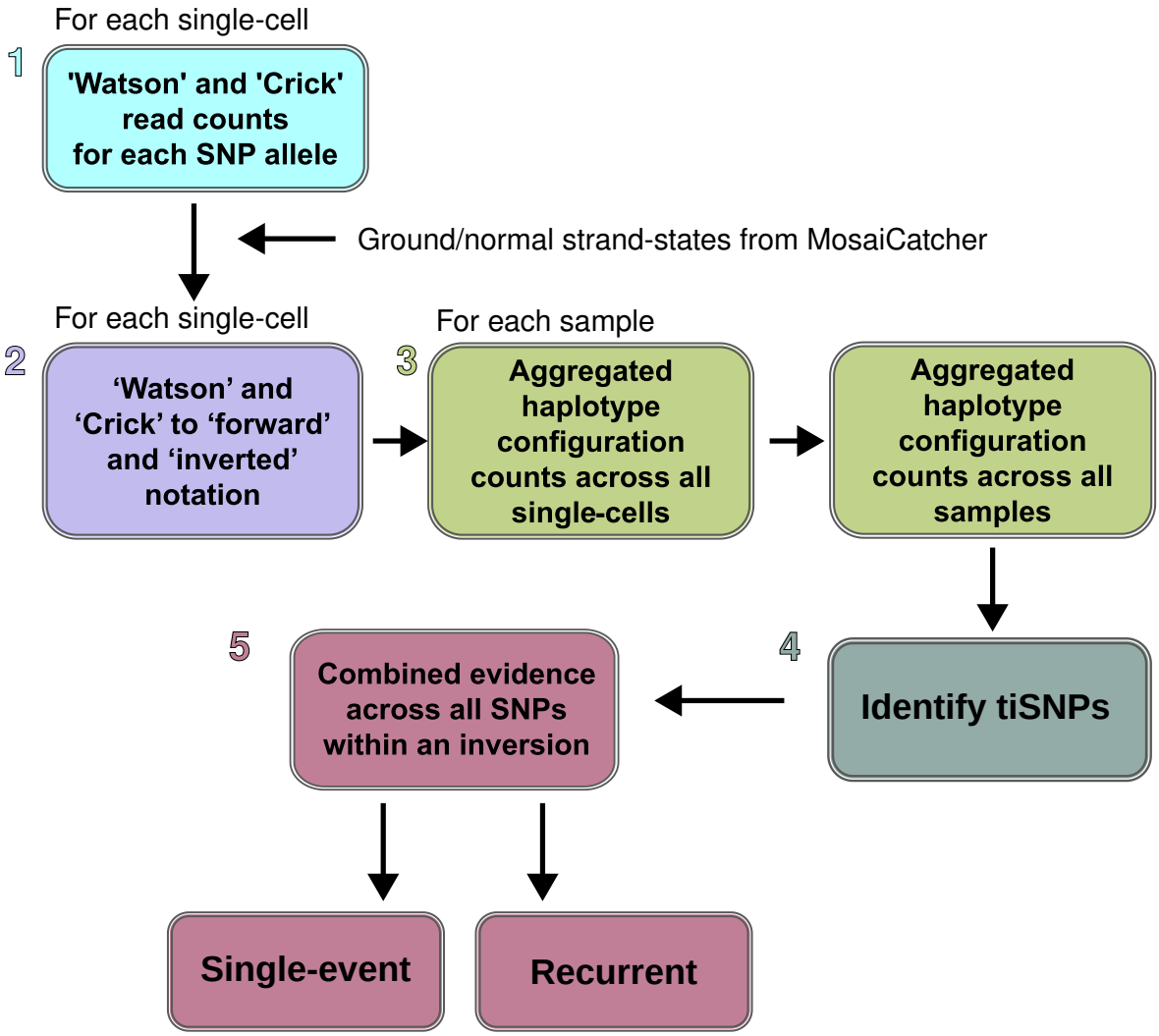


Figure 4.4: Workflow of the tiSNPs-based approach for Strand-seq data.

requiring each SNP-inversion allele configuration to be observed at least three times. In other words, for a SNP to qualify as a tiSNP, each allele must be supported by at least three reads in both the inverted and non-inverted orientations (Figure 4.4 4).

Final verdict on recurrence

For a quantitative assessment, both the number and the positions of tiSNPs across the locus are recorded. The presence of at least one tiSNP within the inverted locus is interpreted as "evidence" of inversion recurrence. The strength of this evidence is further evaluated based on the frequency and distribution of tiSNPs across the inverted region. Specifically, a high proportion of tiSNPs—relative to the total number of analyzed SNPs—distributed throughout the entire locus, serves as a strong indicator of inversion recurrence. Given the substantial variability in the architecture of inverted loci, the definition of what constitutes a "high fraction of tiSNPs" is left as a user-defined parameter. The additional requirement

for tiSNPs to be spread across the entire locus helps filter out cases where the recurrence signal may actually originate from another variant within the inverted region. Additionally, in combination with independent assessment of each SNP, it helps to distinguish true recurrence from recombination or gene conversion events, which could otherwise produce similar signals (Figure 4.4 5).

As previously mentioned, this approach was developed and tested using Strand-seq data, but it can be readily applied to other types of sequencing data, such as long reads or fully phased assemblies. In these alternative settings, however, the method requires both phased SNP genotypes and phased inversion genotypes as input.

4.3 Performance Evaluation and Results

We used the SNP callset from New York Genome Center (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/`) for the tiSNPs-based analysis presented in this study [13]. After filtering out rare SNPs (AF $<5\%$, as described in Section 4.2), we analyzed 252 out of 279 "balanced" inversions located on autosomes and the X chromosome from the callset described in Chapter 3, Section 3.3.2. At least one tiSNP was detected in 49 of the analyzed loci. To further investigate these candidate recurrent inversions, we conducted a series of validation experiments, described in the subsequent sections.

4.3.1 Influence of flanking inverted repeats on inversion recurrence

A key factor distinguishing recurrent from non-recurrent inversions is their flanking sequence architecture. Similar to other classes of recurrent SVs (Section 1.4.2), NAHR between flanking inverted repeat sequences is hypothesized to be the primary driver of inversion recurrence [32]. Accordingly, it is hypothesized that length of the flanking repeats positively correlates with the frequency of NAHR [63, 66] (Section 1.4.3).

To evaluate whether our tiSNPs-based classification aligns with this mechanistic model, we tested whether inversions identified as recurrent (that is, containing $\geq$1 tiSNP) tend to have longer flanking inverted repeat sequences than single-event inversions. To focus on inversion-mediating repeats, we analyzed a 70 kbp flanking window, which included 10 kbp inside each annotated inversion breakpoint. We extracted nucleotide sequences from both flanks using the GRCh38 reference genome and estimated the length of inverted repeats using dot plot alignments. A comparison of the fraction of tiSNPs between recurrent and single-event inversions, in relation to flanking inverted repeat length (Figure 4.5), revealed a clear pattern:

- Inversions showing evidence of recurrence were enriched for longer flanking inverted repeats, with the fraction of tiSNPs increasing as repeat length increased.

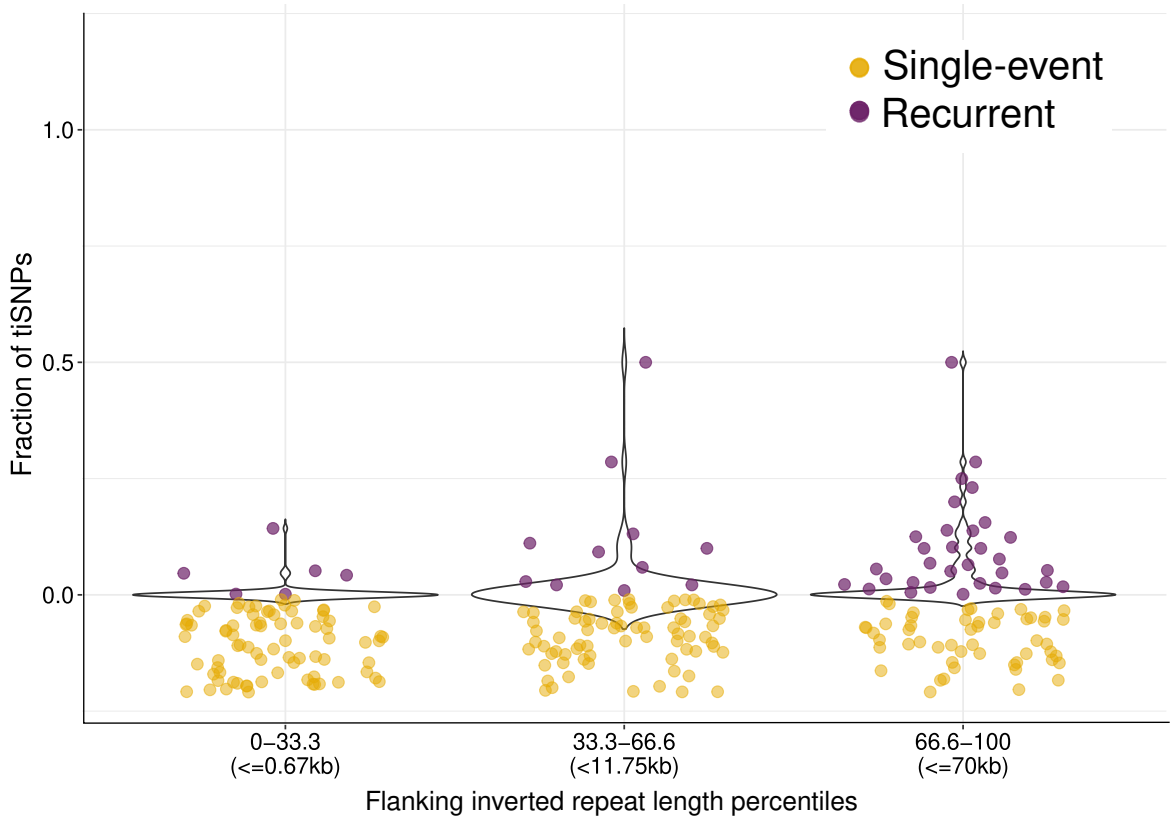


Figure 4.5: Flanking inverted repeat length versus fraction of tiSNPs. Plot showing relationship between the fraction of tiSNPs (vertical axis) observed at each locus and length of the inverted repeat at its flanks (horizontal axis). The points where the fraction of tiSNPs = 0 are jittered in the downwards direction for better visualization. This plot is based on 252 balanced inversions tested for recurrence using the tiSNPs-based approach. For inversions where the tiSNPs-based approach detects evidence of recurrence, an enrichment for longer flanking inverted repeats is observed. Moreover, with increasing length of the flanking inverted repeats, the fraction of tiSNPs clearly appears to increase.

- Inversions without any recurrence signal were more frequently associated with shorter flanking inverted repeats.

These findings indicate that inversions detected as recurrent by the tiSNPs-based method are structurally conforming to NAHR-mediated recurrence requirements, reinforcing the method's validity.

4.3.2 Influence of inversion length on inversion recurrence

A crucial aspect of the quality assessment was to rule out the possibility of the recurrence signal being driven by the inversion length. Since longer inversions tend to contain more SNPs, they inherently have a higher likelihood of showing a recurrence signal compared to shorter inversions with fewer SNPs. Additionally, flanking inverted repeat length is believed to be directly correlated with inversion length [13]. To statistically rule out inversion length

as a potential confounder, we used a multiple linear regression model. This model expressed the fraction of tiSNPs as a function of:

- inversion length
- length of the longest flanking inverted repeat
- major allele frequency (MAF)

We included MAF in the model because inversions with rare alleles have reduced statistical power to detect recurrence compared to inversions with more balanced allele frequencies. The regression results clearly indicated that the recurrence signal, that is, the fraction of tiSNPs is primarily being influenced by the length of flanking inverted repeats and major allele frequency ($\beta = 6.694 \times 10^{-4}$, $p = 7.5 \times 10^{-4}$ and β = -1.032$\times 10^{-1}$, $p = 2.65 \times 10^{-6}$, respectively). Importantly, inversion length was observed to have no significant influence ($\beta = 4.817 \times 10^{-7}$, $p = 8.35 \times 10^{-1}$).

To incorporate an additional layer of validation and assess the robustness of the method to varying allele frequency thresholds used for SNP filtering, we performed the analysis using SNP sets obtained using different allele frequency cutoffs (ranging from no filtering to SNPs with AF $\geq$ 3, 5, 10, 15, 20%). The results from the regression model (described above), applied separately to each SNP set, consistently demonstrated that flanking inverted repeat length and major allele frequency were the primary factors driving the recurrence signal, further confirming the method's reliability and robustness.

4.3.3 Phylogenetic validation

In theory, if an inversion has occurred independently in different haplotype backgrounds, sequence-based clustering would group inverted and non-inverted haplotypes together. In contrast, if the inversion arose only once, we would expect a clear separation between inverted and non-inverted haplotypes in the clustering. To assess whether the loci identified as recurrent through the tiSNPs-based approach exhibit this pattern, we applied centroid hierarchical clustering to generate locus-specific phylogenetic trees based on SNP haplotypes, distinguishing ancestral and derived SNP alleles as determined using the Chimpanzee genome (PanTro6). Moreover, from a mechanistic perspective, if a locus toggled in orientation multiple times, evidence of recurrence should be evenly distributed across the entire locus. Otherwise, it becomes challenging to differentiate true inversion recurrence from recombination patterns. To validate the observed results from this perspective, we additionally generated phylogenetic trees independently for the right, middle and left thirds of each analyzed locus. Although tiSNPs-based approach implicitly produces SNP genotype information, the sparse nature of Strand-seq data renders these haplotypes unsuitable for direct visualization. To address this, we employed the "integrative phasing" strategy offered by WhatsHap [46] to incorporate haplotype information available from PacBio reads (HiFi: 14 samples, CLR: 30 samples) into the sparse Strand-seq derived haplotypes. The

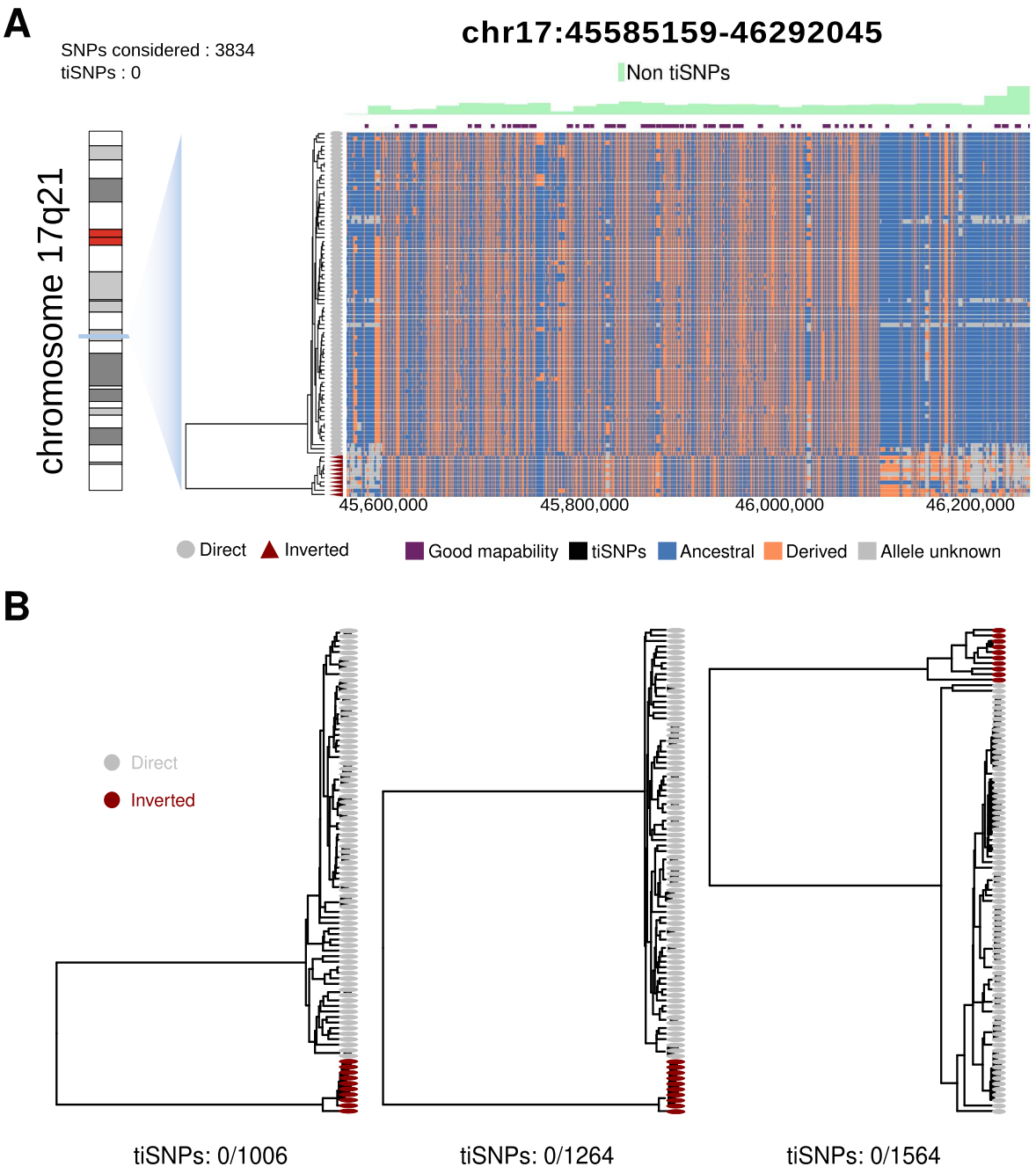


Figure 4.6: Single-event inversion at 17q21. A. Left: Chromosome ideogram showing position of the inverted locus. Middle: Centroid hierarchical-clustering based dendrogram showing relationships among inverted and non-inverted SNP haplotypes shown on the right. Right: Ancestral (blue) versus derived (orange) SNPs, determined using the Chimpanzee genome (PanTro6). Top: Regions with $\geq$75% mappability are marked in purple. Histograms show the distribution of SNPs (no tiSNPs in this case) across the locus. **B.** Phylogenetic trees constructed independently for the right, middle and left thirds of the locus, each mirroring the single-event pattern depicted by the tree constructed for the entire locus.

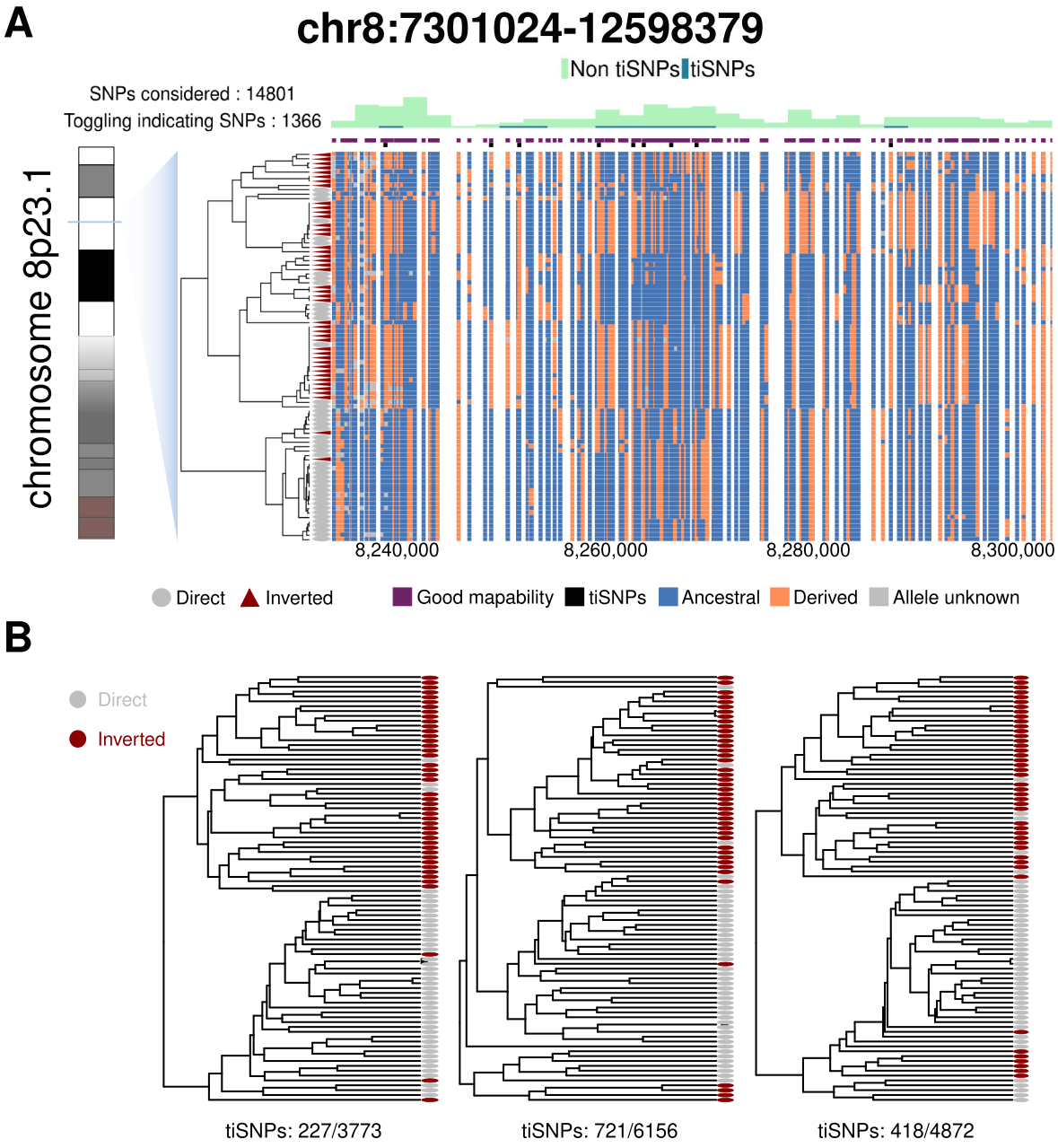


Figure 4.7: Recurrent inversion at 8p23.1. **A.** Left: Chromosome ideogram showing position of the inverted locus. Middle: Centroid hierarchical-clustering based dendrogram showing relationships among inverted and non-inverted SNP haplotypes shown on the right. Right: Ancestral (blue) versus derived (orange) SNPs, determined using the Chimpanzee genome (PanTro6). Top: tiSNPs track shown in black. Regions with $\geq$75% mappability are marked in purple. Histograms show the distribution of toggling vs non toggling-indicating SNPs. Because of the massive size of the 8p23.1 inversion, only a 100 kbp distal region (chr8:8225000-8301024) of the inverted locus is shown in this part of the figure. **B.** Phylogenetic trees constructed independently for the right, middle and left thirds of the entire 8p23.1 inversion locus. Inverted and non-inverted haplotypes can be seen clustering together in each tree, with tiSNPs distributed across the whole locus.

resulting integrated haplotypes were then used for tree construction and the visualizations depicted in Figures 4.6, 4.7 and 4.9.

As a negative control, we analyzed the 706 kbp inversion located on 17q21.31, exhibiting an allele frequency of 11% in our inversion callset. This inversion has been hypothesized, by multiple studies, to have occurred only once in the last 2.3 million years [32, 34, 76]. In line with this expectation, none of the 3,834 analyzed SNPs within this locus were identified as tiSNPs. The hierarchical clustering-based phylogenetic tree for this locus showed that all inverted haplotypes formed a distinct cluster, clearly separated from the non-inverted haplotypes (Figure 4.6A). This pattern remained consistent across trees generated independently for different segments of the locus (Figure 4.6B).

As a positive control, we selected the 5.3 Mbp inversion located at chromosome 8p23.1 which exhibited an allele frequency of 50% in our callset and has previously been reported to undergo recurrence [77, 142]. In contrast to the 17q21.31 inversion, 1,366 out of 14,801 SNPs (9.2%) were identified as tiSNPs, spanning the entire locus. The phylogenetic trees constructed for this region further supported the multiple occurrence of this inversion event in different haplotype backgrounds, as inverted and non-inverted haplotypes appeared together within similar clusters (Figure 4.7).

Cumulatively, the findings from all analyzed cases demonstrated that loci with a high proportion of tiSNPs evenly distributed across the locus exhibited clustering patterns in which inverted and non-inverted haplotypes did not form distinct, separate groups, whether considering the entire locus or its individual subsections, conforming to the expected phylogenetic pattern for a recurrent event.

The following sections mention some of the work contributed by co-authors of the study [13]. The haplotype-based approach was developed and employed by PingHsun Hsieh and Matthias Steinrücken. Pille Hallast analyzed inversion recurrence on chromosome Y. David Porubsky computed the overlap between inversions and morbid CNVs. Wolfram Höps analyzed the SD-architectural changes around inversions using phased assemblies.

4.3.4 Orthogonal support and downstream analyses

As mentioned previously, we additionally developed a complementary haplotype-based approach for inversion recurrence detection. This approach utilizes phylogenetic analyses, including haplotype-based PCA and ancestral recombination graph reconstruction, to estimate inversion recurrence rates [13]. Due to mappability-based SNP filtering, this approach could be applied to 127 balanced autosomal and X-chromosomal inversions from our callset. While the tiSNPs-based approach could analyze 252 loci, we restricted our downstream analyses to the 127 loci that could be tested by both methods for highly reliable findings. Of the 34/127 loci identified as recurrent using the tiSNPs-based detection approach, all but 2 (94%) were also classified as recurrent by the haplotype-based method. We observed that 52% (66/127) of inversions showed evidence for inversion recurrence by at least one of the two approaches. Both methods agreed in 93 out of 127 cases, with 32 (34%) classified as "recurrent" (Table 4.1) and 61 as "single-events". The inversion rates estimated by the haplotype-based approach ranged from 3.4×10^{-6} to 1.4×10^{-4} (median $=1.2\times10^{-5}$). Based on the inferred phylogenetic tree (for 100 kbp distal region, due to size limitation), the haplotype-based method identified 15 independent inversion events for the 8p23.1 inversion with a mutation rate of 1.11×10^{-4} inversions per generation—equivalent to one inversion per 10,000 parent to child transmissions. To put this into perspective, this mutation rate is four orders of magnitude higher than that observed for SNPs [39]. In contrast, we predicted that the 17q21.31 locus has undergone a single inversion event with a calculated mutation rate of 3.47×10^{-6}. These observations, on he one hand, indicate extensive inversion toggling in humans and on the other hand, demonstrate the substantial variation between inversion rates observed across different loci.

In line with the hypothesized NAHR mechanism, we observed that 72% (23/32) of the identified recurrent inversions have $\geq$10 kbp of flanking inverted repeats with $\geq$79% sequence identity as determined using the dot plot alignments of the flanking sequences (Section 4.3.1, Table 4.1). To statistically assess the influence of repeat length and sequence identity on inversion recurrence using the subset of 127 inversions analyzed by both approaches, we computed the correlation between them and observed that both length of the flanking inverted repeat (Pearson's correlation: 0.51; $p=1.7\times10^{-7}$) and its sequence identity (Pearson's correlation: 0.39; $p=1.3\times10^{-4}$) positively correlate with inversion recurrence status (Figure 4.8A, B). However, flanking inverted repeat length and sequence identity themselves appeared to be strongly correlated (Pearson's correlation: 0.63; $p=1\times10^{-11}$), suggesting a potential confounding effect (Figure 4.8C). To account for this, we used a multivariate lo-

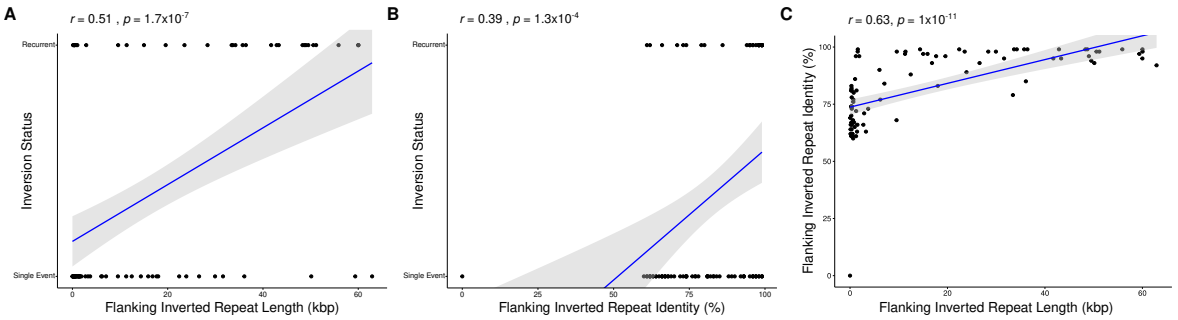


Figure 4.8: Inverted repeat length and sequence identity vs inversion recurrence. Relationship between inversion recurrence status determined by the tiSNPs-based approach and the haplotype-based approach for the consensus set of 93 balanced inversion loci. Blue lines and gray areas are regression lines and their 95% confidence areas, respectively (p-values based on Pearson's correlation).

gistic regression model using inversion length, flanking inverted repeat length and sequence identity as variables, which confirmed that the primary driver for inversion recurrence status is flanking inverted repeat length ($p = 7.2\times10^{-3}$), while neither repeat sequence identity nor inversion length showed any significant influence ($p = 3.1\times10^{-1}$ and $p = 8.6\times10^{-1}$, respectively). Given that length of the inverted repeat is the key parameter facilitating NAHR [63], these observations alongside the results described in Sections 4.3.1 and 4.3.2, support the hypothesis implicating NAHR as the main driver of inversion recurrence.

One particularly noteworthy recurrent candidate was the 169 kbp inversion at chromosome 11p11. The tiSNPs-based approach detected 54 (of 389, that is, 14%) tiSNPs, distributed across the entire inversion, with phylogenetic trees conforming to the pattern expected for a recurrent inversion (Figure 4.9). The haplotype-based approach predicted eight independent inversion events for this locus with the inversion rate estimated as 4×10^{-5}. Furthermore, we observed that six out of 32 recurrent inversions have been reported to have recurred also in great apes [31], suggesting that inversion recurrence is an ancient and widespread phenomenon in both humans and their close relatives.

While the haplotype-based approach could not analyze Y-chromosomal inversions due to its SNP filtering strategy [13], the consensus between the two complementary approaches focused on autosomal and X-chromosomal inversions only. However, we independently analyzed 11 balanced Y-chromosomal inversions for recurrence, using chromosome Y phylogenetic trees and identified eight of them as recurrent, with inversion rates ranging from 1.07×10^{-4} to 2.68×10^{-4} per father-to-son transmission. Additionally, we observed that sex chromosomes are significantly enriched for recurrent inversions compared to autosomes (chi-squared test, $p = 1.2\times10^{-4}$ and $p = 6.6\times10^{-3}$, for X and Y, receptively), with 45% of all recurrent loci identified in our study located on the sex chromosomes. All 40 recurrent inversions, identified in our study, cumulatively cover about 0.6% of the human genome.

Chr band	Position	Size (kbp)	Inverted AF	FIR size (kbp)	FIR identity(%)	Morbid CNVs	tiSNPs	Recurrent events [.95 CI]	Inversion rate ($\times 10^{-4}$) [.95 CI]
chr1p36.21	chr1:13104252-13122521	18.30	0.69	60.00	95	-	4 (2.47%)	13 [7.00 ,13.75]	1.02 [0.272 ,1.21]
chr10q11.22	chr10:46983451-47468232	484.80	0.09	0.42	61	-	41 (4.64%)	7 [5.00 ,7.00]	0.59 [0.326 ,0.799]
chr11p11.12	chr11:50154999-50324102	169.10	0.40	41.72	95	-	54 (13.88%)	8 [6.15 ,9.00]	0.4 [0.328 ,0.571]
chr15q13.2-13.3	chr15:30618103-32153204	1,535.10	0.11	0.34	74	15q11.2, 15q13.3, 15q26	6 (0.17%)	4 [2.00 ,7.00]	0.278 [0.0895 ,0.6]
chr15q25.2	chr15:84373375-84416696	43.30	0.56	34.22	99	15q26	5 (3.45%)	9 [5.30 ,10.00]	0.529 [0.301 ,0.693]
chr16p12.3	chr16:16721273-18073542	1,352.30	0.08	0.37	66	ATR-16	5 (0.13%)	4 [3.00 ,5.00]	0.287 [0.15 ,0.484]
chr16p12.1-11.2	chr16:28471892-28637651	165.80	0.36	23.53	98	ATR-16, 16p11.2-p12.2	4 (1.19%)	6 [3.27 ,6.00]	0.484 [0.264 ,0.661]
chr2p11.1	chr2:91832040-92012663	180.60	0.41	48.31	99	-	10 (6.8%)	19 [10.62 ,19.38]	1.41 [0.931 ,1.85]
chr2q11.1-11.2	chr2:95800191-96024403	224.20	0.08	49.02	96	2q11.2-deletion	3 (1.59%)	4 [2.38 ,5.00]	0.408 [0.234 ,0.681]
chr3q29	chr3:195749463-195980207	230.70	0.26	0.36	73	3p25.3, 3q29	34 (4.22%)	5 [3.00 ,9.00]	0.422 [0.229 ,0.837]
chr7p22.1	chr7:5989046-6735643	746.60	0.10	60.04	98	-	33 (1.75%)	7 [6.00 ,8.00]	0.506 [0.314 ,0.815]
chr7q11.1	chr7:60911891-61578023	666.10	0.52	33.66	99	-	100 (13.77%)	16 [14.10 ,20.00]	0.654 [0.49 ,0.869]
chr7q11.21	chr7:65219157-65531823	312.70	0.33	15.02	97	-	1 (0.13%)	5 [3.00 ,8.00]	0.318 [0.167 ,0.663]
chr7q11.23	chr7:73113989-74799029	1,685.00	0.05	0.75	80	WBS	19 (0.93%)	3 [2.00 ,4.00]	0.262 [0.136 ,0.433]
chr7q11.23	chr7:74869950-75058098	188.10	0.10	43.32	95	-	1 (0.53%)	6 [1.90 ,6.00]	0.57 [0.126 ,0.779]
chr8p23.2	chr8:2343351-2378385	35.00	0.51	55.88	99	8p23.1	32 (12.36%)	17 [3.40 ,17.00]	1.13 [0.33 ,1.53]
chr8p23.1	chr8:7301024-12598379	5,297.40	0.50	1.04	86	8p23.1	1366 (9.23%)	15 [4.75 ,17.00]	1.11 [0.228 ,1.6]
chr1p13.3	chr1:108310642-108383736	73.10	0.57	60.01	99	1p36	3 (1.44%)	5 [5.02 ,5.97]	0.184 [0.184 ,0.194]
chr11q14.3	chr11:89920623-89923848	3.20	0.53	48.70	99	-	3 (25%)	5 [5.05 ,6.95]	0.336 [0.338 ,0.411]
chr16p13.11	chr16:14954790-15100859	146.10	0.77	33.43	79	ATR-16, 16p13.11	5 (2.23%)	3 [3.00 ,8.00]	0.264 [0.191 ,0.832]
chr7q11.21	chr7:62290674-62363143	72.50	0.42	19.58	96	-	12 (5.08%)	10 [5.50 ,10.90]	0.892 [0.598 ,0.896]
chr7q11.21	chr7:62408486-62456444	48.00	0.57	2.90	71	-	12 (5.91%)	18 [9.12 ,19.00]	0.942 [0.458 ,1.24]
chrXq22.2	chrX:103989434-104049428	60.00	0.63	49.52	94	-	2 (2.67%)	5 [2.22 ,5.00]	0.58 [0.308 ,0.651]
chrXq28	chrX:149599490-149655967	56.50	0.08	0.12	62	-	3 (5.17%)	3 [2.00 ,3.00]	0.351 [0.234 ,0.47]
chrXq28	chrX:149681035-149722249	41.20	0.61	28.37	98	-	7 (15.56%)	9 [7.25 ,9.88]	0.85 [0.78 ,1.21]
chrXq28	chrX:153149748-153250226	100.50	0.60	42.88	99	-	46 (20%)	6 [6.00 ,6.00]	0.573 [0.401 ,0.624]
chrXq28	chrX:154347246-154384867	37.60	0.44	11.39	98	Xq28	1 (2.13%)	4 [4.00 ,4.92]	0.613 [0.542 ,0.936]
chrXq28	chrX:154591327-154613096	21.80	0.43	35.74	99	Xq28	1 (5.56%)	3 [3.00 ,5.85]	0.495 [0.475 ,0.785]
chrXq28	chrX:155386727-155453982	67.30	0.15	50.58	98	-	1 (1.25%)	5 [5.00 ,5.00]	0.577 [0.447 ,0.659]
chrXp11.22	chrX:52077120-52176974	99.90	0.36	36.41	99	SHOX, Xp11.22-p11.23	14 (4.71%)	6 [5.40 ,11.00]	0.542 [0.233 ,1.03]
chrXq13.1-13.2	chrX:72997772-73077479	79.70	0.20	9.60	98	SHOX, STS	16 (13.11%)	6 [2.60 ,6.00]	0.548 [0.288 ,0.598]
chrXq28	chrX:152729753-152738707	9.00	0.40	51.16	98	-	5 (12.5%)	5 [5.05 ,6.95]	0.559 [0.352 ,0.553]

[1] FIR, flanking inverted repeat.
[2] CI, central interval.
[3] Number of recurrent events and inversion rates are calculated by the haplotype-based approach devised by PingHsun Hsieh and Matthias Steinrücken.

Table 4.1: Recurrent inversions in the human genome. 32 inversion loci found to be recurrent by both the tiSNPs and the haplotype-based approaches.

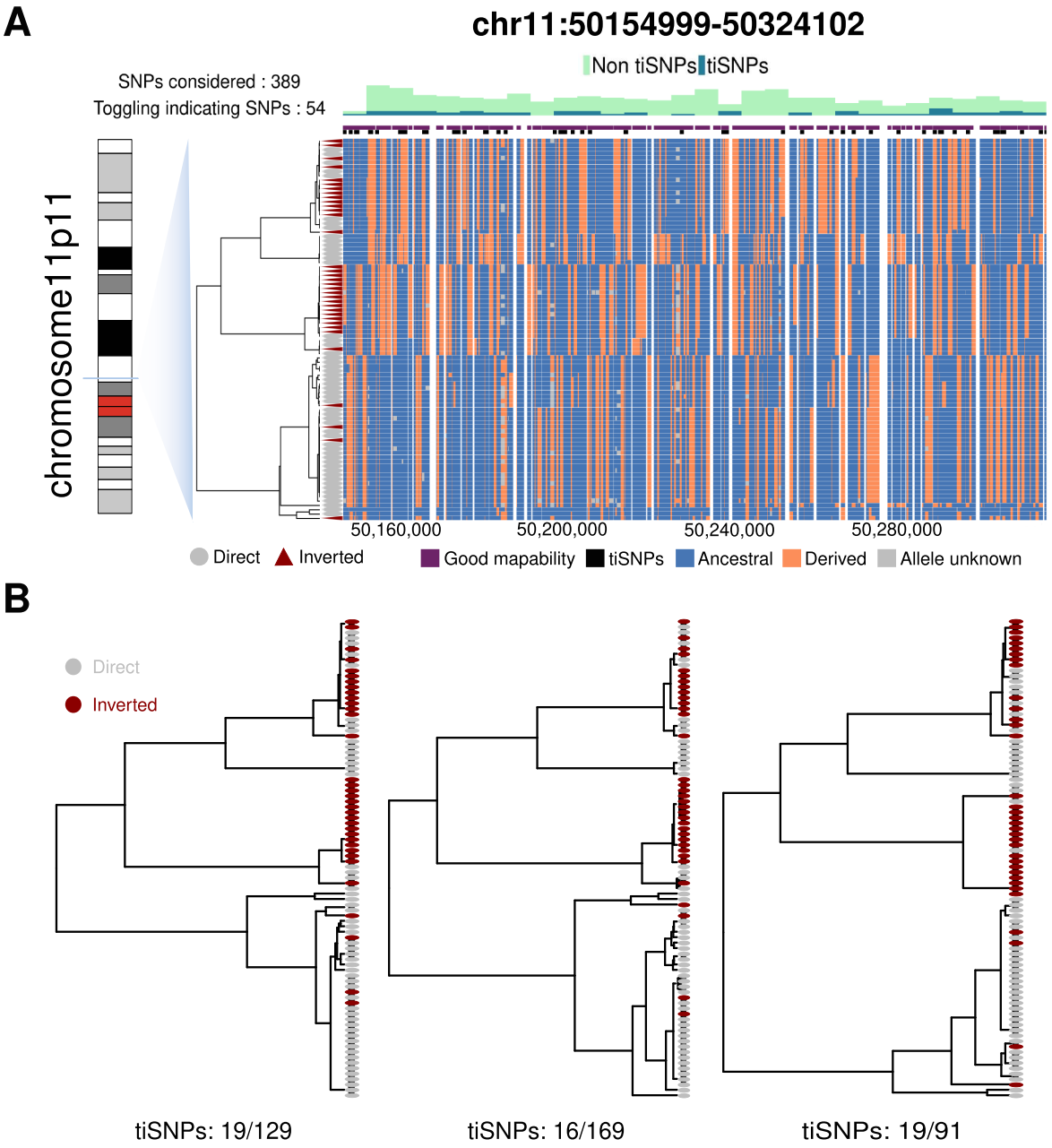


Figure 4.9: Recurrent inversion at 11p11. A. Left: Chromosome ideogram showing position of the inverted locus. Middle: Centroid hierarchical-clustering based dendrogram showing relationships among inverted and non-inverted SNP haplotypes shown on the right. Right: Ancestral (blue) versus derived (orange) SNPs, determined using the Chimpanzee genome (PanTro6). Top: tiSNPs track shown in black. Regions with $\geq$75% mappability are marked in purple. Histograms show the distribution of toggling vs non toggling-indicating SNPs. **B.** Phylogenetic trees constructed independently for the right, middle and left thirds of the whole 11p11 inversion locus. Inverted and non-inverted haplotypes can be seen clustering together in each tree, with tiSNPs distributed across the whole locus.

4.3.5 Recurrent inversions and disease-associated copy number variations

Another interesting aspect related to inversion recurrence is its potential relationship with disease-associated copy number variations (morbid CNVs), also believed to be mediated by NAHR (discussed in detail in Section 1.5.1). We performed an enrichment analysis using morbid CNVs from the Decipher database [143] and observed that the balanced inversions included in our callset, appeared to be significantly co-localized with morbid CNV regions (14% overlap, 2-fold enrichment compared to randomized loci). Notably, when considering only recurrent inversions, the enrichment was even higher (5-fold enrichment, 31% overlap), suggesting a potential association between recurrent inversions and morbid CNVs.

One notable case within this overlap was the inversion overlapping the 7q11.23 deletion/duplication region also known as the Williams-Beuren syndrome (WBS) critical region. It had previously been reported that an inversion predisposes this region to disease formation [33], however, the mutational recurrence of this inversion was previously unknown. We found that this inversion has toggled at least thrice over the course of evolution with an inversion rate of 2.6×10^{-5}, with 19 tiSNPs distributed across this locus. Further studies in patient cohorts may provide insight into whether a subset of these haplotypes act as a pre-mutational state for the syndrome.

Additionally, we analyzed the SD-architecture around inversions using phased assemblies and observed that a recurrent inversion overlapping the 3q29 critical region, may play a protective role by reorienting the morbid CNV triggering SDs in an inverted orientation. We also observed complex haplotype structures for recurrent inversions overlapping the 7q11.23 (WBS) and 15q13.3 microdeletion regions, leading to the hypothesis that these inversions could act as both pre-mutational and protective, depending on whether they reorient the SDs in same or opposite orientation, respectively (Figure 1.6).

4.4 Discussion

This chapter primarily focuses on the tiSNPs-based approach for detecting inversion recurrence, developed as part of a collaborative effort to study recurrent inversion polymorphisms across human populations [13]. Alongside this approach, a complementary haplotype-based approach utilizing phylogenetic analysis contributed to the identification of 32 autosomal and X-chromosomal recurrent inversions. Additionally, an analysis of the Y-phylogenetic trees led to the identification of eight recurrent inversions on the Y chromosome. Together, these 40 recurrent inversions span about 0.6% of the human genome. One of the novel contributions of our study is the quantification of the extent of inversion toggling across human populations, with recurrence rates ranging from 3.4×10^{-6} up to 1.4×10^{-4}. Notably, six of the identified recurrent loci have previously been reported to exhibit toggling in great apes [31]. These findings suggest that inversion recurrence is a more prevalent mutational process compared to other types of genomic variation and has been continuously occurring not only in humans but also in their close relatives for at least the past 15 million years.

We observed a notable enrichment of recurrent inversions on the sex chromosomes, which can potentially be attributed to NAHR- or NHEJ-mediated repair mechanisms outside the XY-homologous pseudoautosomal regions.

With a comprehensive set of inversions and their recurrence status, we were able to further explore the relatively understudied relationship between inversions and pathogenic copy number variations. Our analysis revealed that the previously established link between inversions and morbid CNVs [26, 29, 33, 34, 144] is primarily driven by recurrent inversions. A particularly notable example is the inversion overlapping the 7q11.23 WBS critical region, which is known to predispose carrying individuals to disease [33]. This inversion was identified as a novel recurrent candidate by both of our recurrence detection approaches. One possible explanation for the relationship between recurrent inversions and morbid CNVs can be that inversion recurrence increases the number of heterozygous inversion carriers in the population. Since heterozygous inversions suppress homologous recombination [145], the cellular machinery is forced to rely on error-prone non-homologous repair mechanisms to resolve DNA breaks, thereby increasing the likelihood of deletion and duplication events. This hypothesis is supported by the increased occurrence of inverted duplication and deletion events observed at the 8p23.1 locus, linked to suppressed homologous recombination [146–148]. Interestingly, we found the inversion traversing this locus, showing an AF of 50% in our callset, to have recurred at least 15 times.

Another key factor influencing this relationship could be the alteration of the SD landscape surrounding the inversion due to its repeated toggling, potentially involving different breakpoints. We performed a structural analysis of the flanking region of inversions overlapping the 7q11.23, 3q29, and 15q13.3 critical regions, which suggested that the highly repetitive SD architecture around these inversions creates an ideal environment for NAHR, and consequently, independent mutation events. The impact of these inversions on disease susceptibility depends on how the relative orientation of involved SDs changes within the resulting haplotype. In some cases, an inversion may increase disease risk, while in others, it may offer a protective effect. Notably, for loci exhibiting extensive structural diversity at their flanks, a recurrent inversion could act as protective in certain haplotypes while playing a pre-mutational role in others. With advancements in long-read sequencing technologies and the availability of more refined genome assemblies, providing improved breakpoint resolution, future studies could analyze these loci in patient cohorts with greater precision. This could enable the identification of haplotypes that predispose to or protect against disease, offering new insights into the role of recurrent inversions in disease susceptibility.

In summary, this chapter highlights the widespread yet understudied phenomenon of inversion recurrence in the human genome and describes some of the work we have undertaken in this area [13]. By developing two orthogonal but complementary approaches to identify inversion recurrence, we were able to confidently determine the recurrence status and mutation rates of the analyzed loci. However, considering only the results where two approaches (with different sensitivities) agree, can lead to an underestimation of the true ex-

tent of inversion recurrence. Specifically for the tiSNPs-based approach, the extreme sparseness of Strand-seq data can be a limitation. Moreover, since both methods rely on SNPs for recurrence detection, they inherently have reduced power for detecting recurrence in loci with few SNPs. Additionally, the analyses presented in this chapter are based on a dataset of 41 unrelated samples—a relatively small cohort, particularly for studying population-scale patterns. Owing to that, we use the term "single-event" rather than "non-recurrent" for loci without evidence of recurrence, acknowledging that expanding the dataset with more diverse populations would not only extend the catalog of known inversions but also likely reclassify many (currently) "single-event" loci as "recurrent". Recent advances in the field of pangenomics and the release of the first human pangenome [35], have created new opportunities to achieve this goal. The improved sequence resolution and increased structural diversity captured by a pangenome allow for a more thorough characterization and deeper understanding of inversion recurrence on a larger scale. With this in mind, we have developed a pangenome-friendly recurrence detection method, Pivot, described in detail in Chapter 5, that addresses some of the limitations of the methods discussed in this chapter while leveraging the depth of genomic diversity offered by a pangenome.

Chapter 5

Pivot: Pangenome based analysis of Inversion Toggling

This chapter introduces Pivot, a pangenome-friendly version of the tiSNPs-based approach for inversion recurrence detection described in Chapter 4. The content presented in this chapter is entirely my own work and has not been published yet.

5.1 Motivation

As discussed in Chapter 4, NAHR-mediated inversion recurrence is a widespread phenomenon that has occurred throughout primate evolution [13, 31]. Numerous studies have established a connection between recurrent inversions and genomic regions associated with disease-causing microdeletions and microduplications [13, 26–32]. Chapter 4 presents our investigation of this phenomenon within a framework using a linear reference genome [13]. Specifically, we introduced two complementary approaches for detecting inversion recurrence: a haplotype-based method that employs phylogenetic and coalescent-based recombination graph analyses to infer inversion recurrence, and a toggling-indicating SNPs-based approach that scans inverted regions for SNPs discrepant with a single inversion origin [13]. Both methods are reference genome dependent: the former relies on phased SNPs derived from the integration of Strand-seq and PacBio data, while the latter utilizes Strand-seq read alignments and SNP calls relative to the reference genome [13]. Furthermore, both methods exclusively consider SNPs located within an inversion to assess its recurrence status. Consequently, their sensitivity is reduced when analyzing small inversions or inversions with fewer SNPs.

This chapter introduces a novel inversion recurrence detection approach, Pivot, which is a modified and extended version of the tiSNPs-based method and operates within a pangenomic framework, thereby eliminating reference bias. Rather than relying solely on SNPs, Pivot examines all genetic variants within and surrounding an inverted locus to identify evidence of recurrence, thereby increasing its detection power relative to previous approaches.

Additionally, Pivot does not require SNP or other variant genotypes as input; instead, it works solely with a pangenome graph. In this way, Pivot addresses the limitations of previous approaches while leveraging the extensive genomic variation captured by a pangenome. The graphs generated using the latest high-quality, nearly complete genome assemblies [15], which resolve large inversions that were previously unresolved, present a unique opportunity to analyze inversion recurrence.

5.2 Algorithmic overview

5.2.1 Locating the inverted region in a pangenome

To assess the recurrence of an inversion event using a pangenome graph, the first step is to locate the corresponding region within the graphical structure. Given annotated inversion breakpoints, we can locate this region within the haplotype where the inversion was originally detected. For ease of understanding, we refer to this haplotype as the "reference" haplotype. However, determining the similar region across all other haplotypes in the pangenome is nontrivial due to the absence of haplotype-specific breakpoints. As discussed in Chapters 3 and 4, the flanking sequences of an inverted region play a crucial role in its recurrence. These sequences also contain information about unique characteristics of the specific inversion. Figure 5.1A illustrates a schematic representation of a simple inversion within a pangenome graph, where the inverted region forms a loop-like structure. The flanking inverted repeats serve as entry and exit paths, allowing different haplotypes to traverse the loop in distinct orientations depending on whether they carry the reference or inverted allele (for example, the blue and orange paths in Figure 5.1A). When this loop structure is represented as linear graph paths, specific nodes, which we refer to as "anchor nodes", appear on both the left and right sides of the inverted region, in opposite traversal orientations. Pivot identifies these nodes by scanning the inversion flanks in the graph path of the reference haplotype. In addition to anchor nodes, Pivot tracks unique nodes, termed as "safe nodes", across all haplotypes. Both anchor and safe nodes are formally defined as follows:
Let:

- $G = (V, E)$ be a bidirected graph, where:
  - V is a set of nodes, each with two possible traversal orientations (sides): forward $(+)$ and reverse $(-)$. Each node is represented as a tuple (v, o), where v indicates the node ID and o represents the traversal orientation.
  - $E \subseteq \{((u, o_u), (v, o_v)) \mid u, v \in V;\ o_u, o_v \in \{+, -\}\}$ is a set of edges connecting nodes.
- A haplotype path through G is defined as a sequence $P = [(v_1, o_1), (v_2, o_2), \ldots, (v_n, o_n)]$ of $n \in \mathbb{N}$ nodes, where:

  - $\forall i \in \{1, \ldots, n\}$, $v_i \in V$ and $o_i \in \{+, -\}$.
  - $\forall i \in \{1, \ldots, n-1\}$, $((v_i, o_i), (v_{i+1}, o_{i+1})) \in E$.

- Let $\mathcal{P} = \{P_1, P_2, \ldots, P_m\}$ be a set of m haplotype paths through G.
- Let $P_r \in \mathcal{P}$ be the reference haplotype path.
- For any path $P = [(v_1, o_1), \ldots, (v_n, o_n)]$ and any node $v \in V$, define:

$$\text{count}_P(v, +) = |\{i \in \{1, \ldots, n\} \mid v_i = v \text{ and } o_i = +\}|$$

$$\text{count}_P(v, -) = |\{i \in \{1, \ldots, n\} \mid v_i = v \text{ and } o_i = -\}|$$

$$\text{count}_P(v) = |\{i \in \{1, \ldots, n\} \mid v_i = v\}|$$

- Let $W \subseteq V$ be the set of nodes in a user-specified window in the reference path P_r.
- Let $l \in \mathbb{N}$ be a user-defined minimum length threshold for safe nodes.

Then:

$$v \in W \text{ is an anchor node } \iff \text{count}_{P_r}(v, +) \geq 1 \ \wedge \ \text{count}_{P_r}(v, -) \geq 1$$

$$v \in V \text{ is a safe node } \iff (\forall P_k \in \mathcal{P}, \ \text{count}_{P_k}(v) = 1 \ \wedge \ \text{length}(v) \geq l)$$

That is, an anchor node is a node traversed by the reference haplotype path, both upstream and downstream of the inverted locus, with opposite traversal orientations, while a safe node is a node that represents a sequence longer than a specified value and is traversed by each haplotype path represented in the pangenome graph exactly once.

The anchor nodes represent the SDs flanking the inversion and allow localization of the inverted region in haplotypes lacking precise breakpoint coordinates. Given that inversion length positively correlates with the length of the flanking inverted repeat [13], and that repeat structures vary significantly across loci, the length of the search window for anchor nodes is set as a user-defined parameter. The node IDs and corresponding traversal orientations for safe nodes are recorded with respect to the reference haplotype path (**Figure 5.1**). Additionally, each safe node is annotated with its relative position to the inverted locus—that is, left or right—in the reference haplotype. Since node lengths can vary depending on graph construction parameters and used assemblies, in certain cases, there might be no nodes matching the safe nodes minimum length criterion. As a solution, Pivot allows users to provide upper and lower bounds for the minimum length. It begins the safe node search at the upper bound and iteratively reduces the threshold until qualifying safe nodes are found or the lower bound is reached. Once anchor and safe nodes are identified, Pivot locates the inverted locus across all haplotypes in the pangenome graph by screening haplotype paths for the anchor node structure. In cases where assembly breaks result in multiple contigs covering a region, Pivot extracts relevant segments from each contig using anchor nodes

and concatenates them to construct a "pseudo-contiguous" path, as described in the next section.

5.2.2 Synchronizing the region of interest across the whole haplotype panel

Pivot operates using the Graphical Fragment Assembly (GFA) format representation of the graph. The contigs reported in a GFA file are not necessarily orientation-wise synchronized with each other. In other words, if a specific genomic region is extracted from two different contigs and the extracted sequences appear to be in opposite orientations, it is not possible to determine—based solely on the sequences—whether they are truly inverted with respect to each other or if one of the paths is simply reported as the reverse complement in the GFA. While anchor nodes facilitate the extraction of the region of interest across all haplotypes in the pangenome, the orientation of the extracted paths must be synchronized before further analysis. This is where safe nodes play a crucial role. Pivot looks beyond the anchor structure surrounding the inversion in non-reference haplotypes and attempts to locate the nearest safe nodes on either side of the inversion. Since safe nodes occur exactly once in each haplotype and reside outside the inverted region, they provide a reliable reference for orientation alignment. If two contigs are reported in the same orientation, the traversal direction and relative position (left/right) of the safe nodes should be consistent. Pivot leverages this property to compare the occurrence pattern of safe nodes in each non-reference haplotype against the reference haplotype, ensuring that all extracted haplotype paths are correctly oriented (Figure 5.1B).

For example, consider a reference haplotype where S_1 represents a safe node traversed in forward orientation (+) and located on the left side of the annotated inversion while S_2 represents another safe node traversed in reverse orientation (-) and located on the right side of the annotated inversion. The possible scenarios for other haplotypes, as illustrated in Figure 5.1B, are as follows:

1. The extracted path has the same position and direction for both safe nodes as the reference haplotype. In this case, no adjustments are needed, as the path is already synchronized (Figure 5.1B, H1 and H3).

2. The position and direction of the safe nodes are inverted relative to the reference haplotype. This indicates that the extracted path must be flipped (reverse-complemented) to achieve synchronization (Figure 5.1B, H2).

3. Relevant segments from multiple contigs are extracted. In this case, each segment must be individually synchronized using the rules described in points (1) and (2) and concatenated to construct a pseudo-contiguous path (Figure 5.1B, H4).

Once this synchronization step is completed, Pivot has all the necessary data to proceed to the next phase: identifying evidence of inversion recurrence

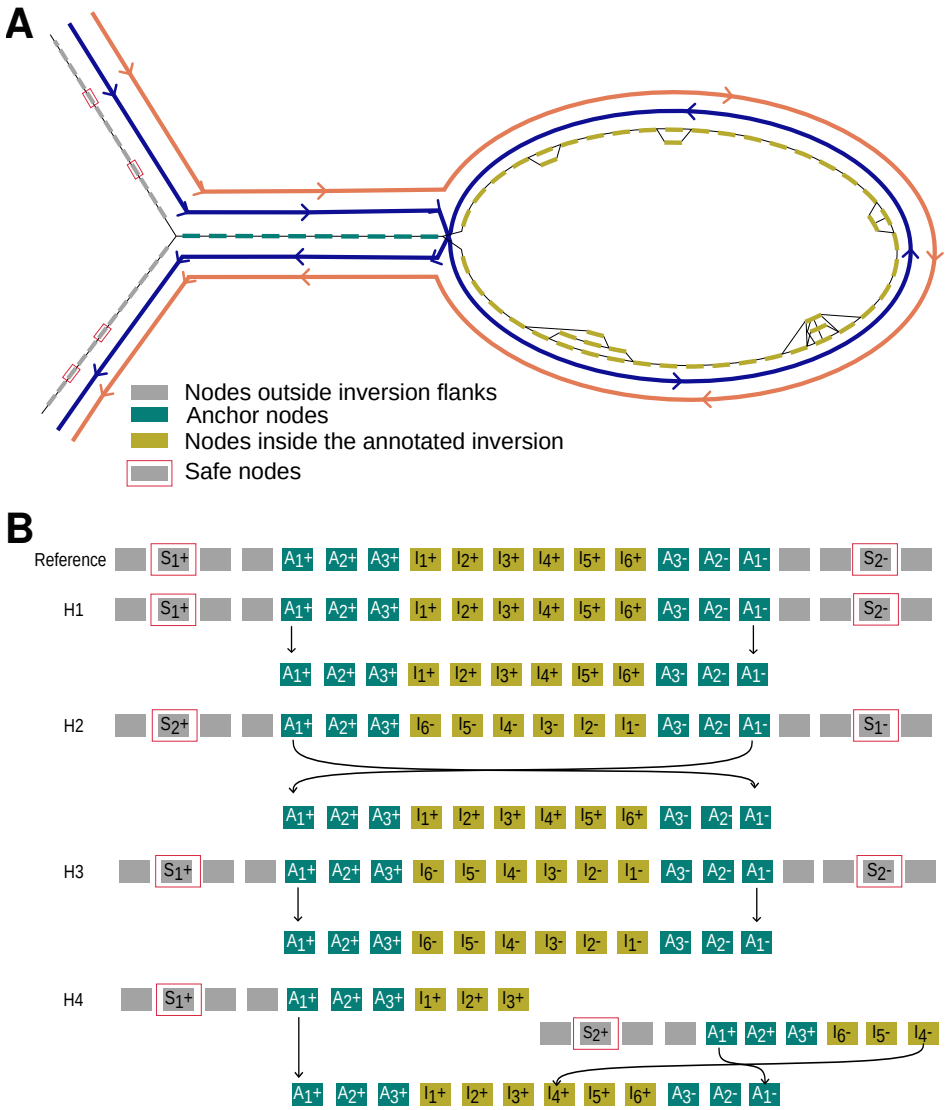
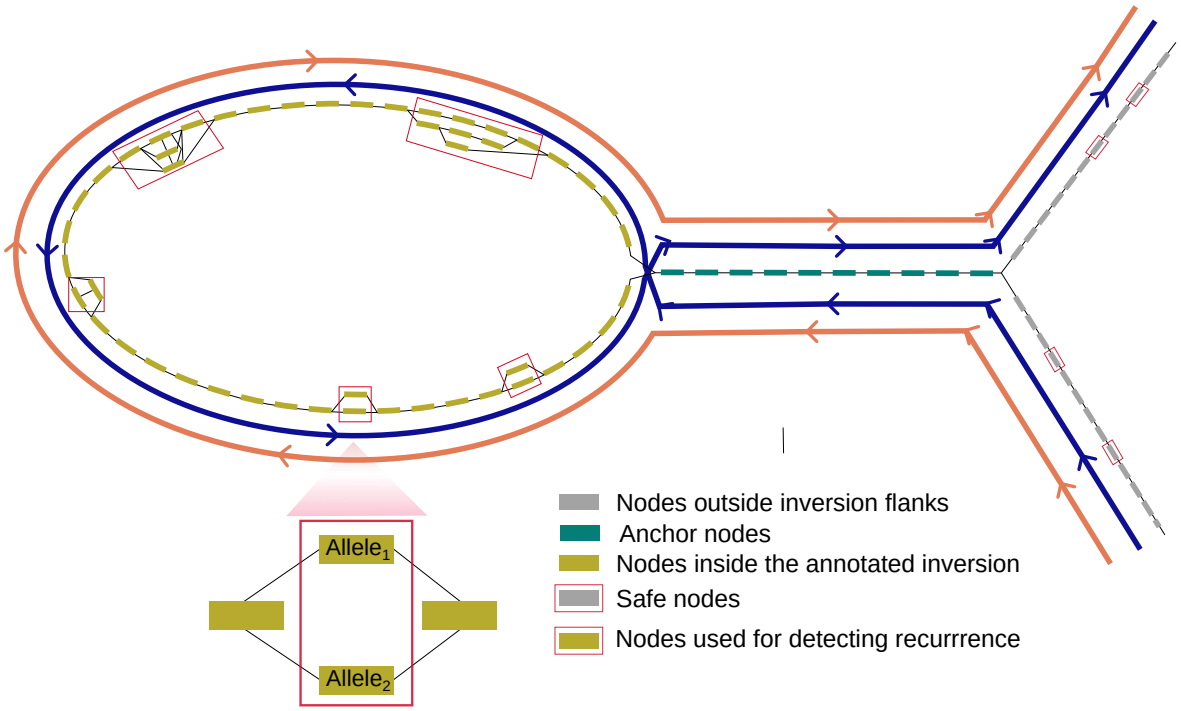


Figure 5.1: Extracting an inverted region from a pangenome graph. A. A schematic representation of an inversion within a pangenome graph. Blue and orange paths indicate haplotypes which are oppositely oriented in the inverted region. **B.** S_n, A_n, and I_n denote the node IDs corresponding to safe nodes, anchor nodes, and nodes within the inverted region, respectively. The symbol "+" indicates forward and "–" indicates reverse traversal direction of a node. H1–H4 represent four haplotype subpaths extracted by Pivot using the anchor nodes. Below each extracted haplotype subpath, the corresponding orientationally synchronized subpaths are shown. Arrows indicate whether an orientation flip is required: straight downward arrows indicate that the subpath is already aligned with the reference haplotype, while curved arrows denote an orientation flip. H1 and H3 are already synchronized with the reference. In contrast, comparing the safe nodes of H2 with the reference reveals the need for an orientation flip. For H4, the inverted region is traversed by two assembly contigs, therefore, each contig is independently synchronized with the reference—left one kept in its original orientation, and the right one flipped. The resulting subpaths are then concatenated.



	Forward Present	Forward Absent	Reverse Present	Reverse Absent
$Allele_1$	>0	0	0	>0
$Allele_2$	0	>0	>0	0

Single Event

	Forward Present	Forward Absent	Reverse Present	Reverse Absent
$Allele_1$	>0	>0	>0	>0
$Allele_2$	>0	>0	>0	>0

Recurrence Evidence

Figure 5.2: Identifying toggling indicating nodes.

5.2.3 Finding evidence of recurrence

The theoretical framework for recurrence detection employed by Pivot builds upon the principles of the tiSNPs-based approach described in Chapter 4. However, rather than focusing solely on SNPs, Pivot extends its framework by incorporating additional variant classes to potentially enhance sensitivity for inversion recurrence detection.

During this phase, Pivot screens all variants within the inverted region and its flanks. As expected for bi-allelic SNPs, in case of a single-event, each variant allele should segregate exclusively with either non-inverted or inverted haplotypes. Conversely, the presence of an allele in both haplotype groups serves as evidence of inversion recurrence. Pivot scans the extracted haplotype paths to identify graph nodes representing such variant alleles. To generalize the concept of tiSNPs, we introduce the term “toggling-indicating nodes” (tinodes) to represent such nodes. As mentioned in Section 1.6, genomic variants are represented as “bubbles” in a pangenome graph [96]. Pivot employs BubbleGun [149] to extract all bubbles represented in the graph. It then traverses the extracted paths and, for each bubble encountered on the way, records the occurrence counts for each inside node. Given that Pivot additionally considers multi-allelic variants, its counting strategy slightly differs

from the tiSNPs-based approach. For each variant allele (represented as inside nodes), Pivot records:

- the number of times the allele appears in forward and reverse orientations
- the number of times the allele is absent in forward and reverse orientations

If a variant allele node has a count $\geq$1 in all four categories, it is classified as a tinode (Figure 5.2). In order to avoid false signals of recurrence generated from multiple traversals of a bubble in the same haplotype path, Pivot only considers bubbles traversed at most once per haplotype in the pangenome graph. The cumulative evidence across the entire inverted region, that is, four occurrence counts for each inside node is recorded, which is then used to determine the recurrence status of each inversion.

To account for the diversity of inversion loci, Pivot does not apply a single, fixed threshold for recurrence classification. Instead, it produces a detailed output across multiple thresholds (ranging from 1 to 4) for the minimum required value for each of the four occurrence counts described above, to classify a node as a tinode. For notational clarity, we refer to these as tinodes$_n$, where n indicates the minimum count threshold used. Pivot's output for an analyzed locus includes the number of analyzed nodes, number of detected tinodes$_{1-4}$, their respective cumulative lengths, and a breakdown of variant types (SNPs, indels or SVs) contributing to the recurrence signal. An important measure for assessing the confidence in the observed recurrence signal is the minimum number of haplotypes that must be removed from the analyzed cohort to entirely eliminate the recurrence signal—that is, to ensure the absence of any tinodes$_1$. For each analyzed node, this minimum is selected from four distinct haplotype sets (contributing to forward present, forward absent, reverse present, and reverse absent occurrence counts, as illustrated in Figure 5.2). In some cases, more than one haplotype sets yield the same minimal value, resulting in a non-trivial optimization problem. By considering all haplotype sets that yield the minimum count, Pivot provides an upper-bound estimate on the minimum number of haplotypes contributing to the recurrence signal. Additionally, Pivot provides graph quality statistics, such as the number of contig breaks within the region and the number of haplotypes where the inverted region could be confidently located. This detailed per-locus information helps to take different factors into account while determining the recurrence status of each locus. It helps to filter out potential noise, such as cases where a large number of tinodes originate from a single haplotype, which might be the result of an assembly error. At the same time, by using all variants inside the inverted region, this approach enhances sensitivity compared to our previous two approaches [13], allowing detection of recurrence for rare inversions or inversions with a low recurrence rate.

5.2.4 Impact of pangenome graph quality on downstream analyses

The pangenome graph builder (PGGB) is a tool designed to construct pangenome graphs from multiple genome assemblies [150]. The graph construction process in PGGB starts by

performing an all-vs-all alignment of the whole-genome nucleotide sequences using wfmash [151], which is a sequence aligner optimized for genome comparison. The resulting pairwise alignments are then passed to seqwish [152], which converts them into a variation graph that represents the sequence variation between the input sequences. Since seqwish's graph generation is lossless, the initial graph may be complex and contain redundant information. Therefore, it is further processed by smoothxg [150], which simplifies and normalizes the pangenome graph by locally realigning sequences. The final output is a GFA-format graph with each input genome represented as a path. Among the state-of-the-art pangenome graph construction algorithms available [35, 150, 153–157], only PGGB allows back traversal of the graph paths, that is, a path can traverse the same nodes more than once and in opposite orientations, for example in the case of repetitive sequences. Since Pivot locates inversions using the inverted flanking repeats represented by the same set of nodes traversed in opposite orientations (anchor nodes), we use PGGB graphs for all analyses presented in this chapter.

The quality of input data plays a vital role in determining the overall performance of any computational method. For methods working in a pangenome graph space, the most important contributing factor is the graphical layout which is primarily influenced by the graph construction parameters. Even with the same genome assemblies, different parameter settings can lead to vastly different graph structures, affecting the resolution and complexity of variant representation. For instance, for PGGB, three key parameters play a central role in shaping the graph structure [150]:

- **"-p"**: Defines the percent identity required for wfmash alignment. Lower values increase alignment sensitivity, leading to more compressed graphs.
- **"-s"**: Sets the minimum sequence length for all-vs-all alignment by wfmash. Higher values produce graphs with longer collinear regions, while lower values increase graph complexity, particularly in repetitive regions.
- **"-k"**: Specifies the minimum match length, below which all matches in wfmash alignments are filtered out during merging by seqwish. Smaller values retain more matches, increasing graph complexity, particularly in challenging regions like alpha-satellites and centromeres, whereas larger values simplify the graph.

Another critical factor affecting pangenome graph quality is the quality of input genome assemblies. In highly repetitive regions, assembly algorithms often introduce breaks, resulting in fragmented sequence representations of individual genomes as separate contigs. Although advancements in sequencing and assembly techniques have significantly reduced such breaks, the likelihood of sequencing an entire haplotype as a continuous stretch, remains low. Such breaks occurring within an SV region add another layer of complexity to the SV analysis. This is particularly challenging in case of inversions because the independent contigs might not be synchronous in terms of the traversal orientation.

As mentioned before in Section 1.6, the graphical layout complexity of the bubble structures varies across different variant types, with simple variation, such as SNPs, represented

as “simple bubbles”, while complex graphical structures, called “superbubbles”, representing complex structural variation. Therefore, the analyses relying on variant information represented in the pangenome graph are highly influenced by the representation accuracy of these bubbles. Pivot attempts to minimize the biases introduced by graph construction parameters, assembly quality, and variant complexity by conducting a comparative analysis across four distinct settings when analyzing inversion recurrence:

1. Considering all variants and all haplotypes to detect recurrence evidence. This is the most lenient setting using all the information represented in the pangenome graph.
2. Considering all haplotypes but restricting only to variants represented as simple bubbles.
3. Considering all variants (regardless of bubble complexity) but excluding haplotypes with contig breaks inside the inverted locus.
4. Restricting the analysis to simple bubbles and haplotypes with a single, unbroken contig covering the whole inverted locus. This is the strictest setting which minimizes the chances of observing false recurrence signals.

5.3 Results

For the set of analyzed inversions, we used the inversion callset of 292 balanced inversions reported in Porubsky et al. [13], described in detail in Chapter 3 and Chapter 4. The rationale behind choosing this callset lies in the availability of high-quality inversion recurrence ground-truth information, based on a detailed analysis using two orthogonal approaches (Chapter 4). Since this callset was based on the GRCh38 reference genome, we used the GRCh38 path in the pangenome graph as the reference path for Pivot’s analysis.

5.3.1 Human Pangenome Reference Consortium (HPRC) graphs

In the first phase of our analysis, we applied Pivot to the PGGB graphs from the first HPRC release [35]. These graphs were constructed using a cohort of 44 individuals with diverse geographical ancestries and two linear reference genomes: GRCh38 and T2T-CHM13. As done for the tiSNPs-based approach [13] (Section 4.3.3, Figure 4.6), as a negative control, we analyzed the 706 kbp inversion at 17q21.31 locus (chr17:45585159-46292045), reported to have occurred only once in the last 2.3 million years [32, 34, 76]. Table 5.1 shows some of the statistics produced by Pivot for this inversion under each of the four settings described in Section 5.2.4.

In the most flexible setting—that is, using all bubbles and contigs for the analysis—Pivot found 155, 41, 19, and 8 tinodes$_{1,2,3,4}$, respectively. The sequence covered by these tinodes ranged from 16 to 210 bp, accounting for only 0.04% and 0.58%, respectively, of the total analyzed length. For the strictest setting (Section 5.2.4)—considering only simple bubbles

Table 5.1: Pivot's statistics for the 17q21.31 inversion.

	haps	analyzed nodes	analyzed length	$tinodes_1$	$tinodes_1$ length	$tinodes_2$	$tinodes_2$ length	$tinodes_3$	$tinodes_3$ length	$tinodes_4$	$tinodes_4$ length	sig_haps
all contigs, all bubbles	90	10292	36055	155	210	41	67	19	34	8	16	16
all contigs, simple bubbles	90	8636	11113	30	30	0	0	0	0	0	0	3
unbroken contigs, all bubbles	81	10024	35516	57	87	34	60	13	24	0	0	11
unbroken contigs, simple bubbles	81	8390	10566	2	2	0	0	0	0	0	0	1

Table 5.2: Pivot's statistics for the 8p23 inversion.

	haps	analyzed nodes	analyzed length	$tinodes_1$	$tinodes_1$ length	$tinodes_2$	$tinodes_2$ length	$tinodes_3$	$tinodes_3$ length	$tinodes_4$	$tinodes_4$ length	sig_haps
all contigs, all bubbles	90	88891	214237	36005	73982	24994	48919	19394	41993	15683	31048	67
all contigs, simple bubbles	90	76511	90447	29468	29705	19978	20115	15316	15410	12308	12388	67

[1] haps: number of haplotypes analyzed
[2] sig_haps: upper bound on the minimum number of haplotypes needed to be removed to get rid of the recurrence signal

and haplotypes with one contig traversing the full inverted locus—Pivot reported only 2 tinodes$_1$, accounting for 0.02% of the analyzed sequence. For the two intermediate settings, a similar behavior was observed, with the overall percentage of tisequence (combined sequence of tinodes) across the four settings being $\leq$0.58%.

Comparison of results across the four settings suggests that the impact of bubble structure on the analysis tends to be stronger than the impact of broken assembly contigs. This is indicated by a larger drop in Pivot's upper-bound estimate of the minimum number of haplotypes that must be removed from the cohort to eliminate the recurrence signal (sig_haps in Table 5.1) when the analysis is restricted from using all bubbles to only simple bubbles, compared to the smaller change observed when excluding haplotypes with contig breaks. For an inverted locus, there are three potential explanations for such behavior: First, this might indicate that the locus has accumulated a lot of complex variation, represented as superbubbles, and this variation contains evidence of its recurrence. Second, the recurrence of another nested variant, represented by a superbubble—for example a recurrent SNP located inside the inverted region—can generate such a signal. Third, this could suggest that some of the complexity represented in the graphical structure of the respective locus is not inherent to the locus but generated during the graph construction process. Given that the locus in question is a well-known single-event inversion, the second and third explanations are more likely to be true in this case. The relatively lower impact of including eight haplotypes with contig breaks inside the inverted locus further highlights that the pseudo-contiguous haplotype construction implemented in Pivot works as expected and is not introducing noise into the analysis.

As a positive control, we selected the 5.3 Mbp inversion located at chromosome 8p23.1 (chr8:7301024-12598379), which was found to be highly recurrent by our orthogonal tiSNPs- and haplotype-based approaches [13] (discussed in Chapter 4), with a mutation rate of 1.11×10^{-4} inversions per generation. Table 5.2 presents some of the statistics produced by Pivot for this inversion under each of the four settings described in Section 5.2.4. In the most flexible setting Pivot identified tinodes ranging from 15,683 (tinodes$_4$) to 36,005 (tinodes$_1$). The base pairs covered by these tinodes ranged from 31,048 to 73,982 bp, accounting for 14.5% and 34.5%, respectively, of the total analyzed length. For this inversion, the region of interest was traversed by a single contig across all 90 haplotypes, therefore, only the bubble-based stratification is presented in Table 5.2.

In contrast to the 17q21.31 inversion, even when limiting the analysis to simple bubbles, Pivot still reported a large number of tinodes, ranging from 12,308 to 29,468, accounting for 13.7% and 32.8% of the analyzed sequence. Additionally, in both settings, 67 sig_haps, that is, 74% of the haplotypes, were observed, consistent with the high mutation rate of this inversion. These findings indicate that the recurrence signal detected by Pivot is clearly pronounced for true recurrent inversions compared to single-event inversions. Moreover, the persistence of this pronounced behavior even under stricter settings highlights both the sensitivity and robustness of Pivot.

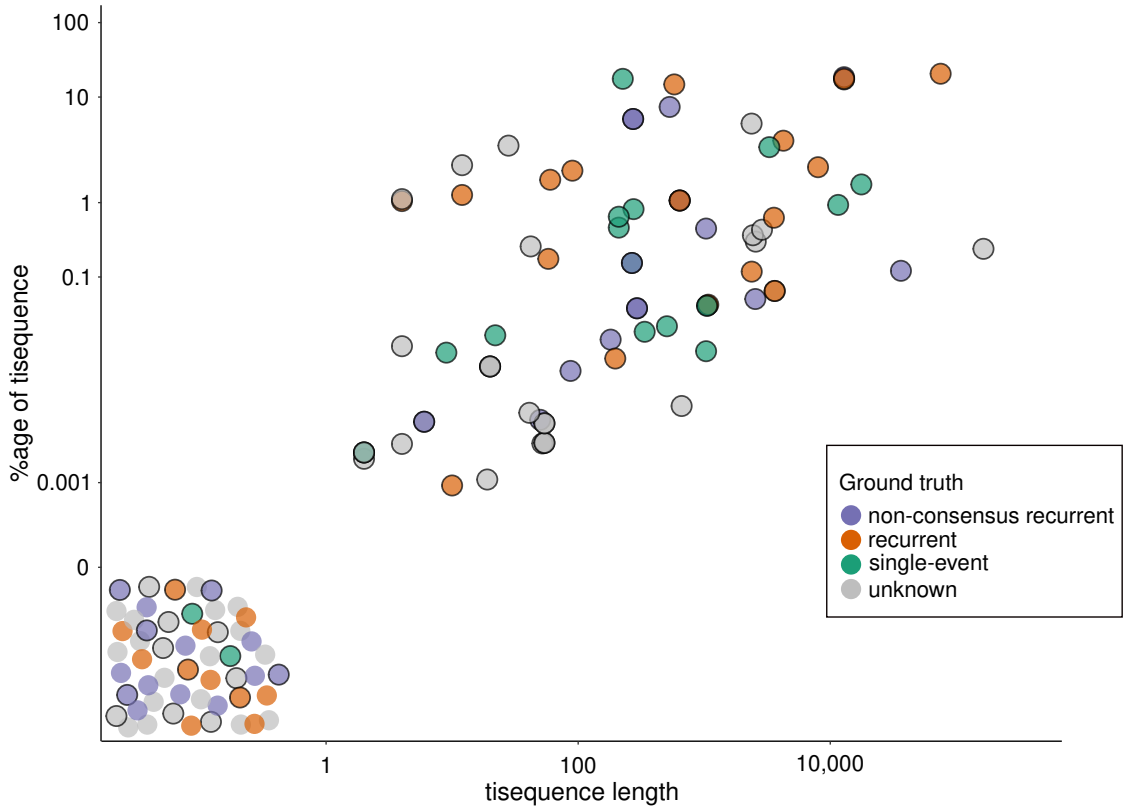


Figure 5.3: Pivot's inversion recurrence analysis using HPRC graphs. Plot showing Pivot's assessment of recurrence using HPRC graphs. The horizontal axis shows the sum of length of observed tinodes and the vertical axis represents what percentage of the total analyzed sequence indicated toggling, that is, (tisequence length/analyzed sequence length)$\times$100. All 129 inversions for which Pivot was able to locate the inverted region in the graph using anchor nodes are shown. The 95 inversions that could be assessed for recurrence are indicated by black outline. To enhance visibility, a jitter has been applied to the cluster of points in the bottom-left corner of the plot, corresponding to 0 on both the horizontal and vertical axes. Recurrent: inversion found recurrent by both haplotype- and tiSNPs-based approaches. Single-event: no inversion recurrence evidence found by either of the two approaches. Non-consensus recurrent: inversion found recurrent by only one of the two approaches. Unknown: previous assessment for inversion recurrence status was not available.

Genome-wide recurrence analysis

Overall, with a 1 Mbp flanking region used to search for the anchor structure and a minimum safe node length of 15 bp, Pivot successfully identified the anchor structure and extracted inverted loci across the haplotype panel for 129 out of 292 inversions. Based on the dot plot alignments of the flanking sequences, we observed that 125 of the remaining 163 inversions (77%), were flanked by short inverted repeats ($\leq$10 kbp). The collapsing of the graph in repetitive regions, inherently reduces the ability to identify such loci using anchor nodes. Interestingly, these 163 inversions contained 13 out of 35 X-chromosomal inversions, implying that the graph structure did not depict the inversion neighborhood for this

chromosome with sufficient granularity. This interpretation is further supported by the fact that, for all 22 X-chromosomal inversions where Pivot successfully located the anchor structure, the analysis could not proceed further due to the absence of safe nodes necessary to synchronize the orientation of the extracted paths. To address this, we re-ran Pivot with the minimum safe node length reduced to 5 bp, which allowed the identification of some safe nodes across the entire X chromosome graph. However, this adjustment still did not facilitate further analysis, as the safe nodes did not belong to the same contig as the inversion in any haplotype, except for GRCh38 and T2T-CHM13, both of which are represented as single contigs. A similar trend was observed for Y-chromosomal inversions. Although Pivot extracted the inverted locus across all haplotypes for 11 out of 13 cases, no safe nodes were found even after reducing the minimum allowed length to 5 bp.

After additionally removing one inversion on chromosome 2 where no non-repetitive bubbles were found inside the inverted locus, a total of 95 inversions were analyzed for recurrence (Figure 5.3, outlined points). Among these, 20 inversions had been previously classified as recurrent by both the tiSNPs- and haplotype-based approaches [13] (outlined orange points in Figure 5.3). Pivot detected some evidence of recurrence (at least one tinode_1) for 17 (85%) of these inversions. For the remaining three inversions, the maximum number of haplotypes contributing to the haplotype configuration counts across all bubbles was relatively low (10, 16 and 49 out of 90), thereby limiting the power to detect recurrence.

Additionally, Pivot identified recurrence evidence for 16 out of 21 inversions that had previously been classified as recurrent by either the tiSNPs- or haplotype-based approach, but not both (outlined purple points in Figure 5.3) [13]. Furthermore, Pivot detected 44 recurrence candidates for which prior methods had not reported any evidence of recurrence (outlined green and gray points in Figure 5.3). Among these, 15 showed strong evidence of recurrence ($\geq$50 tinodes_1 and $\geq$1 tinodes_4). These findings suggest that Pivot exhibits higher sensitivity for inversion recurrence as compared to our previous approaches.

5.3.2 Human Genome Structural Variation Consortium (HGSVC) graphs

Soon after HPRC released the first human pangenome reference [35], HGSVC, in its production phase 3 (HGSVC3), produced nearly complete genome assemblies for 65 individuals [15]. This enabled the analysis of inversion recurrence in a larger cohort using pangenome graphs derived from higher-quality genome assemblies compared to those from HPRC.

Graph construction

Unlike HPRC, HGSVC did not provide ready-to-use graphs, so the first step was to generate PGGB graphs. Before constructing the graphs, we assigned each assembly contig to a chromosome using alignment quality statistics derived from minimap2 [158] alignments against the T2T-CHM13 reference genome. After contig assignment, we generated chromosome-specific PGGB graphs using the parameters -s = 50k, -p = 98%, and -k = 79, as recommended

by the PGGB developers. These parameters differ from those used for the HPRC graph construction (-s = 100k, -p = 98%, and -k = 311) [35]. The rationale for this adjustment (detailed in Section 5.2.4) is that higher-quality genome assemblies allow for fine-tuning of graph construction parameters to produce a more granular graph. The resulting graphs included a total of 132 haplotypes—130 from HGSVC3 samples and two reference genomes, GRCh38 and T2T-CHM13.

Pivot's recurrence assessment

After generating the PGGB graphs, we analyzed the same set of 292 inversions, examined in the HPRC graphs, using Pivot. Using a 1 Mbp flanking region for anchor node search and a minimum safe node length of 15 bp, Pivot successfully identified the anchor structure for 129 out of 292 inversions. Among these, 110 were also located in the HPRC graphs. Of the 19 inversions that could not be located, 11 belonged to chromosome X and six to chromosome Y, suggesting that even with improved assembly quality, the graphical representation still struggles to accurately depict inversions and their surrounding regions on sex chromosomes. The remaining two small inversions were on chromosome 5 and had flanking inverted repeats shorter than 1 kbp. The 19 inversions that were identified in the HGSVC3 graphs, but not in the HPRC ones, comprised 17 autosomal and two X-chromosomal inversions. Notably, 58% (11/19) of these were flanked by long inverted repeats ($\geq$5 kbp).

We observed a decrease in the number of extracted inversion loci that could not be further analyzed due to the absence of safe nodes, dropping to 15 out of 129 (11.6%), compared to 32 out of 129 (25%) in the HPRC graphs, hence enabling the analysis of more inversions overall. After excluding two additional inversions that lacked non-repetitive bubbles, we proceeded with 112 (107 autosomal, 5 Y-chromosomal) inversions for recurrence analysis. Among these, 21 had previously been identified as recurrent by both tiSNPs- and haplotype-based approaches (outlined orange points in Figure 5.4) [13]. Pivot found recurrence evidence ($\geq$1 tinodes$_1$) for 19 of these (90%). For the remaining two, the detection power was limited—one involved only four (out of 132) haplotypes contributing to haplotype configuration counts, while the other contained only eight nodes (12 bp) that were part of any bubble.

Additionally, the analyzed set included 26 inversions previously classified as recurrent by only one of the tiSNPs-, haplotype-, or Y-phylogeny-based approaches (outlined purple points in Figure 5.4) [13]. Pivot identified recurrence evidence ($\geq$ tinodes$_1$) for 22 of them (85%). Furthermore, Pivot detected recurrence evidence ($\geq$1 tinodes$_1$) for 52 loci where none of the previously used approaches had reported recurrence evidence (outlined green and gray points in Figure 5.4). Among these, 22 (42%) showed particularly strong recurrence signals ($\geq$50 tinodes$_1$ and $\geq$1 tinodes$_4$).

Figure 5.4 depicts a noticeable change, where the previously dense cluster of inversions showing no signs of recurrence (bottom-left of Figure 5.3) became less pronounced. This shift mainly included inversions where prior indications of recurrence existed (orange and

purple points) but were not classified as recurrent using HPRC graphs. Moreover, the recurrence evidence became quantitatively stronger after using the HGSVC3 graphs, as seen in the increased fraction of the analyzed sequence indicating recurrence (tisequence), particularly for inversions with prior recurrence evidence. This is reflected in the tighter clustering of orange points towards the upper right corner of Figure 5.4, compared to their more scattered distribution in Figure 5.3. These improvements likely stem from the higher-quality, more contiguous genome assemblies generating better graphical layout. Another potential contributing factor is the larger number of individuals included in the analyzed cohort.

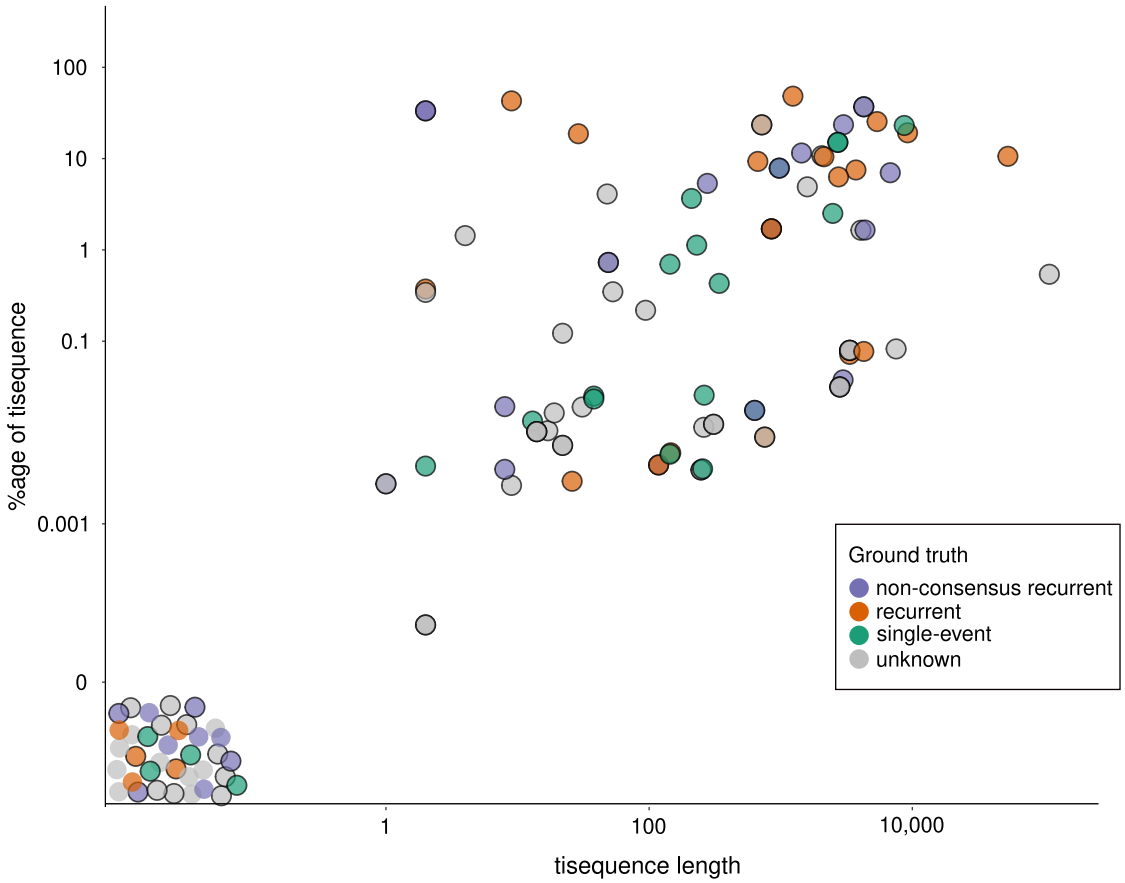


Figure 5.4: Pivot's inversion recurrence analysis using HGSVC3 graphs. Plot showing Pivot's assessment of recurrence using HGSVC3 graphs. The horizontal axis shows the sum of length of observed tinodes and the vertical axis represents what percentage of the total analyzed sequence indicated toggling, that is, (tisequence length/analyzed sequence length)$\times$100. All 129 inversions for which Pivot was able to locate the inverted region in the graph using anchor nodes are shown. The 112 inversions that could be assessed for recurrence are indicated by black outline. To enhance visibility, a jitter has been applied to the cluster of points in the bottom-left corner of the plot, corresponding to 0 on both the horizontal and vertical axes. Recurrent: inversion found recurrent by both haplotype- and tiSNPs-based approaches. Single-event: no inversion recurrence evidence found by either of the two approaches. Non-consensus recurrent: inversion found recurrent by only one of the two approaches. Unknown: previous assessment for inversion recurrence status was not available.

5.3.3 Recurrent inversions overlapping disease-critical regions

In order to find potential disease associations of the novel recurrent candidates (with weak or no recurrence evidence observed before) identified using the HGSVC graphs, we computed their overlap with known morbid CNV regions [13, 80] and observed 18 inversions located within 50 kbp of a known morbid CNV locus.

Smith-Magenis/Potocki-Lupski critical region

One of the notable cases include a 1.56 Mbp inversion overlapping the 17p11.2 microdeletion/microduplication critical region. Recurrent 3.7 Mbp microdeletions at this locus are associated with a contiguous gene syndrome known as Smith-Magenis syndrome (SMS) [159]. SMS is a complex neurobehavioral disorder characterized by mild to moderate intellectual disability, sleep disturbance, craniofacial anomalies, and developmental delay [159, 160]. It is considered to be one of the most frequently observed human microdeletion syndromes [161, 162]. The reciprocal 3.7 Mbp, preferentially paternal, microduplication of the same locus results in another developmental delay disorder known as Potocki-Lupski syndrome (PTLS). PTLS shows a milder clinical phenotype than its deletion counterpart [159, 161, 163]. SMS has an estimated prevalence of 1 in 15,000 to 25,000 births [162]. Most of the phenotypic traits associated with SMS are caused by the haploinsufficiency of the retinoic acid-induced 1 gene (RAI1), resulting either from the 17p11.2 microdeletion encompassing the RAI1 gene (90% of the cases) or its mutation in patients without the deletion (10%) [162, 164, 165]. Differently sized deletions with varying breakpoints have been observed in SMS patients, however, approximately 70% of the patients carry a 3.7 Mbp long deletion of the same genetic markers, referred to as the "common" SMS deletion [161–163].

NAHR between large and complex flanking repeats, called SMS-REPs, is considered to be the major rearrangement mechanism mediating the SMS deletions and reciprocal duplications [66, 166, 167]. There are three copies of SMS-REPs: a proximal, a distal, and a middle SMS-REP (Figure 5.5A). The proximal copy (SMS-REPP) is the longest (approximately 256 kbp) and is located in the same orientation as the distal copy (SMS-REPD) which is about 176 kbp long [166, 167]. Both these copies share a high sequence identity (about 98.7%) and mediate the 3.7 Mbp long SMS common deletion and reciprocal common duplication via NAHR ([166, 167]. The middle SMS-REP copy (SMS-REPM) is located between the proximal and distal copies and is inverted [166, 167]. It is about 241 kbp long and is hypothesized to have been derived from the SMS-REPP. Each of these three SMS-REP copies have been reported to contain 14 genes or pseudogenes [166]. Based on the observed involvement of this locus in several other genomic rearrangements [168–171] alongside studies reporting somatic mosaicism for SMS deletions [172], the 17p11.2 locus is considered to be a highly unstable region in the human genome. This instability can potentially be explained by the presence of highly homologous SMS-REPs making the locus prone to both meiotic and mitotic rearrangements [166].

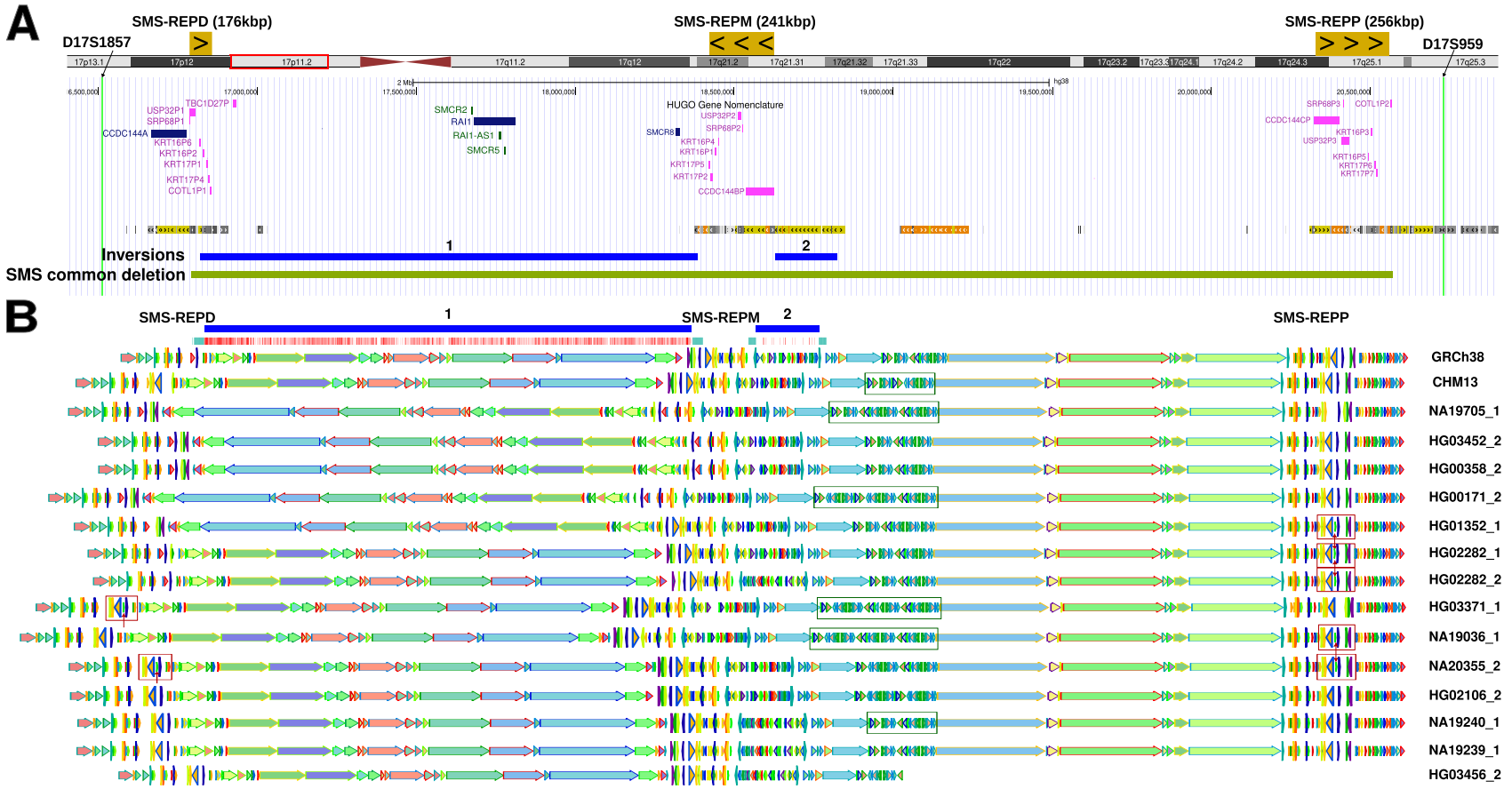


Figure 5.5: Recurrent inversions overlapping Smith-Magenis syndrome critical region. A. The UCSC Genome Browser [173] view showing the SMS common deletion locus that occurs between D17S959 and D17S1857 markers [174]. The three SD-copies (SMS-REPs) associated with this region are shown on the top. For ease of visualization, only genes associated with these SD copies or with the syndrome are shown in the track below the ideogram. The track below the genes shows segmental duplications. **B.** Haplotype structures for the SMS locus generated using pgr-tk [175] are shown as a sequence of colored components with arrows indicating the orientation of each component. Repetitive structures seen across the three SMS-REPS and in the inversions flank have a higher height. Red boxes indicate regions showing transfer of genetic material (pointed to by arrows). Green boxes represent the region exhibiting copy number variation. The positions of tinodes observed for both inversions (shown in A) in GRCh38 reference are shown at the top using red vertical lines, while the green vertical lines represent the anchor nodes for each inversion.

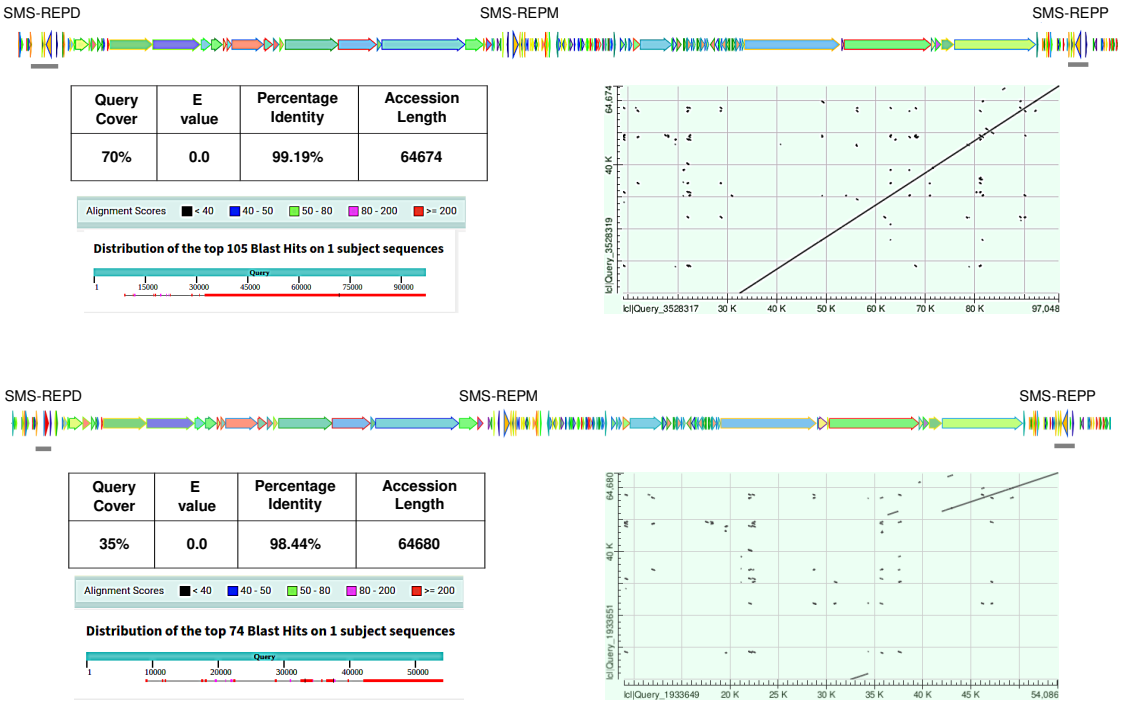


Figure 5.6: Sequence homology between distal and proximal SMS-REPs. pgr-tk [175] based haplotype structures for two haplotypes with varying SMS-REPD structures. The respective sequence homology statistics obtained using blastn [176] are shown below each haplotype.

The 1.56 Mbp inversion identified as recurrent by Pivot lies between the distal and middle SMS-REPs and traverses the RAI1 gene (marked "1" in Figure 5.5). In our cohort of 65 individuals, only four haplotypes (HG03452_2, HG00358_2, HG00171_2, and HG01352_1) showed an inversion at this locus [15]. In the cohort of 44 individuals, analyzed in our earlier *Cell* study [13], only two individuals (HG01352, HG00171) were observed to be carrying the heterozygous inversion. Given this extremely low allele frequency (2.3%) of the inverted allele in the analyzed cohort, both the tiSNPs- and haplotype-based approaches had limited power to detect recurrence. In the current cohort of 65 samples, although the inverted allele frequency is still low (3.07%), Pivot found convincing evidence of recurrence for this inversion with 23% (8685/37676) of the analyzed sequence indicating toggling (tinodes_1). The upper-bound estimate on the minimum number of haplotypes needed to be removed from the cohort in order to get rid of the recurrence signal was observed to be ten, indicating variation within the inverted haplotypes. A breakdown based on node lengths showed that 67.3% (5844/8685 bp) of the recurrence signal was coming from SNPs, 13.4% (1165/8685 bp) from small variants (>1 bp and <50 bp) while 19.3% (1676/8685 bp) was originating from SVs (≥ 50 bp). Even with such few inverted haplotypes, the recurrence signal was still quite strong while restricting only to tinodes_2 (9.2%, 3448/37676 bp). All of the 132 haplotypes included in the graph traversed this locus and its flanking region as one contig, indicating high quality of the assemblies and the graph at this locus.

We additionally analyzed the architecture of all the haplotypes at this locus. The haplotype paths extracted by Pivot were translated back to nucleotide sequences using the node sequence information available in the GFA and the resulting sequences were visualized using the PanGenome Research Tool Kit (pgr-tk)—a tool designed to analyze complex haplotype structures and sequences [175] (Figure 5.5B). We observed tinodes distributed across the whole inverted locus (red vertical lines at the top of GRCh38 haplotype in Figure 5.5B), potentially ruling out wrong signals generated by recombination or gene conversion events. In addition to the four haplotypes (HG03452_2, HG00358_2, HG00171_2, and HG01352_1) genotyped as inverted by ArbiGent, we observed that the sample NA19705, for which ArbiGent predicted a copy number variation "4000" (details about how these genotype codes are defined are discussed in Section 3.2.2), is also heterozygous for this inversion.

This locus showed a diverse SD architecture across haplotypes, particularly for the distal and middle SMS-REPs flanking the inversion. Figure 5.5B shows that the inversion is present in at least two distinct repeat backgrounds as indicated by the SMS-REPM structure of HG01352_1 haplotype showing structural differences compared to other haplotypes with the inversion. Based on this observation, we hypothesize that with the inclusion of more individuals into the analyzed cohort, more distinct SD structures flanking the inversion would be revealed. Additionally, we observed non-inverted haplotypes with highly similar SD architecture as the inverted ones, for example, the SMS-REPD structures observed for HG02282_1 and HG02282_2 (non-inverted) appear more similar to the HG01352_1 (inverted) haplotype, than to other haplotypes not carrying the inversion (Figure 5.5B). Furthermore, we observed positional switch of one structural component between SMS-REPD and SMS-REPP (marked by red arrows inside regions shown in red boxes in Figure 5.5B). HG01352_1, HG02282_1, HG02282_2, and NA19036_1 contain this structure only in the SMS-REPP, HG03371_1 has it only inside the SMS-REPD, while NA20355_2 haplotype shows it in both of them.

Both SMS-REPD and SMS-REPM architectures exhibit considerable variation across the analyzed haplotypes, some of which are shown in Figure 5.5B, while SMS-REPP maintained a relatively similar structure. In order to determine the effect of this variation on the frequency of NAHR and hence the predisposition to the common SMS deletion, we analyzed the sequence homology between each of the two most prevalent SMS-REPD structures and the SMS-REPP structure of the respective haplotype. Figure 5.6 shows that the SMS-REPD structure observed across all of the inverted and some of the non-inverted haplotypes (carrying red triangles instead of yellow ones), exhibits considerably less sequence homology with the SMS-REPP structure of the respective haplotype as compared to the other prevalent SMS-REPD structure (with yellow triangles). As sequence homology is one of the primary factors driving NAHR, we hypothesize that certain SMS-REPD structures, for example the one carried by all the inverted haplotypes shown in Figure 5.5B would have a lower chance of mediating the SMS common deletion, hence playing a protective role. We additionally analyzed the node structure of the inversion flanks (anchor nodes) in the PGGB graph. For each

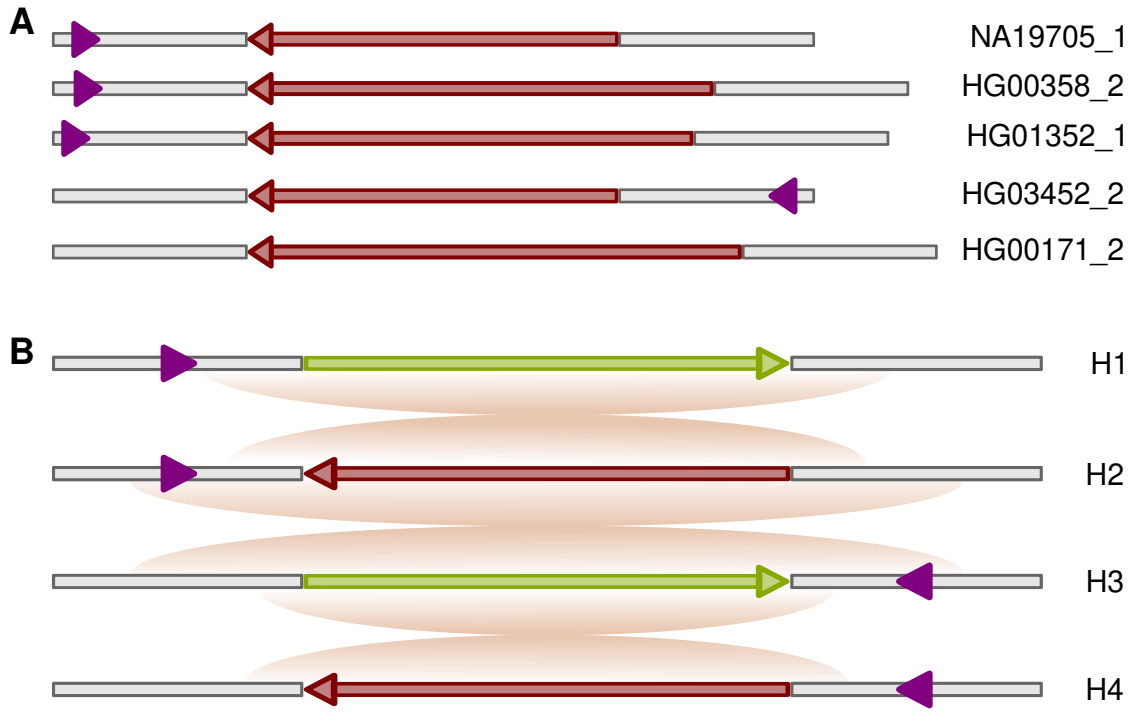


Figure 5.7: Graph nodes switching positions between left and right flank of the longer SMS inversion. A. Visualization of a graph node (purple triangles) observed in either the left or right flank, or absent in both flanks of the inverted region, across all haplotypes. Triangle orientation indicates traversal direction: forward-pointing for forward, and backward-pointing for reverse traversal orientations within the respective haplotype path. Red arrows denote the longer SMS inversion allele (Figure 5.5), while gray bars represent flanking SDs (anchor nodes). Only haplotypes carrying the inversion are shown. Among all 132 haplotypes analyzed, this node appeared in the right flank of the inversion locus exclusively in HG03452_2. **B.** Schematic showing a potential sequence of inversion events that could explain the node's shift between inversion flanks (shown in A). Green arrows represent the reference (non-inverted) allele. Beginning with haplotype H1, three independent inversions involving distinct breakpoints result in haplotype H4 (HG03452_2).

of the haplotypes, we individually searched for unique nodes on each side of the inversion, that is, a node that is present either on the left or right side of the inversion in the respective haplotype. Aggregating the search results over all haplotypes, and filtering for nodes that are unique in some but not all haplotypes, we identified one node that was present at most once across all the haplotypes. Interestingly, we observed that this node was present in the anchor structure on the left side for three of the haplotypes carrying the inversion and on the right side in one of them with a flipped traversal direction (Figure 5.7A). This implies that in one of the haplotypes the inversion breakpoints included the sequence represented by the respective node while in the other three they did not. This pattern serves as a recurrent NAHR footprint indicating that in our analyzed cohort, at least three independent inversion events occurred at this locus with shifted breakpoints (Figure 5.7B).

In addition to this inversion, Pivot also identified a neighboring 201 kbp inversion (marked

"2" in Figure 5.5), as recurrent, with tinodes observed across the whole locus (red vertical lines above the respective locus in Figure 5.5). One breakpoint of this inversion lies in SMS-REPM. This inversion also showed a low allele frequency (3.07%) with NA19240_1, NA19239_1, HG03456_2, and HG02282_2 carrying the inverted allele. Notably, none of these haplotypes carried the longer inversion (marked "1" in Figure 5.5). This inversion is located in an SD-rich background as shown in Figure 5.5B, serving as a perfect substrate for recurrent NAHR. Proximal to this inversion, we observed another highly repetitive region showing duplication in some haplotypes (green boxes in Figure 5.5B). Interestingly, neither in our initial inversion callset, which was based on 44 samples [14], nor in the newly analyzed cohort of 65 samples, was an inversion between SMS-REPM and SMS-REPP observed. Inclusion of a larger number of individuals would help identify if it is just a limitation of the analyzed cohort or if external factors, for example selection, are preventing it.

Prader-Willi/Angelman Syndrome critical region

In our inversion callset [13], we reported four inversions overlapping the 15q11-q13 critical region (Figure 5.8A). A deletion at this locus in the paternal chromosome results in Prader-Willi syndrome (PWS), while a deletion in the maternal copy causes Angelman syndrome (AS) [177]. Both of these are neurodevelopmental genetic disorders, associated with developmental delay and intellectual disability. The Prader-Willi/Angelman syndrome (PWAS) critical region has three breakpoints—BP1, BP2 and BP3—mapped by HERC2 and other duplicons [178, 179] (Figure 5.8A). Recurring deletions involving these breakpoints result in two common deletions associated with PWAS. Type 1 deletion, found in approximately 40% of PWAS patients, involves BP1 and BP3 and is 6 Mbp long, while Type 2 deletion, found in approximately 60% of the cases, involves BP2 and BP3 and is 5.3 Mbp long [179–181].

The four inversions in our callset traversing this locus, shown as Inversion 1–4 in Figure 5.8, are 512 kbp, 22 kbp, 5 Mbp, and 325 kbp long, respectively. In our cohort of 65 individuals, the first and third inversions were each observed in only one haplotype (HG03732_2 and HG02492_2, respectively [15]), making it impossible to observe recurrence evidence for these inversions within the current cohort (Figure 5.8B). The inverted allele frequency observed for the second and fourth inversions was 70% and 16%, respectively [15]. Because of the highly repetitive architecture of this region, with many shared SD copies (hence shared anchor nodes) distributed across this locus, we analyzed the entire locus together using Pivot. Visual analysis of the haplotype structures at this locus alongside Pivot's detected recurrence evidence revealed tinodes dispersed across the whole locus instead of being associated with a specific inversion (positions marked by red circles in Figure 5.8B). Interestingly, most of the observed tinodes belonged to the flanking regions of the inversions. Figure 5.8B shows tinodes present in the region between the first and second inversion. Since the first inversion is present only in one haplotype in our analyzed cohort, we attribute this recurrence evidence to the second inversion. This tinodes-containing intermediate region is flanked by inverted repeats and also appears to invert independently across haplotypes

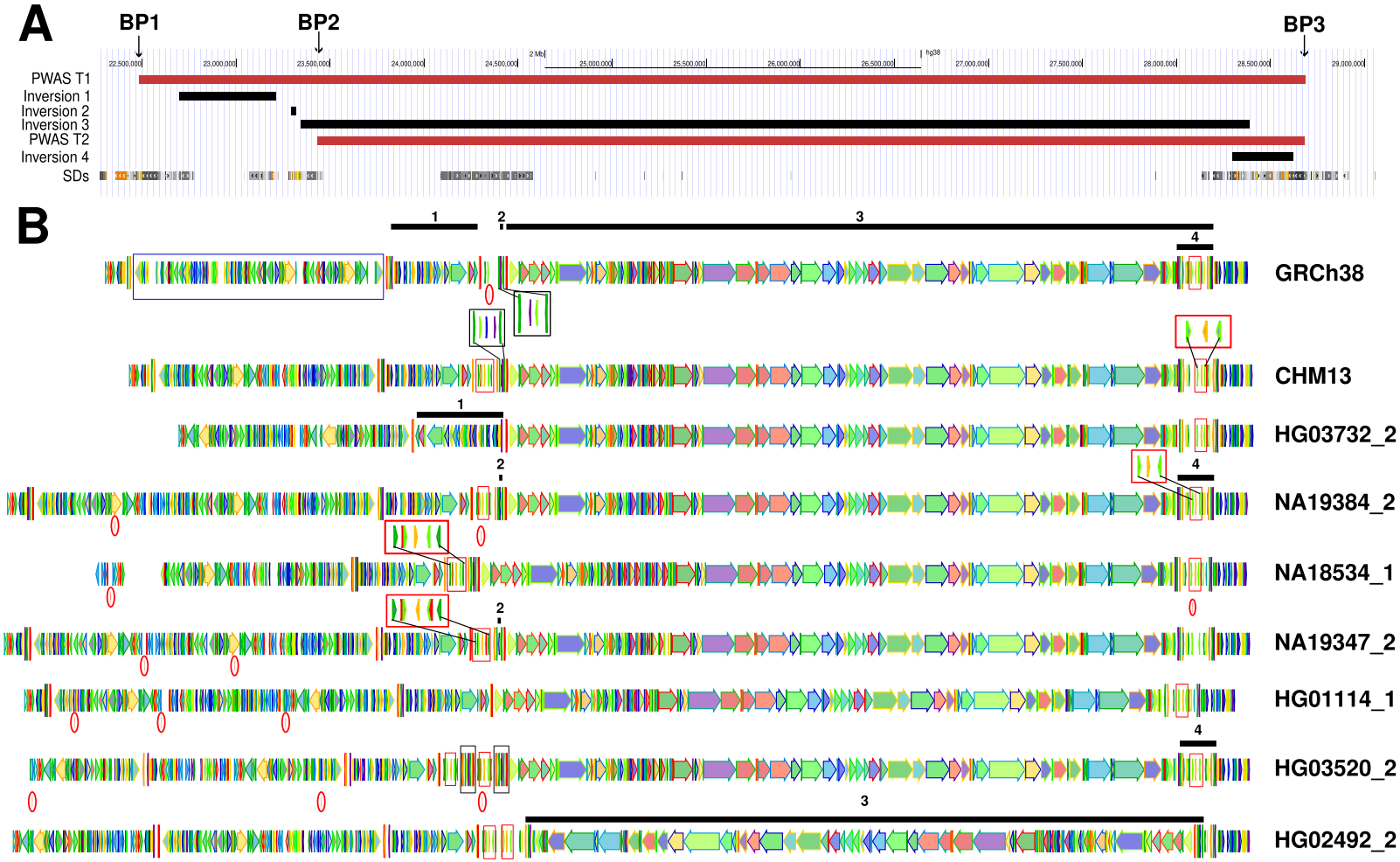


Figure 5.8: Recurrent rearrangements at the Prader-Willi/Angelman syndrome critical region. A. The UCSC Genome Browser [173] view showing the PWAS locus with breakpoints (BP) associated with the syndrome shown at the top. The regions associated with Type 1 and Type 2 deletions (red) and the overlapping inversions (black) are also annotated. **B.** A subset of pgr-tk [175] based haplotype structures at the PWAS locus are shown as a sequence of colored components with arrows indicating the orientation of each component. The four inversions are marked by black bars, with numbers indicating the order of the inversion at the locus (as shown in A). The red lines surrounded by circles represent the location of tinodes. Blue box represents the region undergoing extensive variation across the haplotypes. Red boxes represent regions sharing SD copies and exhibiting inversions and duplications across haplotypes.

(red boxes in Figure 5.8B). Moreover, the architecture of this region is highly identical to the fourth inversion, which also contains tinodes in some haplotypes (Figure 5.8B). Notably, we observed that the region carrying the second inversion flips along with the third 5 Mbp inversion (HG02492_2, Figure 5.8B). This ability of the second inversion locus to invert individually and along with the neighboring inversion event, indicates its recurrent nature. We hypothesize that analysis of more haplotypes carrying the third inversion might reveal cases where the second locus does not invert alongside the third one, indicating another recurrent formation. It is important to note that GRCh38 and CHM13 sequences have opposite orientation for the second inversion locus. As the orientation of this region in haplotype HG02492_2 indicates that GRCh38 is potentially misoriented in this region, we considered CHM13 state as the "reference/non-inverted" state in this case.

In addition to inversion, this region also showed copy number variations, for example, HG01114_1 showed a loss while HG03520_2 showed a copy number gain at the second inversion locus (Figure 5.8). Furthermore, Pivot picked up some signals of recurrence from the region distal to the first inversion, marked by blue rectangle in Figure 5.8B. This region exhibited highly varying haplotype patterns involving inversions and copy number changes, some of which are shown in Figure 5.8B. Overall, we observed dispersed recurrence evidence and extensive structural variation at this locus and hypothesize that with the inclusion of more haplotypes, particularly those carrying the first and third inversions, might reveal more recurrent rearrangement patterns at this locus.

Additional candidates

In addition to the inversions discussed above, other newly identified recurrent candidates overlapping morbid CNV regions include:

- Two inversions overlapping 1q21.1 recurrent microdeletion/microduplication region associated with TAR syndrome and neurodevelopmental disorders [182].
- Two inversions overlapping the 2q13 deletion syndrome critical region associated mainly with developmental delay and cardiac and urogenital malformations [183].
- An inversion overlapping 10q23 deletion region associated with juvenile polyposis and developmental delays [184, 185].
- An inversion overlapping the 3q29 microdeletion/microduplication syndrome critical region associated with mental retardation and microcephaly [182].
- Five chromosome 16 inversions overlapping the 16p11.2–p12.2 microdeletion/microduplication syndrome region associated with neurodevelopmental disorders [182]
- An inversion overlapping the 17q12 microdeletion/microduplication critical region associated with kidney issues, diabetes, and neuro-developmental disorders [186].

5.4 Discussion

This chapter introduces Pivot—an approach to detect inversion recurrence in a pangenome graph. Building upon the theoretical framework of the tiSNPs-based approach, presented in Chapter 4, Pivot extends the analysis to all variants within the inverted locus and its flanking regions, thereby increasing the power to detect recurrence. Additionally, instead of being reliant on the sequencing reads aligned to a linear reference genome, Pivot utilizes the genome assemblies laid out in a pangenome graphical format. On the one hand, this helps improve the quality of the analysis due to the better sequence resolution of genome assemblies and the depth of genomic information captured by a pangenome, while on the other hand, it helps eliminate reference bias. This property also makes Pivot readily usable for assessing the recurrence status of inversions called at the individual haplotype level.

We first applied Pivot to the HPRC year 1 PGGB graphs, which consist of 90 haplotypes [35], to assess the recurrence status of the 292 balanced inversions reported in our inversion callset from the *Cell* study [13]. We observed Pivot's results to be consistent with previous findings for the 17q21.31 single-event and 8p23.1 highly recurrent inversions. For the 17q21.31 inversion, we observed at most 0.58% of the analyzed sequence being indicative of toggling, while this percentage was observed to be at least 14.5% for the 8p23.1 inversion. Results shown in Table 5.1 indicate that even with a higher number of analyzed variants as compared to the tiSNPs-based approach, in stricter settings, Pivot is not picking up false recurrence signals. Consistent with its high recurrence rate of 1.11×10^{-4} [13], a large number of haplotypes (74%) were observed to be contributing to the recurrence evidence for the 8p23.1 inversion(Table 5.2). Based on the overall results observed across all inversions, we observed that while confirming the recurrence status of the already known recurrent inversions, Pivot was able to detect recurrence evidence for inversions where previously no or weak recurrence signal was observed [13], indicating its higher sensitivity. This increase in sensitivity likely stems from the inclusion of additional analyzed variants and the improved sequence resolution of the genome assemblies.

Although Pivot showed an overall improvement in inversion recurrence detection, we observed that 56% of the inversions, primarily consisting of those flanked by short inverted repeats, could not be assessed because of the absence of a flanking anchor structure in the graph. As explained in Section 5.2.4, certain graph construction parameter settings can result in over-simplification of the graphical structure by collapsing repetitive regions. In regions of smaller repeats, these collapses can lead to a graphical layout where the repetitive structure is no longer traceable, resulting in the absence of anchor nodes at these loci. We hypothesize that a locus-specific tweaking of the graph construction parameters can help in generating a more accurate layout for these cases. Particularly, we observed that none of the inversions belonging to X and Y chromosomes could be assessed either due to the absence of anchor structure (as described above) or the absence of safe nodes that help to synchronize the extracted sub-paths across haplotypes. The latter can potentially result

from low-quality fragmented assemblies containing contigs not large enough to traverse the inverted loci alongside the nearest safe nodes. Overall, these observations indicate that the graphical structure plays a vital role in the performance of any method solely reliant on the pangenome graphs.

Afterwards, we analyzed the same set of inversions using graphs generated from higher-quality HGSVC3 assemblies for 65 human individuals [15]. As compared to HPRC graphs, the number of inversions that Pivot could assess in the HGSVC3 graphs increased by 18%. This improvement can be attributed in part to the higher quality of the input assemblies, and in part to the graph construction parameters, which were tweaked to reduce collapses and produce a more granular graphical structure. Concurrently, the number of novel recurrence candidates with strong evidence of recurrence also increased as compared to HPRC graphs (42% vs 34%). Comparison of plots shown in Figure 5.3 and Figure 5.4 clearly depicts that not only more inversions could be assessed in the HGSVC3 graphs, the recurrence evidence, quantified as the percentage of tisequence, also became stronger. This suggests that better quality pangenome graphs can improve inversion recurrence assessment both qualitatively and quantitatively. However, chromosome X still remained an exception, with no inversions assessed for recurrence. Based on this consistent behavior even with better quality assemblies, we hypothesize that it is a limitation of the graph construction process and the parameters working well for autosomes might not necessarily work well for sex chromosomes. Theoretically, the graphical layout of a pangenome should mirror the varying genetic patterns observed across different chromosomes and within the same chromosomes. However, in practice, most of the pangenome graphs are constructed based on a "one-size-fits-all" approach—using the same parameter settings to construct the whole genome. Particularly for inversions, this means that a graph generated with parameters tuned for low complexity may fail to lay out the flanking inverted repeat structures in sufficient detail—resulting in the absence of the loop structure (represented by anchor nodes) required by Pivot to identify an inversion. Additionally, much of the variation in and around the inversion—crucial for gathering recurrence evidence—may be collapsed. On the other hand, a more granular graph that resolves these repeat structures can introduce considerable complexity within and outside the inverted region, thereby increasing noise in the input. Accounting for locus- and SV-specific characteristics during graph construction, or using region-specific graphs, may help mitigate these issues to some extent.

Of the newly identified recurrent candidates in the HGSVC3 graphs, 18 overlap with disease associated critical regions. One notable case is the 1.56 Mbp long inversion overlapping the 17p11.2 critical region associated with Smith-Magenis/Potocki-Lupski syndrome [159]. Even with only five haplotypes carrying the inversion, we found 23% of the analyzed sequence to be toggling-indicating, with tinodes distributed across the whole locus. We additionally observed that the SD architecture of the inversion flanks varies within the inverted haplotypes, indicating independent inversion occurrence in distinct repeat contexts (Figure 5.5B). Concurrently, we observed a graph node shifting position from left to right

flank in one of the haplotypes carrying the inversion, suggesting independent occurrences of this inversion with slightly shifted breakpoints (Figure 5.7B). A comparison of sequence homology between two most prevalent structures observed for SMS-REPD and the respective SMS-REPP copies revealed that one has significantly less sequence identity than the other (Figure 5.7A). Since these two SD copies mediate the common SMS deletion, the haplotypes with less homology between the two would have a lower likelihood of NAHR-mediated deletion as compared to their counterparts. Interestingly, all the inverted haplotypes carry the SMS-REPD sequence exhibiting lower sequence identity. We expect that analysis of this locus in a larger cohort of individuals, including SMS patients, can help establish the potential link between these varying haplotype structures and the disease.

Additionally, we found inversion recurrence evidence distributed across the SD-rich regions at the 15q11-q13 locus associated with Prader-Willi/Angelman syndrome [177]. Visually inspecting the structure of regions containing tinodes revealed an SD-rich architecture mediating different inversion and copy number events (Figure 5.8B). For example, we observed a 22 kbp inversion inverting in place and alongside a 5 Mbp neighboring inversion, with tinodes observed in the flanking SD-rich region (Figure 5.8B). Additionally, we observed tinodes at the distal part of this locus exhibiting extensive copy number variation with nested small inversions. Based on the presence of many SD copies across the locus, along with the wide distribution of tinodes, we hypothesize that this region is predisposed to multiple NAHR-mediated recurrent events including inversions, deletions and duplications. A detailed analysis of such a structurally complex locus needs a much bigger cohort of individuals than we currently analyzed. Furthermore, we found recurrent inversions overlapping 1q21.1, 2q13, 10q23, 3q29, 16p11.2-p12.2, and 17q12 regions.

In summary, we observed that Pivot exhibits a higher sensitivity as compared to our previous approaches, identifying novel recurrent candidates and increasing the confidence for already known cases. Given the improvement seen by using higher quality assemblies for graph generation, we plan to use Pivot to detect inversion recurrence in the newly released HPRC graphs generated with even higher quality assemblies and a cohort of more than 200 samples. With a bigger cohort, we hope to be able to better understand the recurrent rearrangements at structurally complex disease-associated regions, for example the PWAS or the 3q29 region. Additionally, since Pivot is reference-independent, we also plan to extend it to the analysis of recurrence for inversions called on any individual (non-reference) haplotypes.

Chapter 6

Analyzing homology-mediated recurrent deletion polymorphisms

This chapter describes my work on identifying recurrent deletions across a cohort of 1,019 samples from the 1000 Genomes Project. This work was conducted as part of a collaborative study, provisionally accepted for publication in Nature [38]. Relevant text and figures from the publication are included in this chapter, with proper references where necessary. For publication details and author contributions, please refer to Section C.6.

6.1 Motivation

The SV breakpoint analysis described in this section was performed by Carsten Hain.

Chapters 4 and 5 primarily focused on the study of recurrent inversion polymorphisms across the human genome. Although inversion recurrence plays an important role in disease-associated copy number variations due to flipping SD orientations, the phenomenon of recurrence is not confined solely to inversions. The presence of repetitive sequences across the genome can initiate homology-mediated recombination [55], giving rise to different types of SVs, depending on the orientation of the involved repeats and the rearrangement pattern. As discussed in the preceding chapters, homology-mediated rearrangements have been identified as a key mechanism driving inversion recurrence; they are similarly hypothesized to contribute to the recurrence of other repeat-mediated SVs [21, 24, 56, 57]..

This chapter presents research conducted in this context, as part of a study aimed at SV characterization using ONT sequencing in a cohort of 1,019 samples, representing 26 populations from the 1KGP [38]. An analysis of the breakpoints for the SVs (66,198 deletions and 75,238 insertions) showed that 35% of deletions and 28.7% of insertions had over 50 bp of sequence homology at their flanks, indicating a potential role for homology-directed repair (HDR) in their formation. Among these, 10.8% deletions and 6.7% insertions had flanking homology of at least 200 bp, which is indicative of NAHR being the poten-

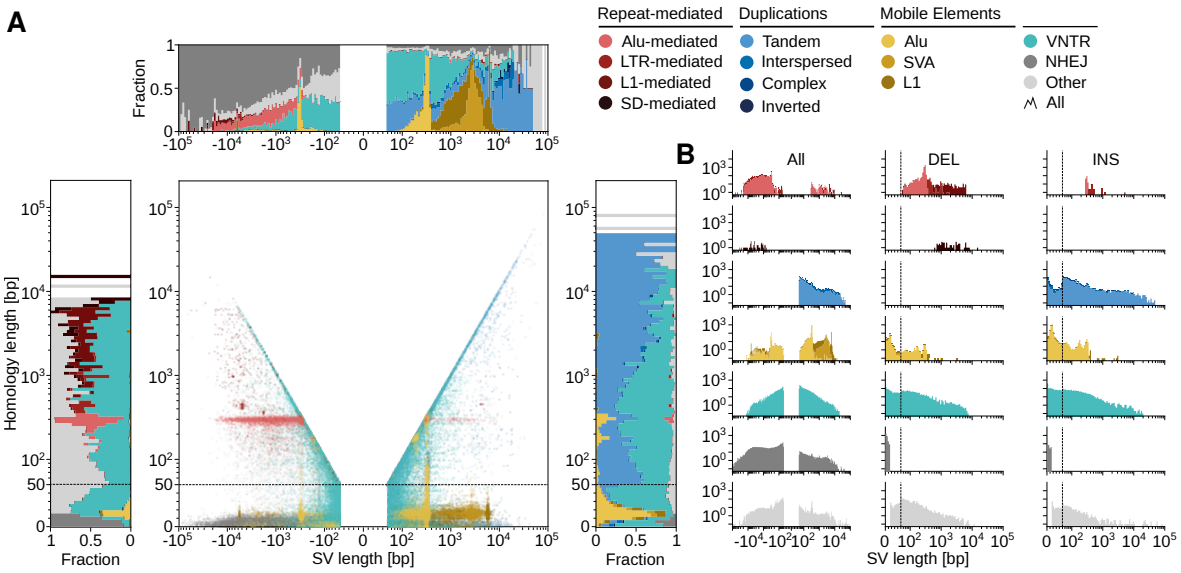


Figure 6.1: SV breakpoint homology. A. SV length (horizontal axis) versus flank homology length (vertical axis) for all SVs classes (colored based on the categories shown on top right). SV length <0 indicates deletions while SV length >0 indicates insertions. Marginal plots show the size-binned fraction of SV classes, with deletions and insertions, individually shown at the left and right, respectively. NHEJ: SVs exhibiting $\leq$15 bp microhomology or blunt ended breakpoints. Other: Unclassified SVs. **B.** SV length (horizontal axis) versus homology length (vertical axis) for all, only deletions, and only insertions. Figure adapted from [38].

tial mediating mechanism. Notably, many of these homology-flanked SVs were flanked by annotated transposable elements (TEs)—specifically Alu, L1, and LTR elements—in the T2T-CHM13 reference genome (Figure 6.1). This enrichment was particularly evident in deletions (3,260 instances) compared to insertions (80 instances). Among this subset of TE flanked SVs, 89.3% were observed to be flanked by full-length Alu elements, showing a sharp breakpoint homology length peak at 295 bp in Figure 6.1A. AluY and AluSx were observed to be the predominant Alu-subclasses appearing in different configurations, correlating with their counts in the reference genome. These Alu-mediated SVs varied widely in size, ranging from approximately 300 bp up to 20.4 kbp for deletions and up to 9.5 kbp for insertions. These findings are consistent with Alu-based transposable element-mediated rearrangement (TEMR), previously known to be a prevalent SV formation process [187]. Following Alu elements, L1 elements were the next most common mobile elements flanking large SVs, again more frequent in deletions (219 cases) than insertions (1 case). SVs flanked by L1 elements were typically larger, reaching up to 62.9 kbp in length (Figure 6.1B).

Overall, these observations suggest that homology-mediated processes, utilizing a diverse spectrum of repeat lengths, play a significant role in the formation of SVs, particularly deletions. Additionally, TEs can potentially create new substrates for deletions due to their ability to jump around in the genome. With these aspects in mind, we conducted a more detailed analysis of the deletions flanked by TEs to identify potential instances of repeat-

mediated recurrence. To this end, we developed an approach analogous to the "tiSNPs-based approach" for detecting inversion recurrence, described in Chapter 4 and applied it to search for "recurrence-indicating SNPs" in deletion flanking regions. Since the breakpoints of recurrent events cluster in the mediating LCRs, the flanking region can potentially contain patterns indicative of recurrence (Section 1.4.2, Figure 1.4). Additionally, we performed phylogenetic validation using hierarchical clustering based trees to identify the recurrence status for each of the analyzed locus.

6.2 Deletion recurrence analysis: Methodological framework

6.2.1 Identifying potentially recurrent deletions

Before testing for recurrence, we applied a filtering strategy to focus the analysis on cases with the potential for recurrence. The filtering process was structured as follows:

- Given the relatively shorter length of TEs compared to other repeat classes, such as SDs, they are not expected to mediate large events. Therefore, we limited our analysis to $\leq$5 kbp long deletions.
- Concurrently, deletions with flanking sequence homology outside the range of 200–9,000 bp were excluded. The lower bound of 200 bp was selected based on prior findings indicating the minimal processing length of flanking repeat sequences necessary for NAHR [188, 189]. The upper bound of 9,000 bp was chosen to include events mediated by small to intermediate sized SDs.
- In order to limit to cases where we have enough power to detect recurrence, deletions with allele frequency outside the range of 40-60% were excluded from the analysis.
- Additionally, we used Hardy-Weinberg equilibrium and Mendelian consistency statistics to exclude events with potential genotyping errors.
- Finally, we filtered out deletions lacking phasing information.

6.2.2 Recurrence detection

To identify potential recurrence of all deletions shortlisted after applying the filtering criteria mentioned in Section 6.2.1, we used a modified version of the tiSNPs-based approach, designed for detecting inversion recurrence [13] (described in Chapter 4). Consistent with the behavior expected in case of inversions (Chapters 4, 5), observing both alleles of a bi-allelic SNP in haplotypes with and without the deletion provides evidence that the deletion arose independently in different haplotype backgrounds over the course of evolution. Because the loss of sequence makes it impossible to search for such "recurrence-indicating SNPs" within the deletion locus itself, we focused on the 20 kbp flanking region on each

side of the deletion locus and screened bi-allelic SNPs with at least 10% allele frequency to identify signals of recurrence.

Additionally, we applied centroid hierarchical clustering to analyze the clustering patterns observed between haplotypes with and without the deletion. Specifically, we aimed to determine whether haplotypes carrying the deletion allele group together in clusters with haplotypes carrying the reference allele, which would indicate that the deletion event recurred across human populations. For this analysis, we again relied on the flanking sequences, constructing a phylogenetic tree using SNPs occurring within a 100 kbp window on each side of the deletion while excluding the sequence within the annotated deletion breakpoints. The difference in analyzed flanking region lengths—20 kbp for SNP screening versus 100 kbp for tree construction—was intended to minimize potential noise. Mechanistically, if the deletion recurred multiple times, the recurrence-indicating SNPs should be located in close proximity of the event. Signals from SNPs farther away cannot be confidently attributed to the locus of interest. Conversely, a broader region is necessary for phylogenetic tree construction to prevent clustering patterns from being driven by unrelated variation or recombination near the deleted locus. Furthermore, we visually inspected the SNP haplotypes alongside the phylogenetic tree to filter out noise. This involved ensuring that the recurrence-indicating SNPs were present on both sides of the deletion and ruling out cases where recombination appeared to be driving the clustering pattern.

Because accurate interpretation of the clustering patterns depends on the correct genotype assignment for each haplotype, we also validated genotype accuracy in cases where both SNPs and the phylogenetic tree provided evidence of recurrence. To do this, the read alignments from individuals carrying the deletion and the reference allele who appeared together in similar clusters in the phylogenetic tree, were visually examined using IGV [190].

6.3 Results

Using the filtering criteria outlined in Section 6.2.1, we shortlisted 42 potentially recurrent deletion candidates. For each of these candidates, we first searched for evidence of recurrence by looking for recurrence-indicating SNPs in deletion flanks, using phased SNP genotypes generated for the full cohort [38]. Afterwards, we analyzed whether haplotypes with and without the deletion appeared within the same phylogenetic groups, based on SNP haplotypes from the flanking regions. Finally, after verifying genotyping accuracy as an additional quality check, we identified six deletion candidates with the most compelling evidence of recurrence. All six sites displayed recurrence-indicating SNPs on both sides of the deletion and showed clustering patterns consistent with the expectations for a recurrent locus. Notably, all six cases belong to the class of potential TEMRs [187], likely driven by HDR or NAHR using the flanking retrotransposons sequences. Among these, one site involved L2/LINE sequences, while the remaining five were flanked by Alu elements. These findings suggest that TEMRs potentially play a vital role in generating recurrent SV poly-

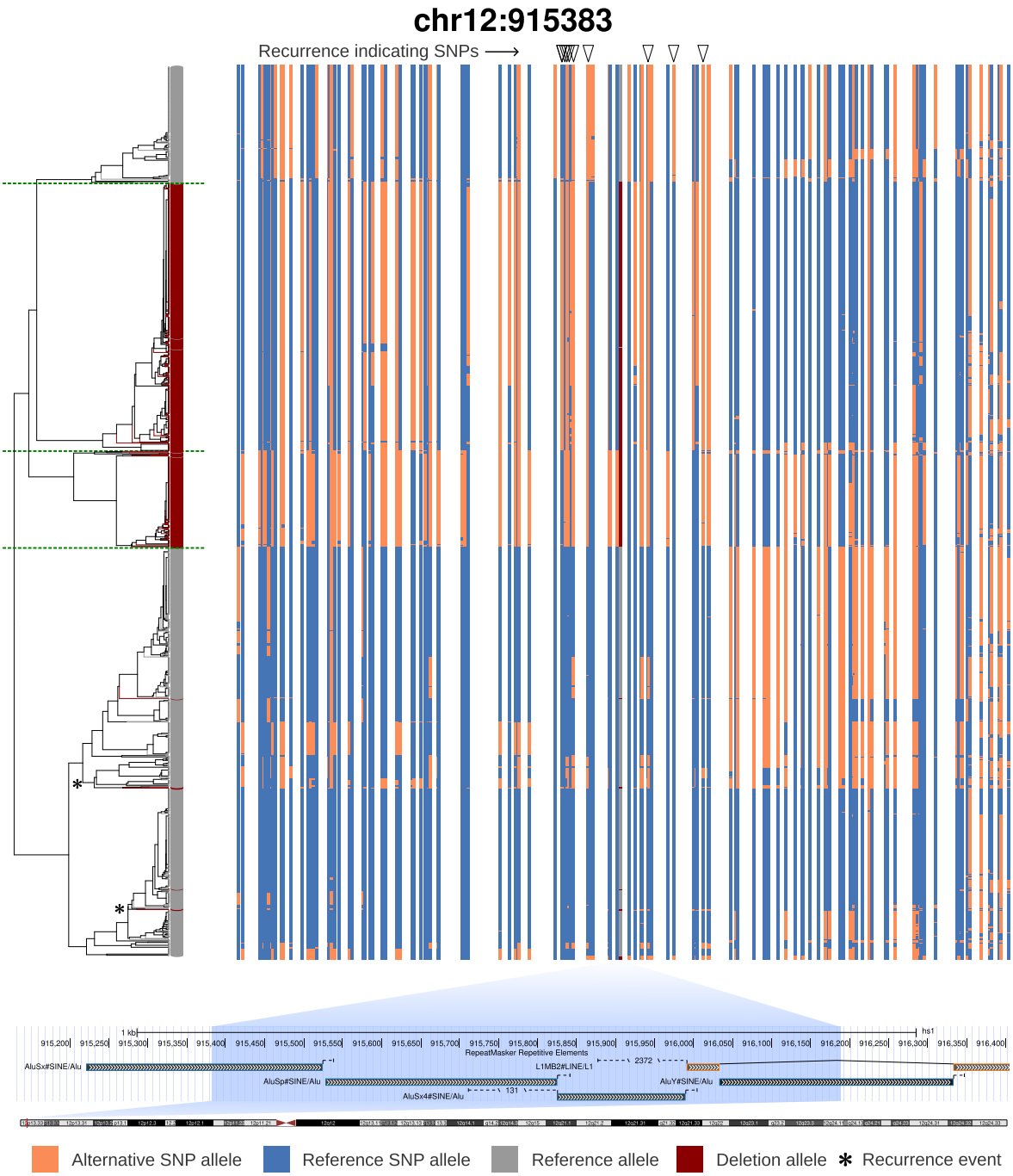


Figure 6.2: Alu-mediated recurrent deletion at 12p13.3. A predicted 806 bp recurrent deletion at 12p13.3, mediated by an AluSx-AluY pair. The figure shows the variation of haplotypes in a 100 kb window centered around the deletion and the relationship between haplotypes with (red) and without the deletion (gray). Dendrograms of haplotypes are plotted using centroid hierarchical clustering algorithm. Green dashed lines represent the separation of four haplotype groups. In each SNP haplotype, reference and alternative alleles are shown in blue and orange, respectively. Recurrence-indicating SNPs within 20 kbp window around the deletion are marked by triangles at the top. Two predicted independent occurrences of the deletion event are denoted by *. The UCSC Genome Browser [173] view, with the region of interest, highlighted in blue, is shown at the bottom of the figure.

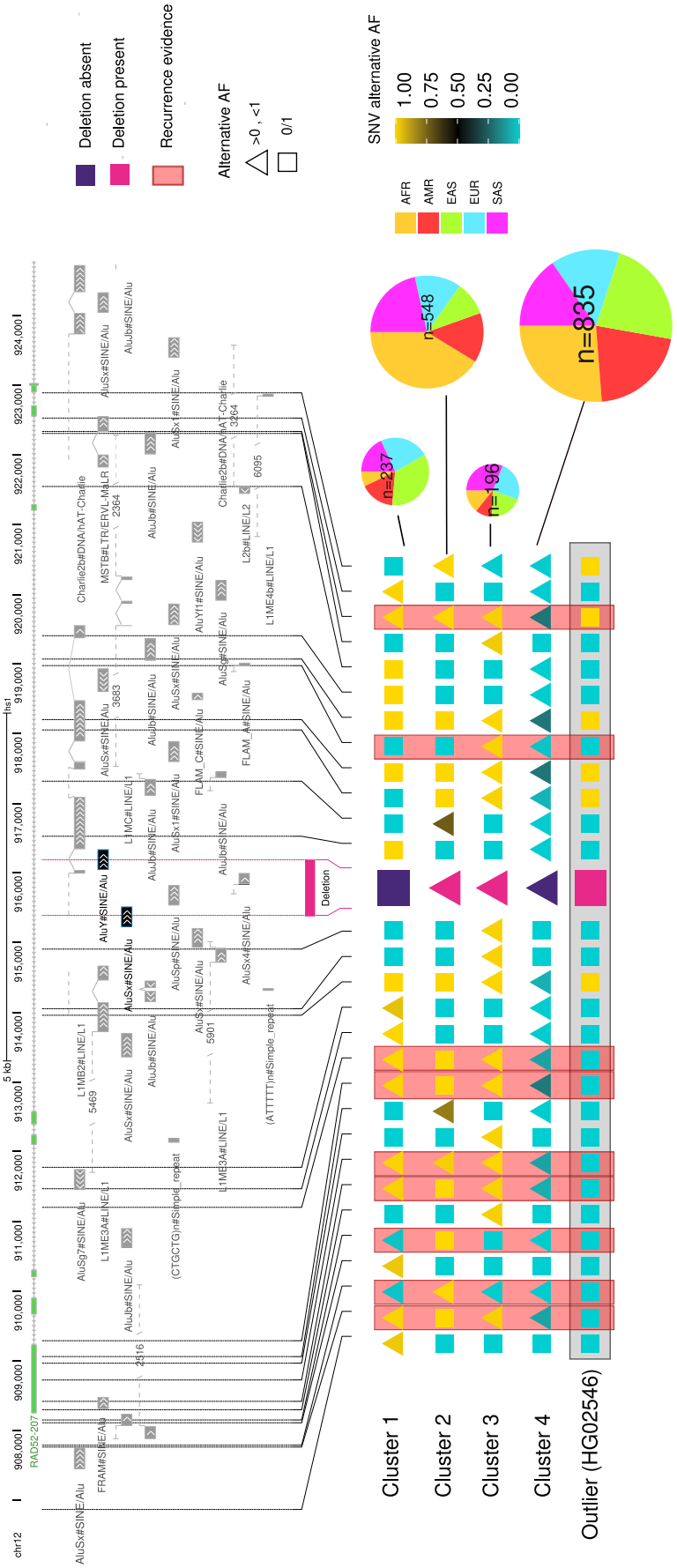


Figure 6.3: Haplotype consensus and geographical ancestries at the 12p13.3 locus. A 806 bp deletion located at 12p13.3, mediated by an AluSx-AluY pair, with evidence for recurrent formation. Clusters 1-4 are obtained from SNP-based clustering of the haplotypes in a 100 kbp window centered around the deletion. Pie charts represent geographical ancestries for each of the four clusters. For ease of visualization, a consensus haplotype in an 18 kbp window centered around the deletion is represented for each cluster. Squares represent an allele frequency of 0 (yellow) or 1 (cyan) within a cluster, while triangles represent allele frequencies in between. A haplotype from sample "HG02546", grouping with Cluster 4, is represented as an outlier. Recurrence-indicating SNPs lying within the 18 kbp window are marked by red vertical bars.

morphisms in human populations. Given the high abundance of mobile elements in the genome, such events are expected to occur frequently.

Figure 6.2 shows an 806 bp recurrent deletion located on chromosome 12p13.3, driven by an AluSx-AluY pair. Within a 20 kbp window centered around the deletion, we identified ten recurrence-indicating SNPs (shown as black triangles). Hierarchical clustering of the SNP haplotypes across a 100 kbp window around the deletion revealed strong evidence of recurrence, with haplotypes carrying the deletion allele clustering alongside those with the reference allele. The resulting dendrogram displayed four distinct haplotype groups, three of which showed a clear mixing of deleted and non-deleted haplotypes (Figure 6.2, green dashed lines). Through read alignment visualization, we confirmed the genotypes for five haplotypes carrying the deletion, but appearing in the fourth cluster, which predominantly consists of haplotypes without the deletion (marked with *). To investigate further, we constructed "pseudo-haplotypes"—consensus sequences where each position reflects the most frequent allele observed—for each of the four clusters (Figure 6.3). Additionally, superpopulation stratification for each cluster was performed, which recapitulated the geographical ancestries: Cluster 1 primarily included East Asian (EAS) samples, Clusters 2 and 4 were dominated by African (AFR) samples, and South Asian (SAS) samples were most common in Cluster 3 (cluster numbers based on their order in Figure 6.2). Notably, the deletion allele was present in different haplotype backgrounds in Clusters 2 and 3 and in an outlier haplotype from Cluster 4, where the reference allele is otherwise the consensus at this locus. The presence of recurrence-indicating SNPs on both sides of the deletion suggests either multiple independent deletion events or extensive local recombination near the deletion site. Given the proximity of these SNPs and their association with specific haplotype groups, the most likely explanation is recurrent TEMR formation, with alternative scenarios involving recombination on either side of the event being less plausible.

The other five potential recurrent candidates include:

- deletion at 19p13.13 mediated by an AluSx-AluSx pair (Figure B.1)
- deletion at 11p15.4 mediated by an AluY-AluSx pair (Figure B.2)
- 9p24.3 deletion locus mediated by AluY-AluY pair (Figure B.3)
- 9p13.3 deletion mediated by an AluSx-AluSx pair (Figure B.4)
- a deletion at 9q22.1 locus mediated by an L2-L2 pair (Figure B.5)

Each of these events depicted strong evidence of recurrence, with numerous recurrence-indicating SNPs found on both sides of the deletion. Phylogenetic trees, supported by manual genotype validation, revealed distinct clustering patterns consistent with multiple independent origins of these deletions.

6.4 Discussion

As part of our study aimed towards SV characterization in a cohort of 1,019 samples, using ONT reads [38], we found that mobile elements account for an average of 2.26 Mbp of sequence variation detected per individual, driven by several mechanisms including mobile element insertions (MEIs), TEMRs, transductions and processed pseudogenes. Analyzing the SV breakpoints and their surrounding regions revealed flanking homology lengths ranging from 50 to thousands of base pairs (Figure 6.1). This suggests that a diverse array of homology-associated processes, mediated by varying homology lengths and different genomic repeat contexts, shape the SV landscape in humans.

When examining SVs involving flanking mobile element sequences, we observed that the majority of TEMRs (89.9%) represent deletions. Among the different classes of flanking mobile elements, Alu elements were the most prevalent, consistent with previous findings that Alu-based TEMRs are a dominant SV formation process [187]. Given that these homology-mediated rearrangements can occur independently in distinct haplotype backgrounds, we hypothesized that they play a crucial role in SV recurrence. While precise recurrence rate estimates for these events require much larger sample cohorts, we attempted to evaluate the recurrence status for some of the deletion events.

We analyzed a set of 42 deletions that met the criteria described in Section 6.2.1 to determine whether they arose more than once over the course of evolution. Applying a recurrence detection approach built on the theoretical framework used for inversion recurrence analysis, discussed in detail in Chapters 4 and 5, we searched for recurrence-indicating SNPs in the flanking region of each deletion. Beyond identifying such SNPs, we sought additional phylogenetic support by constructing centroid hierarchical clustering-based dendrograms using SNP haplotypes from the 100 kbp flanking regions of each event. By integrating these analyses with manual inspection of aligned sequencing reads for genotype validation, we identified six deletions with compelling evidence of recurrence. In each case, phylogenetic trees consistently revealed co-clustering of haplotypes with and without the deletion. Moreover, the presence of recurrence-indicating SNPs in close proximity to both breakpoints reinforced the recurrent nature of these deletions. Among these six events, one was flanked by LINE elements, while the remaining five were flanked by Alu elements, highlighting the potential role of TEMRs in facilitating SV recurrence. These findings lay the foundation for future studies using larger sample cohorts to explore this phenomenon across various SV types. Future directions include examining the linkage disequilibrium patterns around these TEMRs and estimating the precise recurrence rates for such events.

Overall, the results presented in this chapter, alongside those from Chapters 4 and 5, establish mutational SV recurrence as a widespread phenomenon across different SV classes, regulated by diverse mechanisms. This underscores the importance of using large, geographically diverse cohorts and developing sophisticated methodologies to gain a population-wide perspective on these events, determine their mutational recurrence rates and understand

their potential disease associations.

Summary and Conclusion

Recent advancements in sequencing technologies have significantly improved our ability to analyze genomic variation. While short-read sequencing provides high accuracy and precision, long-read sequencing technologies, although capable of capturing larger structural variations, are often susceptible to higher rates of systematic errors. These errors limit the full potential of long reads, making it difficult to fully leverage them for identifying and characterizing genomic variation. In Chapter 2 of this thesis, a novel allele detection approach, *k*-merald, was introduced. *k*-merald uses platform-specific *k*-mer-based sequencing error profiles to calculate read alignment costs, thereby enhancing alignment reliability and, in turn, enhancing allele detection accuracy, particularly for error-prone long reads. We demonstrated that *k*-merald outperforms conventional edit-distance-based allele detection, resulting in a prominent reduction in genotyping error rates for SNPs and indels across both ONT and PacBio CLR datasets. This improvement was especially evident in low-coverage sequencing data. Additionally, we showed that *k*-merald does not rely on sample-specific error profiles. This allows a single error profile to be used across multiple samples—provided the sequencing data originated from the same platform—thereby reducing computational runtime.

Although all classes of genomic variation contribute to our understanding of genetic diversity, structural variants are particularly important due to the larger genomic regions they typically affect. Among these, inversions represent an especially intriguing subclass; unlike deletions or duplications, they are usually copy-neutral and preserve genomic content. However, their frequent occurrence in highly repetitive regions renders them difficult to detect. Chapter 3 of this thesis introduced ArbiGent, a tool for genotyping inversions and copy number variants using strand-state inheritance patterns obtained from Strand-seq data. ArbiGent corrects for alignment artifacts by normalizing Strand-seq read counts based on the mappability of each locus. Validation experiments, including Mendelian consistency and Hardy-Weinberg equilibrium analyses, confirmed the high quality of the produced genotypes. Furthermore, a downsampling experiment demonstrated that ArbiGent maintains reliable genotyping performance even at low Strand-seq coverage. This chapter also detailed ArbiGent's role in generating a multi-platform based inversion callset, comprising 399 inversions derived from 44 individuals from the 1KG Project. Analysis of this callset revealed that inversions span approximately twice the genomic length affected by indels and four times that affected by SNPs. Notably, a significantly higher number of inversions were found to

occur at $>5\%$ minor allele frequency compared to other SV types. Moreover, we observed SD-rich regions on chromosomes 1, 2, 7, 10, 15, 16, and 17 to be the "hotspots" for inversion formation.

The enrichment of inversions in highly-repetitive regions makes them prone to NAHR-mediated recurrent formation. However, distinguishing a single ancestral inversion event from multiple independent events at the same locus requires haplotype-resolved data along with robust phylogenetic and population-based analysis. Chapter 4 presented the novel concept of toggling-indicating SNPs (tiSNPs) and outlined a strategy for assessing inversion recurrence using these markers. By analyzing allelic patterns of within-inversion SNPs across both inverted and non-inverted haplotypes, this approach allowed us to classify inversions as recurrent or single-event. In combination with orthogonal haplotype-based and Y-chromosome phylogenetic analyses, we identified 40 recurrent inversions, collectively spanning approximately 0.6% of the human genome. Recurrence rates reached up to one in 10,000 haplotype transmissions—four orders of magnitude higher than those observed for SNPs—highlighting the widespread nature of inversion recurrence in human genomes. Among these 40 recurrent inversions, we identified one overlapping the 7q11.23 Williams-Beuren syndrome critical region, previously associated with disease predisposition. Additionally, we demonstrated that inversions overlapping 7q11.23, 3q29 and 15q13.3 critical regions altered the surrounding SD architecture, suggesting a potential mechanistic role in disease predisposition. However, exact characterization of these associations requires larger case-control cohorts and high-quality, haplotype-resolved genome assemblies.

With the emergence of pangenomics, this goal is becoming increasingly attainable. Chapter 5 of this thesis presented Pivot, a pangenome-friendly inversion recurrence detection method that builds upon the theoretical framework of the tiSNPs-based approach, while extending it by incorporating all within-inversion variant types—not just SNPs—and operates directly on pangenome graphs, thereby mitigating reference bias. We applied Pivot to analyze inversion recurrence using HPRC and HGSVC pangenome graphs, which represent 88 and 130 (non-reference) haplotypes, respectively. Pivot not only identified novel recurrent inversions but also provided stronger evidence for known recurrent candidates. As expected, it performed better on HGSVC graphs, constructed from higher quality genome assemblies as compared to the HPRC graphs, emphasizing the importance of assembly quality and larger sample sizes for recurrence detection. Pivot revealed 18 novel recurrent inversions overlapping disease-associated regions, including a 1.56 Mbp inversion at the 17p11.2 locus encompassing the Smith-Magenis/Potocki-Lupski syndrome region. We demonstrated that the flanking SD architectures varied between inversion-carrying haplotypes, indicating their independent origins. Additionally, evidence for inversion recurrence was observed across the Prader-Willi/Angelman syndrome critical region (15q11–q13), where four distinct inversions share similar SD-flanking structures. Future work will involve applying Pivot to even better-quality pangenome graphs, representing more samples—such as those recently released by the HPRC—to refine recurrence detection, explore complex loci in greater detail,

and assess the contribution of recurrent inversions to disease risk.

Inversions are not the only SVs subject to recurrence. Chapter 6 explored deletion recurrence, proposing an approach to scan deletion flanks for recurrence-indicating SNPs. Using this approach, in combination with phylogenetic validation, we identified six recurrent deletions within a cohort of 1,019 samples from the 1KGP. Five were flanked by Alu elements and one by LINE elements, suggesting a role for transposable element-mediated NAHR in their recurrence. Moving forward, we plan to investigate linkage disequilibrium patterns around such events, estimate precise recurrence rates in larger cohorts, and extend these analyses to other SV classes.

The work presented in this thesis underscored both the importance and the intricate complexity of human genomic variation, offering just a glimpse into a much larger and still unfolding narrative. While significant progress has been made in this field, a great deal remains to be explored. With the rapid evolution of high-throughput sequencing technologies, the development of advanced computational methods, and the recent surge in pangenomics, we are now better equipped than ever to explore the full spectrum of human genomic variation. These innovations are paving the way toward advancements in personalized medicine, uncovering the genetic basis of disease phenotypes, and deepening our understanding of human evolution.

Bibliography

[1] The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001. doi: 10.1038/35057062. URL https://doi.org/10.1038/35057062.

[2] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[3] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[4] International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851, 2007.

[5] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[6] Nathan D Olson, Justin Wagner, Jennifer McDaniel, Sarah H Stephens, Samuel T Westreich, Anish G Prasanna, Elaine Johanson, Emily Boja, Ezekiel J Maier, Omar Serang, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell genomics*, 2(5), 2022.

[7] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, 17(6):333–351, 2016.

[8] Glennis A Logsdon, Mitchell R Vollger, and Evan E Eichler. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10):597–614, 2020.

[9] Wenjuan Yu, Haohui Luo, Jinbao Yang, Shengchen Zhang, Heling Jiang, Xianjia Zhao, Xingqi Hui, Da Sun, Liang Li, Xiu-qing Wei, et al. Comprehensive assessment of 11 de novo HiFi assemblers on complex eukaryotic genomes and metagenomes. *Genome Research*, 34(2):326–340, 2024.

[10] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[11] Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J Gardner, Oscar L Rodriguez, Li Guo, Ryan L Collins, et al. Multi-platform

discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1):1784, 2019.

[12] Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537):eabf7117, 2021.

[13] David Porubsky, Wolfram Höps, Hufsah Ashraf, PingHsun Hsieh, Bernardo Rodriguez-Martin, Feyza Yilmaz, Jana Ebler, Pille Hallast, Flavia Angela Maria Maggiolini, William T Harvey, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, 185(11):1986–2005, 2022.

[14] David Porubsky, William T Harvey, Allison N Rozanski, Jana Ebler, Wolfram Höps, Hufsah Ashraf, Patrick Hasenfeld, Human Pangenome Reference Consortium (HPRC), Human Genome Structural Variation Consortium (HGSVC), Benedict Paten, et al. Inversion polymorphism in a complete human genome assembly. *Genome Biology*, 24(1):100, 2023.

[15] Glennis A Logsdon, Peter Ebert, Peter A Audano, Mark Loftus, David Porubsky, Jana Ebler, Feyza Yilmaz, Pille Hallast, Timofey Prodanov, DongAhn Yoo, et al. Complex genetic variation in nearly complete human genomes. *bioRxiv*, 2024.

[16] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

[17] Joachim Weischenfeldt, Orsolya Symmons, François Spitz, and Jan O Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*, 14(2):125–138, 2013.

[18] Evan E Eichler. Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine*, 381(1):64–74, 2019.

[19] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews genetics*, 11(6):446–450, 2010.

[20] Matthew E Hurles, Emmanouil T Dermitzakis, and Chris Tyler-Smith. The functional impact of structural variation in humans. *Trends in Genetics*, 24(5):238–245, 2008.

[21] James R Lupski and Pawel Stankiewicz. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS genetics*, 1(6):e49, 2005.

[22] Max Levitan and Ashley Montagu. Textbook of human genetics. *Genetics*, 450:65, 1971.

[23] Pawel Stankiewicz and James R Lupski. Genome architecture, rearrangements and genomic disorders. *TRENDS in Genetics*, 18(2):74–82, 2002.

[24] Claudia MB Carvalho and James R Lupski. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17(4):224–238, 2016.

[25] Lisenka ELM Vissers and Paweł Stankiewicz. Microdeletion and microduplication syndromes. *Genomic Structural Variants: Methods and Protocols*, pages 29–75, 2012.

[26] Francesca Antonacci, Megan Y Dennis, John Huddleston, Peter H Sudmant, Karyn Meltz Steinberg, Jill A Rosenfeld, Mattia Miroballo, Tina A Graves, Laura Vives, Maika Malig, et al. Palindromic golga8 core duplicons promote chromosome 15q13. 3 microdeletion and evolutionary instability. *Nature genetics*, 46(12):1293–1302, 2014.

[27] Claudia Rita Catacchio, Flavia Angela Maria Maggiolini, Pietro D'addabbo, Miriana Bitonto, Oronzo Capozzi, Martina Lepore Signorile, Mattia Miroballo, Nicoletta Archidiacono, Evan E Eichler, Mario Ventura, et al. Inversion variants in human and primate genomes. *Genome Research*, 28(6):910–920, 2018.

[28] Jay N Lozier, Amalia Dutra, Evgenia Pak, Nan Zhou, Zhili Zheng, Timothy C Nichols, Dwight A Bellinger, Marjorie Read, and Richard A Morgan. The chapel hill hemophilia a dog colony exhibits a factor viii gene inversion. *Proceedings of the National Academy of Sciences*, 99(20): 12991–12996, 2002.

[29] Flavia AM Maggiolini, Stuart Cantsilieris, Pietro D'Addabbo, Michele Manganelli, Bradley P Coe, Beth L Dumont, Ashley D Sanders, Andy Wing Chun Pang, Mitchell R Vollger, Orazio Palumbo, et al. Genomic inversions and golga core duplicons underlie disease instability at the 15q25 locus. *PLoS Genetics*, 15(3):e1008075, 2019.

[30] Flavia Angela Maria Maggiolini, Ashley D Sanders, Colin James Shew, Arvis Sulovari, Yafei Mao, Marta Puig, Claudia Rita Catacchio, Maria Dellino, Donato Palmisano, Ludovica Mercuri, et al. Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome research*, 30(11):1680–1693, 2020.

[31] David Porubsky, Ashley D Sanders, Wolfram Höps, PingHsun Hsieh, Arvis Sulovari, Ruiyang Li, Ludovica Mercuri, Melanie Sorensen, Shwetha C Murali, David Gordon, et al. Recurrent inversion toggling and great ape genome evolution. *Nature genetics*, 52(8):849–858, 2020.

[32] Michael C Zody, Zhaoshi Jiang, Hon-Chung Fung, Francesca Antonacci, LaDeana W Hillier, Maria Francesca Cardone, Tina A Graves, Jeffrey M Kidd, Ze Cheng, Amr Abouelleil, et al. Evolutionary toggling of the mapt 17q21. 31 inversion region. *Nature genetics*, 40(9):1076–1083, 2008.

[33] Lucy R Osborne, Martin Li, Barbara Pober, David Chitayat, Joann Bodurtha, Ariane Mandel, Teresa Costa, Theresa Grebe, Sarah Cox, Lap-Chee Tsui, et al. A 1.5 million–base pair inversion polymorphism in families with williams-beuren syndrome. *Nature genetics*, 29(3):321–325, 2001.

[34] David A Koolen, Lisenka ELM Vissers, Rolph Pfundt, Nicole De Leeuw, Samantha JL Knight, Regina Regan, R Frank Kooy, Edwin Reyniers, Corrado Romano, Marco Fichera, et al. A new chromosome 17q21. 31 microdeletion syndrome associated with a common inversion polymorphism. *Nature genetics*, 38(9):999–1001, 2006.

[35] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K Lucas, Jean Monlong, Haley J Abel, et al. A draft human pangenome reference. *Nature*, 617(7960):312–324, 2023.

[36] Hufsah Ashraf, Jana Ebler, and Tobias Marschall. Allele detection using k-mer-based sequencing error profiles. *Bioinformatics Advances*, 3(1):vbad149, 2023.

[37] Ashley D Sanders, Sascha Meiers, Maryam Ghareghani, David Porubsky, Hyobin Jeong, M Alexandra CC van Vliet, Tobias Rausch, Paulina Richter-Pechańska, Joachim B Kunz, Silvia Jenni, et al. Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nature biotechnology*, 38(3):343–354, 2020.

[38] Siegfried Schloissnig, Samarendra Pani, Bernardo Rodriguez-Martin, Jana Ebler, Carsten Hain, Vasiliki Tsapalou, Arda Söylev, Patrick Hüther, Hufsah Ashraf, Timofey Prodanov, et al. Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 genomes project. *bioRxiv*, pages 2024–04, 2024.

[39] Brett Trost, Livia O Loureiro, and Stephen W Scherer. Discovery of genomic variation across a generation. *Human Molecular Genetics*, 30(R2):R174–R186, 2021.

[40] Anthony Rhoads and Kin Fai Au. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

[41] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3): 443–453, 1970.

[42] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[43] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, 36(9):875–879, October 2018.

[44] Murray Patterson, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W Klau, and Alexander Schönhuth. WhatsHap: Weighted haplotype assembly for Future-Generation sequencing reads. *J. Comput. Biol.*, 22(6):498–509, June 2015.

[45] Marcel Martin, Murray Patterson, Shilpa Garg, Sarah Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schöenhuth, and Tobias Marschall. WhatsHap: fast and accurate read-based phasing. *BioRxiv*, page 085050, 2016.

[46] David Porubsky, Shilpa Garg, Ashley D Sanders, Jan O Korbel, Victor Guryev, Peter M Lansdorp, and Tobias Marschall. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature communications*, 8(1):1293, 2017.

[47] George H Perry, Fengtang Yang, Tomas Marques-Bonet, Carly Murphy, Tomas Fitzgerald, Arthur S Lee, Courtney Hyland, Anne C Stone, Matthew E Hurles, Chris Tyler-Smith, et al. Copy number variation and evolution in humans and chimpanzees. *Genome research*, 18(11): 1698–1710, 2008.

[48] Maren Wellenreuther, Claire Mérot, Emma Berdan, and Louis Bernatchez. Going beyond snps: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular ecology*, 28(6), 2019.

[49] Linyi Zhang, Radka Reifová, Zuzana Halenková, and Zachariah Gompert. How important are structural variants for speciation? *Genes*, 12(7):1084, 2021.

[50] Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, et al. A structural variation reference for medical and population genetics. *Nature*, 581(7809):444–451, 2020.

[51] Marco Raffaele Cosenza, Bernardo Rodriguez-Martin, and Jan O Korbel. Structural variation in cancer: role, prevalence, and mechanisms. *Annual review of genomics and human genetics*, 23(1):123–152, 2022.

[52] Zhichao Liu, Ruth Roberts, Timothy R Mercer, Joshua Xu, Fritz J Sedlazeck, and Weida Tong. Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome biology*, 23(1):68, 2022.

[53] Andrew J Sharp, Devin P Locke, Sean D McGrath, Ze Cheng, Jeffrey A Bailey, Rhea U Vallente, Lisa M Pertz, Royden A Clark, Stuart Schwartz, Rick Segraves, et al. Segmental duplications and copy-number variation in the human genome. *The American Journal of Human Genetics*, 77(1):78–88, 2005.

[54] Jeffrey A Bailey, Zhiping Gu, Royden A Clark, Knut Reinert, Rhea V Samonte, Stuart Schwartz, Mark D Adams, Eugene W Myers, Peter W Li, and Evan E Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, 2002.

[55] Mitchell R Vollger, Xavi Guitart, Philip C Dishuck, Ludovica Mercuri, William T Harvey, Ariel Gershman, Mark Diekhans, Arvis Sulovari, Katherine M Munson, Alexandra P Lewis, et al. Segmental duplications and their variation in a complete human genome. *Science*, 376(6588): eabj6965, 2022.

[56] Lisanne Vervoort and Joris R Vermeesch. Low copy repeats in the genome: From neglected to respected. *Exploration of Medicine*, 4(2):166–175, 2023.

[57] Wenli Gu, Feng Zhang, and James R Lupski. Mechanisms for human genomic rearrangements. *Pathogenetics*, 1:1–17, 2008.

[58] Ondrej Pös, Jan Radvanszky, Gergely Buglyó, Zuzana Pös, Diana Rusnakova, Bálint Nagy, and Tomas Szemes. Dna copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomedical Journal*, 44(5):548, 2021.

[59] Claudia MB Carvalho, Feng Zhang, and James R Lupski. Structural variation of the human genome: mechanisms, assays, and role in male infertility. *Systems biology in reproductive medicine*, 57(1-2):3–16, 2011.

[60] Philip J Hastings, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8):551–564, 2009.

[61] Òscar Molina, Ester Anton, Francesca Vidal, and Joan Blanco. High rates of de novo 15q11q13 inversions in human spermatozoa. *Molecular Cytogenetics*, 5:1–9, 2012.

[62] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[63] James R Lupski. Genomic disorders: structural features of the genome can lead to dna rearrangements and human disease traits. *Trends in genetics*, 14(10):417–422, 1998.

[64] Florencia Pratto, Kevin Brick, Pavel Khil, Fatima Smagulova, Galina V Petukhova, and R Daniel Camerini-Otero. Recombination initiation maps of individual human genomes. *Science*, 346(6211):1256442, 2014.

[65] Kenneth Paigen and Petko M Petkov. Prdm9 and its role in genetic recombination. *Trends in Genetics*, 34(4):291–300, 2018.

[66] Pengfei Liu, Melanie Lacaria, Feng Zhang, Marjorie Withers, PJ Hastings, and James R Lupski. Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *The American Journal of Human Genetics*, 89(4):580–588, 2011.

[67] Alfred Henry Sturtevant. A case of rearrangement of genes in drosophila. *Proceedings of the National Academy of Sciences*, 7(8):235–237, 1921.

[68] Costas B Krimbas and Jeffrey R Powell. *Drosophila inversion polymorphism*. CRC press, 1992.

[69] Mark Kirkpatrick. How and why chromosome inversions evolve. *PLoS biology*, 8(9):e1000501, 2010.

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

[71] Maren Wellenreuther and Louis Bernatchez. Eco-evolutionary genomics of chromosomal inversions. *Trends in ecology & evolution*, 33(6):427–440, 2018.

[72] Dorcas J Orengo, Eva Puerma, Unai Cereijo, and Montserrat Aguadé. The molecular genealogy of sequential overlapping inversions implies both homologous chromosomes of a heterokaryotype in an inversion origin. *Scientific reports*, 9(1):17009, 2019.

[73] Diarmaid Hughes. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome biology*, 1:1–8, 2000.

[74] Deborah L Nagle, Christine A Kozak, Hiroyuki Mano, Verne M Chapman, and Maja Bućan. Physical mapping of the Tec and Gabrb1 loci reveals that the W sh mutation on mouse chromosome 5 is associated with an inversion. *Human molecular genetics*, 4(11):2073–2079, 1995.

[75] Kathleen M Davis, Steven A Smith, and Ira F Greenbaum. Evolutionary implications of chromosomal polymorphisms in Peromyscus boylii from southwestern Mexico. *Evolution*, pages 645–649, 1986.

[76] Hreinn Stefansson, Agnar Helgason, Gudmar Thorleifsson, Valgerdur Steinthorsdottir, Gisli Masson, John Barnard, Adam Baker, Aslaug Jonasdottir, Andres Ingason, Vala G Gudnadottir, et al. A common inversion under selection in europeans. *Nature genetics*, 37(2):129–137, 2005.

[77] Maximilian PA Salm, Stuart D Horswell, Claire E Hutchison, Helen E Speedy, Xia Yang, Liming Liang, Eric E Schadt, William O Cookson, Anthony S Wierzbicki, Rossi P Naoumova, et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome research*, 22(6):1144–1153, 2012.

[78] Zachary L Fuller, Christopher J Leonard, Randee E Young, Stephen W Schaeffer, and Nitin Phadnis. Ancestral polymorphisms explain the role of chromosomal inversions in speciation. *PLoS genetics*, 14(7):e1007526, 2018.

[79] Ryan L Collins, Harrison Brand, Claire E Redin, Carrie Hanscom, Caroline Antolik, Matthew R Stone, Joseph T Glessner, Tamara Mason, Giulia Pregno, Naghmeh Dorrani, et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome biology*, 18:1–21, 2017.

[80] Gregory M Cooper, Bradley P Coe, Santhosh Girirajan, Jill A Rosenfeld, Tiffany H Vu, Carl Baker, Charles Williams, Heather Stalker, Rizwan Hamid, Vickie Hannig, et al. A copy number variation morbidity map of developmental delay. *Nature genetics*, 43(9):838–846, 2011.

[81] Delia Lakich, Haig H Kazazian Jr, Stylianos E Antonarakis, and Jane Gitschier. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature genetics*, 5(3):236–241, 1993.

[82] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[83] Marta Puig, Sonia Casillas, Sergi Villatoro, and Mario Cáceres. Human inversions and their functional consequences. *Briefings in functional genomics*, 14(5):369–379, 2015.

[84] Cristina Aguado, Magdalena Gaya-Vidal, Sergi Villatoro, Meritxell Oliva, David Izquierdo, Carla Giner-Delgado, Víctor Montalvo, Judit Garcia-Gonzalez, Alexander Martinez-Fundichely, Laia Capilla, et al. Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS Genetics*, 10(3):e1004208, 2014.

[85] Carla Giner-Delgado, Sergi Villatoro, Jon Lerga-Jaso, Magdalena Gayà-Vidal, Meritxell Oliva, David Castellano, Lorena Pantano, Bárbara D Bitarello, David Izquierdo, Isaac Noguera, et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature communications*, 10(1):4222, 2019.

[86] Piotr Dittwald, Tomasz Gambin, Claudia Gonzaga-Jauregui, Claudia MB Carvalho, James R Lupski, Paweł Stankiewicz, and Anna Gambin. Inverted low-copy repeats and genome instability—a genome-wide analysis. *Human mutation*, 34(1):210–220, 2013.

[87] Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. Structural variant calling: the long and the short of it. *Genome biology*, 20:1–14, 2019.

[88] Jana Ebler, Peter Ebert, Wayne E Clarke, Tobias Rausch, Peter A Audano, Torsten Houwaart, Yafei Mao, Jan O Korbel, Evan E Eichler, Michael C Zody, Alexander T Dilthey, and Tobias Marschall. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.*, 54(4):518–525, 2022.

[89] Yu Chen, Amy Y Wang, Courtney A Barkley, Yixin Zhang, Xinyang Zhao, Min Gao, Mick D Edmonds, and Zechen Chong. Deciphering the exact breakpoints of structural variations using long sequencing reads with debreak. *Nature Communications*, 14(1):283, 2023.

[90] Ashley D Sanders, Sascha Meiers, Maryam Ghareghani, David Porubsky, Hyobin Jeong, M Alexandra CC van Vliet, Tobias Rausch, Paulina Richter-Pechańska, Joachim B Kunz, Silvia Jenni, et al. Single cell tri-channel-processing reveals structural variation landscapes and complex rearrangement processes. *bioRxiv*, page 849604, 2019.

[91] Ester Falconer, Mark Hills, Ulrike Naumann, Steven SS Poon, Elizabeth A Chavez, Ashley D Sanders, Yongjun Zhao, Martin Hirst, and Peter M Lansdorp. Dna template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature methods*, 9(11):1107–1112, 2012.

[92] Ashley D Sanders, Ester Falconer, Mark Hills, Diana CJ Spierings, and Peter M Lansdorp. Single-cell template strand sequencing by strand-seq enables the characterization of individual homologs. *Nature Protocols*, 12(6):1151–1176, 2017.

[93] Christina Gros, Ashley D Sanders, Jan O Korbel, Tobias Marschall, and Peter Ebert. ASHLEYS: automated quality control for single-cell Strand-seq data. *Bioinformatics*, 37(19):3356–3357, 2021.

[94] Maryam Ghareghani, David Porubskỳ, Ashley D Sanders, Sascha Meiers, Evan E Eichler, Jan O Korbel, and Tobias Marschall. Strand-seq enables reliable separation of long reads by chromosome via expectation maximization. *Bioinformatics*, 34(13):i115–i123, 2018.

[95] Definitions — humanpangenome.org. `https://humanpangenome.org/definitions/`.

[96] Jordan M Eizenga, Adam M Novak, Jonas A Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, Xian Chang, Josiah D Seaman, Robin Rounthwaite, Jana Ebler, et al. Pangenome graphs. *Annual review of genomics and human genetics*, 21(1):139–162, 2020.

[97] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102 (39):13950–13955, 2005.

[98] Sean P Gordon, Bruno Contreras-Moreira, Daniel P Woods, David L Des Marais, Diane Burgess, Shengqiang Shu, Christoph Stritt, Anne C Roulin, Wendy Schackwitz, Ludmila Tyler, et al. Extensive gene content variation in the brachypodium distachyon pan-genome correlates with population structure. *Nature communications*, 8(1):2184, 2017.

[99] Sean P Gordon, Bruno Contreras-Moreira, Joshua J Levy, Armin Djamei, Angelika Czedik-Eysenberg, Virginia S Tartaglio, Adam Session, Joel Martin, Amy Cartwright, Andrew Katz, et al. Gradual polyploid genome evolution revealed by pan-genomic analysis of brachypodium hybridum and its diploid progenitors. *Nature communications*, 11(1):3670, 2020.

[100] Lei Gao, Itay Gonda, Honghe Sun, Qiyue Ma, Kan Bao, Denise M Tieman, Elizabeth A Burzynski-Chang, Tara L Fish, Kaitlin A Stromberg, Gavin L Sacks, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics*, 51(6): 1044–1051, 2019.

[101] Murukarthick Jayakodi, Sudharsan Padmarasu, Georg Haberer, Venkata Suresh Bonthala, Heidrun Gundlach, Cécile Monat, Thomas Lux, Nadia Kamal, Daniel Lang, Axel Himmelbach, et al. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, 588 (7837):284–289, 2020.

[102] Charley GP McCarthy and David A Fitzpatrick. Pan-genome analyses of model fungal species. *Microbial genomics*, 5(2):e000243, 2019.

[103] Yang Zhou, Lv Yang, Xiaotao Han, Jiazheng Han, Yan Hu, Fan Li, Han Xia, Lingwei Peng, Clarissa Boschiero, Benjamin D Rosen, et al. Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. *Genome research*, 32(8):1585–1601, 2022.

[104] Ran Li, Mian Gong, Xinmiao Zhang, Fei Wang, Zhenyu Liu, Lei Zhang, Qimeng Yang, Yuan Xu, Mengsi Xu, Huanhuan Zhang, et al. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Research*, 33(3):463–477, 2023.

[105] Peipei Bian, Jiaxin Li, Shishuo Zhou, Xingquan Wang, Mian Gong, Xi Guo, Yudong Cai, Qimeng Yang, Jiaqi Fu, Rongrong Li, et al. A graph-based goat pangenome reveals structural variations involved in domestication and adaptation. *Molecular Biology and Evolution*, 41(12): msae251, 2024.

[106] Dong Li, Yulong Wang, Tiantian Yuan, Minghao Cao, Yulin He, Lin Zhang, Xiang Li, Yifan Jiang, Ke Li, Jingchun Sun, et al. Pangenome and genome variation analyses of pigs unveil genomic facets for their adaptation and agronomic characteristics. *iMeta*, 3(6):e257, 2024.

[107] Yao-zhong Zhang, Arda Akdemir, Georg Tremmel, Seiya Imoto, Satoru Miyano, Tetsuo Shibuya, and Rui Yamaguchi. Nanopore basecalling from a perspective of instance segmentation. *BMC bioinformatics*, 21(3):1–9, 2020.

[108] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10):1155–1162, 2019.

[109] Jana Ebler, Marina Haukness, Trevor Pesout, Tobias Marschall, and Benedict Paten. Haplotype-aware diplotyping from noisy long reads. *Genome Biol.*, 20(1):116, June 2019.

[110] Matei David, Lewis Jonathan Dursi, Delia Yao, Paul C Boutros, and Jared T Simpson. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, 33(1):49–55, 2017.

[111] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. *PloS one*, 12(6):e0178751, 2017.

[112] Ryan R Wick, Louise M Judd, and Kathryn E Holt. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome biology*, 20(1):1–10, 2019.

[113] Gagandeep Singh, Mohammed Alser, Kristof Denolf, Can Firtina, Alireza Khodamoradi, Meryem Banu Cavlak, Henk Corporaal, and Onur Mutlu. Rubicon: a framework for designing efficient deep learning-based genomic basecallers. *Genome Biology*, 25(1):49, 2024.

[114] Marc Pagès-Gallego and Jeroen de Ridder. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biology*, 24(1):71, 2023.

[115] Yuansheng Liu, Yichen Li, Enlian Chen, Jialu Xu, Wenhai Zhang, Xiangxiang Zeng, and Xiao Luo. Repeat and haplotype aware error correction in nanopore sequencing reads with dechat. *Communications Biology*, 7(1):1678, 2024.

[116] Zhimeng Xu, Yuting Mai, Denghui Liu, Wenjun He, Xinyuan Lin, Chi Xu, Lei Zhang, Xin Meng, Joseph Mafofo, Walid Abbas Zaher, et al. Fast-bonito: A faster deep learning based basecaller for nanopore sequencing. *Artificial Intelligence in the Life Sciences*, 1:100011, 2021.

[117] Bert Bogaerts, An Van den Bossche, Bavo Verhaegen, Laurence Delbrassinne, Wesley Mattheus, Stéphanie Nouws, Maxime Godfroid, Stefan Hoffman, Nancy HC Roosens, Sigrid CJ De Keersmaecker, et al. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *Journal of Clinical Microbiology*, 62(5):e01576–23, 2024.

[118] Ekaterina Polkhovskaya, Evgeniy Moskalev, Pavel Merkulov, Ksenia Dudnikova, Maxim Dudnikov, Ivan Gruzdev, Yakov Demurin, Alexander Soloviev, and Ilya Kirov. Cost-Effective Detection of SNPs and Structural Variations in Full-Length Genes of Wheat and Sunflower Using Multiplex PCR and Rapid Nanopore Kit. *Biology*, 14(2):138, 2025.

[119] Mantas Sereika, Rasmus Hansen Kirkegaard, Søren Michael Karst, Thomas Yssing Michaelsen, Emil Aarre Sørensen, Rasmus Dam Wollenberg, and Mads Albertsen. Oxford Nanopore R10. 4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature methods*, 19(7): 823–826, 2022.

[120] Q system data analysis — nanoporetech.com. `https://nanoporetech.com/document/q-system-data-analysis`.

[121] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.

[122] Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.*, 19(1):118–135, January 2018.

[123] Jordan M Eizenga, Adam M Novak, Jonas A Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, Xian Chang, Josiah D Seaman, Robin Rounthwaite, Jana Ebler, Mikko Rautiainen, Shilpa Garg, Benedict Paten, Tobias Marschall, Jouni Sirén, and Erik Garrison. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.*, May 2020.

[124] Jouni Sirén, Jean Monlong, Xian Chang, Adam M Novak, Jordan M Eizenga, Charles Markello, Jonas A Sibbesen, Glenn Hickey, Pi-Chuan Chang, Andrew Carroll, Namrata Gupta, Stacey Gabriel, Thomas W Blackwell, Aakrosh Ratan, Kent D Taylor, Stephen S Rich, Jerome I Rotter, David Haussler, Erik Garrison, and Benedict Paten. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574):abg8871, December 2021.

[125] Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14 Suppl 5:S1, April 2013.

[126] David Porubsky, Shilpa Garg, Ashley D Sanders, Jan O Korbel, Victor Guryev, Peter M Lansdorp, and Tobias Marschall. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.*, 8(1):1293, November 2017.

[127] Zhenxian Zheng, Shumin Li, Junhao Su, Amy Wing-Sze Leung, Tak-Wah Lam, and Ruibang Luo. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science*, 2(12):797–803, 2022.

[128] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, et al. A universal snp and small-indel variant caller using deep neural networks. *Nature biotechnology*, 36(10):983–987, 2018.

[129] Kishwar Shafin, Trevor Pesout, Pi-Chuan Chang, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Gunjan Baid, Mikhail Kolmogorov, Jordan M Eizenga, Karen H Miga, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature methods*, 18(11):1322–1332, 2021.

[130] Justin Wagner, Nathan D Olson, Lindsay Harris, Ziad Khan, Jesse Farek, Medhat Mahmoud, Ana Stankovic, Vladimir Kovacevic, Aaron M Wenger, William J Rowell, et al. Benchmarking challenging small variants with linked and long reads. *BioRxiv*, 2020.

[131] John G Cleary, Ross Braithwaite, Kurt Gaastra, Brian S Hilbush, Stuart Inglis, Sean A Irvine, Alan Jackson, Richard Littin, Mehul Rathod, David Ware, et al. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv*, page 023754, 2015.

[132] Gustavo Glusman, Hannah C Cox, and Jared C Roach. Whole-genome haplotyping approaches and genomic medicine. *Genome Med.*, 6(9):73, September 2014.

[133] Ashley D Sanders, Mark Hills, David Porubskỳ, Victor Guryev, Ester Falconer, and Peter M Lansdorp. Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome research*, 26(11):1575–1587, 2016.

[134] David Porubsky, Ashley D Sanders, Aaron Taudt, Maria Colomé-Tatché, Peter M Lansdorp, and Victor Guryev. breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics*, 36(4):1260–1261, 2020.

[135] Peter A Audano, Carolyn Paisie, Human Genome Structural Variation Consortium, and Christine R Beck. Large complex structural rearrangements in human genomes harbor cryptic structures. *bioRxiv*, pages 2024–12, 2024.

[136] Ernest T Lam, Alex Hastie, Chin Lin, Dean Ehrlich, Somes K Das, Michael D Austin, Paru Deshpande, Han Cao, Niranjan Nagarajan, Ming Xiao, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology*, 30(8): 771–776, 2012.

[137] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 176(3): 663–675, 2019.

[138] Vincent CT Hanlon, Carl-Adam Mattsson, Diana CJ Spierings, Victor Guryev, and Peter M Lansdorp. Invertyper: Bayesian inversion genotyping with strand-seq data. *BMC genomics*, 22:1–8, 2021.

[139] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

[140] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[141] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440, 2022.

[142] Kiana Mohajeri, Stuart Cantsilieris, John Huddleston, Bradley J Nelson, Bradley P Coe, Catarina D Campbell, Carl Baker, Lana Harshman, Katherine M Munson, Zev N Kronenberg, et al. Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the chromosome 8p23. 1 region. *Genome research*, 26(11):1453–1467, 2016.

[143] Eugene Bragin, Eleni A Chatzimichali, Caroline F Wright, Matthew E Hurles, Helen V Firth, A Paul Bevan, and G Jawahar Swaminathan. Decipher: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic acids research*, 42(D1):D993–D1000, 2014.

[144] Francesca Antonacci, Jeffrey M Kidd, Tomas Marques-Bonet, Mario Ventura, Priscillia Siswara, Zhaoshi Jiang, and Evan E Eichler. Characterization of six human disease-associated inversion polymorphisms. *Human molecular genetics*, 18(14):2555–2566, 2009.

[145] Alfred H Sturtevant. Genetic factors affecting the strength of linkage in drosophila. *Proceedings of the National Academy of Sciences*, 3(9):555–558, 1917.

[146] Roberto Ciccone, T Mattina, R Giorda, MC Bonaglia, Mariano Rocchi, T Pramparo, and O Zuffardi. Inversion polymorphisms and non-contiguous terminal deletions: the cause and the (unpredicted) effect of our genome architecture. *Journal of medical genetics*, 43(5):e19–e19, 2006.

[147] Sabrina Giglio, Karl W Broman, Naomichi Matsumoto, Vladimiro Calvari, Giorgio Gimelli, Thomas Neumann, Hirofumi Ohashi, Lucille Voullaire, Daniela Larizza, Roberto Giorda, et al. Olfactory receptor–gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *The American Journal of Human Genetics*, 68(4):874–883, 2001.

[148] Roberto Giorda, Roberto Ciccone, Giorgio Gimelli, Tiziano Pramparo, Silvana Beri, Maria Clara Bonaglia, Sabrina Giglio, Maurizio Genuardi, Jesus Argente, Mariano Rocchi, et al. Two classes of low-copy repeats comediate a new recurrent rearrangement consisting of duplication at 8p23. 1 and triplication at 8p23. 2. *Human mutation*, 28(5):459–468, 2007.

[149] Fawaz Dabbaghie, Jana Ebler, and Tobias Marschall. Bubblegun: enumerating bubbles and superbubbles in genome graphs. *Bioinformatics*, 38(17):4217–4219, 2022.

[150] Erik Garrison, Andrea Guarracino, Simon Heumos, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, Sebastian Vorbrugg, Santiago Marco-Sola, Christian Kubica, et al. Building pangenome graphs. *Nature Methods*, pages 1–5, 2024.

[151] Andrea Guarracino, Njagi Mwaniki, Santiago Marco-Sola, and Erik Garrison. wfmash: whole-chromosome pairwise alignment using the hierarchical wavefront algorithm, 2021. URL `https://github.com/ekg/wfmash`. Released on 2021-09-09.

[152] Erik Garrison and Andrea Guarracino. Unbiased pangenome graphs. *Bioinformatics*, 39(1): btac743, 2023.

[153] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome biology*, 21:1–19, 2020.

[154] Joel Armstrong, Glenn Hickey, Mark Diekhans, Ian T Fiddes, Adam M Novak, Alden Deran, Qi Fang, Duo Xie, Shaohong Feng, Josefin Stiller, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, 587(7833):246–251, 2020.

[155] Guillaume Holley and Páll Melsted. Bifrost: highly parallel construction and indexing of colored and compacted de bruijn graphs. *Genome biology*, 21:1–20, 2020.

[156] Glenn Hickey, Jean Monlong, Jana Ebler, Adam M Novak, Jordan M Eizenga, Yan Gao, Tobias Marschall, Heng Li, and Benedict Paten. Pangenome graph construction from genome alignments with minigraph-cactus. *Nature biotechnology*, 42(4):663–673, 2024.

[157] Barış Ekim, Bonnie Berger, and Rayan Chikhi. Minimizer-space de bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer. *Cell systems*, 12(10): 958–968, 2021.

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

[159] Jennifer A Lee and James R Lupski. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, 52(1):103–121, 2006.

[160] Ann CM Smith and Andrea L Gropman. Smith–Magenis syndrome. *Cassidy and Allanson's management of genetic syndromes*, pages 863–893, 2021.

[161] Ken-Shiung Chen, Prasad Manian, Thearith Koeuth, Lorraine Potocki, Qi Zhao, A Craig Chinault, Cheng Chi Lee, and James R Lupski. Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nature genetics*, 17(2):154–163, 1997.

[162] Sarah H Elsea and Santhosh Girirajan. Smith–Magenis syndrome. *European Journal of Human Genetics*, 16(4):412–421, 2008.

[163] Lorraine Potocki, Ken-Shiung Chen, Sung-Sup Park, Doreen E Osterholm, Marjorie A Withers, Virginia Kimonis, Anne M Summers, Wendy S Meschino, Kwame Anyane-Yeboa, Catherine D Kashork, et al. Molecular mechanism for duplication 17p11. 2—the homologous recombination reciprocal of the Smith-Magenis microdeletion. *Nature genetics*, 24(1):84–87, 2000.

[164] Mariateresa Falco, Sonia Amabile, and Fabio Acquaviva. RAI1 gene mutations: mechanisms of Smith–Magenis syndrome. *The application of clinical genetics*, pages 85–94, 2017.

[165] Weimin Bi, G Mustafa Saifi, Santhosh Girirajan, Xin Shi, Barbara Szomju, Helen Firth, R Ellen Magenis, Lorraine Potocki, Sarah H Elsea, and James R Lupski. RAI1 point mutations, CAG repeat variation, and SNP analysis in non-deletion Smith–Magenis syndrome. *American Journal of Medical Genetics Part A*, 140(22):2454–2463, 2006.

[166] Sung-Sup Park, Paweł Stankiewicz, Weimin Bi, Christine Shaw, Jessica Lehoczky, Ken Dewar, Bruce Birren, and James R Lupski. Structure and evolution of the Smith-Magenis syndrome repeat gene clusters, SMS-REPs. *Genome research*, 12(5):729–738, 2002.

[167] Pawel Stankiewicz, Weimin Bi, and James R Lupski. Smith-Magenis syndrome deletion, reciprocal duplication dup (17)(p11. 2p11. 2), and other proximal 17p rearrangements. *Genomic Disorders: The Genomic Basis of Disease*, pages 179–191, 2006.

[168] Thoas Fioretos, Bodil Strömbeck, Therese Sandberg, Bertil Johansson, Rolf Billstrom, Åke Borg, Per-Gunnar Nilsson, Herman Van Den Berghe, Anne Hagemeijer, Felix Mitelman, et al. Isochromosome 17q in blast crisis of chronic myeloid leukemia and in other hematologic malignancies is the result of clustered breakpoints in 17p11 and is not associated with coding TP53 mutations. *Blood, The Journal of the American Society of Hematology*, 94(1):225–232, 1999.

[169] Wolfram G Scheurlen, Georg C Schwabe, Peter Seranski, Stefan Joos, Jochen Harbott, Simone Metzke, Hartmut Döhner, Annemarie Poustka, Klaus Wilgenbus, and Oskar A Haas. Mapping of the breakpoints on the short arm of chromosome 17 in neoplasms with an i (17q). *Genes, Chromosomes and Cancer*, 25(3):230–240, 1999.

[170] Maija Tarkkanen, Ritva Karhu, Anne Kallioniemi, Inkeri Elomaa, Aarne H Kivioja, Juha Nevalainen, Tom Böhling, Erkki Karaharju, Eija Hyytinen, Sakari Knuutila, et al. Gains and losses of DNA sequences in osteosarcomas by comparative genomic hybridization. *Cancer research*, 55(6):1334–1338, 1995.

[171] Michael C Frühwald, M Sue O'Dorisio, Zunyan Dai, Laura J Rush, Ralf Krahe, Dominic J Smiraglia, Torsten Pietsch, Sarah H Elsea, and Christoph Plass. Aberrant hypermethylation of the major breakpoint cluster region in 17p11. 2 in medulloblastomas but not supratentorial PNETs. *Genes, Chromosomes and Cancer*, 30(1):38–47, 2001.

[172] Ramesh C Juyal, Akira Kuwano, Ikuko Kondo, Federico Zara, Antonio Baldini, and Pragna I Patel. Mosaicism for del (17)(p11. 2p11. 2) underlying the Smith-Magenis syndrome. *American journal of medical genetics*, 66(2):193–196, 1996.

[173] Gerardo Perez, Galt P Barber, Anna Benet-Pages, Jonathan Casper, Hiram Clawson, Mark Diekhans, Clay Fischer, Jairo Navarro Gonzalez, Angie S Hinrichs, Christopher M Lee, et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Research*, 53(D1):D1243–D1249, 2025.

[174] Weimin Bi, Jiong Yan, Paweł Stankiewicz, Sung-Sup Park, Katherina Walz, Cornelius F Boerkoel, Lorraine Potocki, Lisa G Shaffer, Koen Devriendt, Małgorzata JM Nowaczyk, et al. Genes in a refined Smith-Magenis syndrome critical deletion interval on chromosome 17p11. 2 and the syntenic region of the mouse. *Genome research*, 12(5):713–728, 2002.

[175] Chen-Shan Chin, Sairam Behera, Asif Khalak, Fritz J Sedlazeck, Peter H Sudmant, Justin Wagner, and Justin M Zook. Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nature Methods*, 20(8):1213–1221, 2023.

[176] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. NCBI BLAST: a better web interface. *Nucleic acids research*, 36(suppl_2):W5–W9, 2008.

[177] Van K Ma, Rong Mao, Jessica N Toth, Makenzie L Fulmer, Alena S Egense, and Suma P Shankar. Prader-Willi and Angelman syndromes: mechanisms and management. *The application of clinical genetics*, pages 41–52, 2023.

[178] JH Chai, DP Locke, JM Greally, JHM Knoll, T Ohta, J Dunai, A Yavor, EE Eichler, and RD Nicholls. Identification of four highly conserved genes between breakpoint hotspots BP1 and BP2 of the Prader-Willi/Angelman syndromes deletion region that have undergone evolutionary transposition mediated by flanking duplicons. *The American Journal of Human Genetics*, 73(4):898–925, 2003.

[179] Douglas C Bittel, Nataliya Kibiryeva, and Merlin G Butler. Expression of 4 genes between chromosome 15 breakpoints 1 and 2 and behavioral outcomes in Prader-Willi syndrome. *Pediatrics*, 118(4):e1276–e1283, 2006.

[180] Hiago Azevedo Cintra, Danielle Nascimento Rocha, Ana Carolina Carioca da Costa, Latife Salomão Tyszler, Silvia Freitas, Leonardo Abreu de Araujo, Lisanne Incoutto Crozoe, Luísa Ribeiro de Paula, Patricia Santana Correia, Leonardo Henrique Ferreira Gomes, et al. Investigating the correlation between genotype and phenotype in Prader-Willi syndrome: a study of 45 cases from Brazil. *Orphanet Journal of Rare Diseases*, 19(1):240, 2024.

[181] Marius Keute, Meghan T Miller, Michelle L Krishnan, Anjali Sadhwani, Stormy Chamberlain, Ronald L Thibert, Wen-Hann Tan, Lynne M Bird, and Joerg F Hipp. Angelman syndrome genotypes manifest varying degrees of clinical severity and developmental impairment. *Molecular Psychiatry*, 26(7):3625–3633, 2021.

[182] Helen V Firth, Shola M Richards, A Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter. DECIPHER: database

of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4):524–533, 2009.

[183] MC Digilio, ML Dentici, S Loddo, L Laino, G Calcagni, S Genovese, R Capolino, I Bottillo, G Calvieri, B Dallapiccola, et al. Congenital heart defects in the recurrent 2q13 deletion syndrome. *European Journal of Medical Genetics*, 65(1):104381, 2022.

[184] Seth Septer, Lei Zhang, Caitlin E Lawson, Jose Cocjin, Thomas Attard, and Holly H Ardinger. Aggressive juvenile polyposis in children with chromosome 10q23 deletion. *World Journal of Gastroenterology: WJG*, 19(14):2286, 2013.

[185] Piero Pavone, Andrea D Praticò, Corrado Campisi, and Raffaele Falsaperla. A mild phenotype associated with a de novo microdeletion 10q23. 1-q23. 2: a new patient with a novel feature. *Case Reports*, 2016:bcr2016214388, 2016.

[186] Shanning Wan, Yunyun Zheng, Yinghui Dang, Tingting Song, Biliang Chen, and Jianfang Zhang. Prenatal diagnosis of 17q12 microdeletion and microduplication syndrome in fetuses with congenital renal abnormalities. *Molecular Cytogenetics*, 12:1–4, 2019.

[187] Parithi Balachandran, Isha A Walawalkar, Jacob I Flores, Jacob N Dayton, Peter A Audano, and Christine R Beck. Transposable element-mediated rearrangements are prevalent in human genomes. *Nature Communications*, 13(1):7115, 2022.

[188] Alan S Waldman and R Michael Liskay. Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Molecular and cellular biology*, 8(12):5350–5357, 1988.

[189] Jeffrey Rubnitz and Suresh Subramani. The minimum amount of homology required for homologous recombination in mammalian cells. *Molecular and cellular biology*, 4(11):2253–2258, 1984.

[190] James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24, 2011.

Appendix A

k-merald: Allele detection using *k*-mer based sequencing error profiles

A.1 Data Availability

Following data has been used for generation of the results described in Section 2.3.

PacBio CLR

HG002: `https://s3-us-west-2.amazonaws.com/human-pangenomics/NHGRI_UCSC_panel/HG002/hpp_HG002_NA24385_son_v1/PacBio_CLR/PB_HG002-CLR-SvDetection/m64070_190824_163708.subreads.bam`
pbmm2 (https://github.com/PacificBiosciences/pbmm2) was used for alignment of sequencing reads to GRCh38 reference

```
pbmm2 align GRCh38.fa m64070_190824_163708.subreads.bam alignments.bam --sort
↪  --median-filter --sample HG002
```

PacBio Hifi

HG002: `https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC_PLUS/HG002/analysis/aligned_reads/hifi/GRCh38/HG002_aligned_GRCh38_winnowmap.sorted.bam`

Oxford Nanopore

HG002: `https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/Ultralong_OxfordNanopore/guppy-V3.4.5/HG002_GRCh38_ONT-UL_GIAB_20200204.bam`

HG001: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Ultralong_OxfordNanopore/NA12878-minion-ul_GRCh38.bam

Variant callsets

HG002: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh38/HG002_GRCh38_1_22_v4.2.1_benchmark.vcf.gz
HG001: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh38/HG001_GRCh38_1_22_v4.2.1_benchmark.bed

Genome stratifications

https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/genome-stratifications/v3.0/GRCh38/

Appendix B

Analyzing homology-mediated recurrent deletion polymorphisms

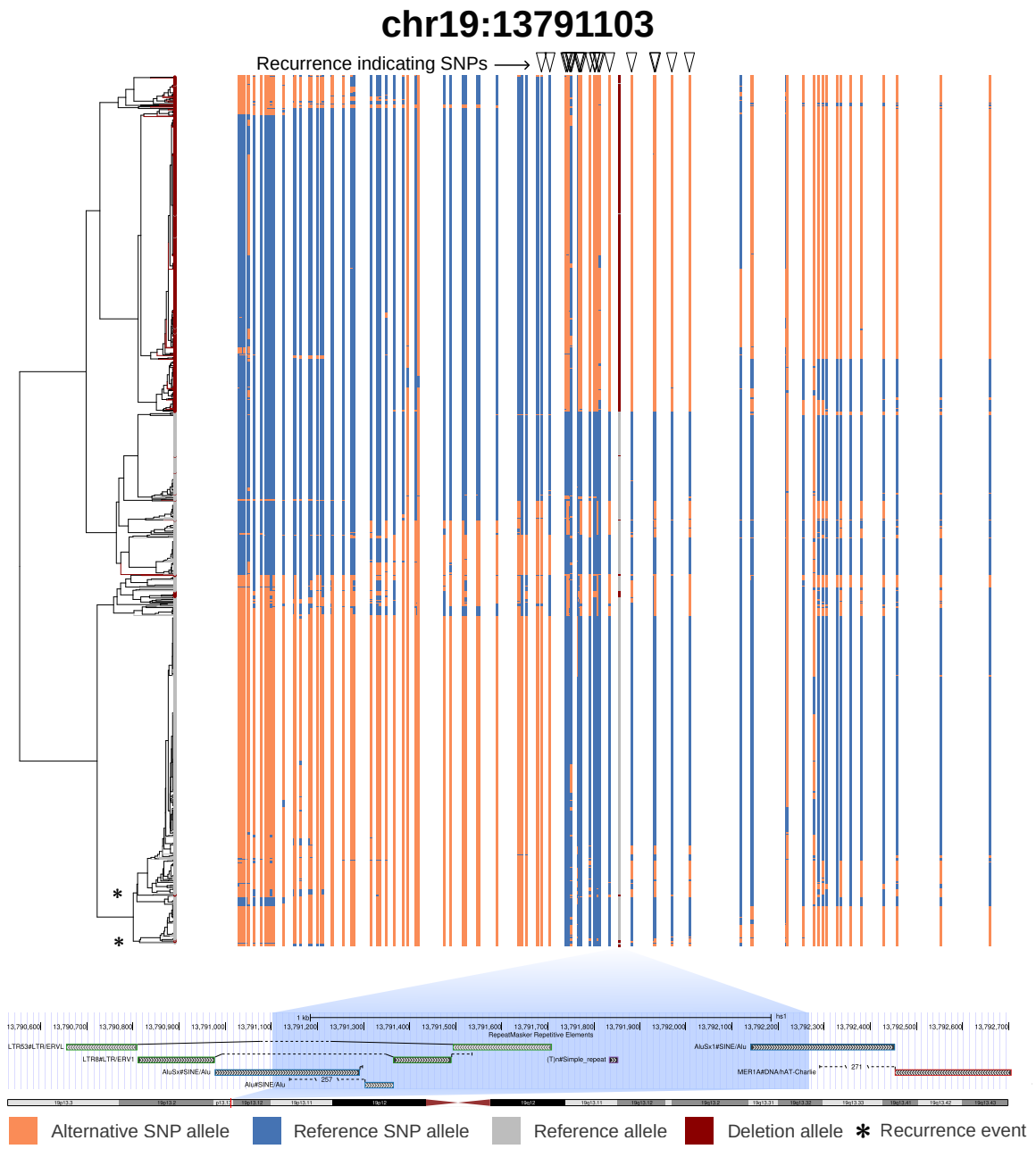


Figure B.1: Alu-mediated recurrent deletion at 19p13.13. A predicted recurrent deletion at 19p13.13 mediated by an AluSx-AluSx pair. The figure represents the variation of haplotypes in a 100 kbp window centered around the deletion and the relationship between haplotypes with (red) and without the deletion (gray). Dendrograms of haplotypes are plotted using centroid hierarchical clustering algorithm. In each haplotype, reference and alternative alleles are shown in blue and orange, respectively. SNPs within 20 kbp around the deletion showing evidence of deletion recurrence are marked by triangles at the top. Two predicted independent occurrences of the deletion event are marked by *. The UCSC Genome Browser [173] view, with the region of interest highlighted in blue, is shown at the bottom of the figure.

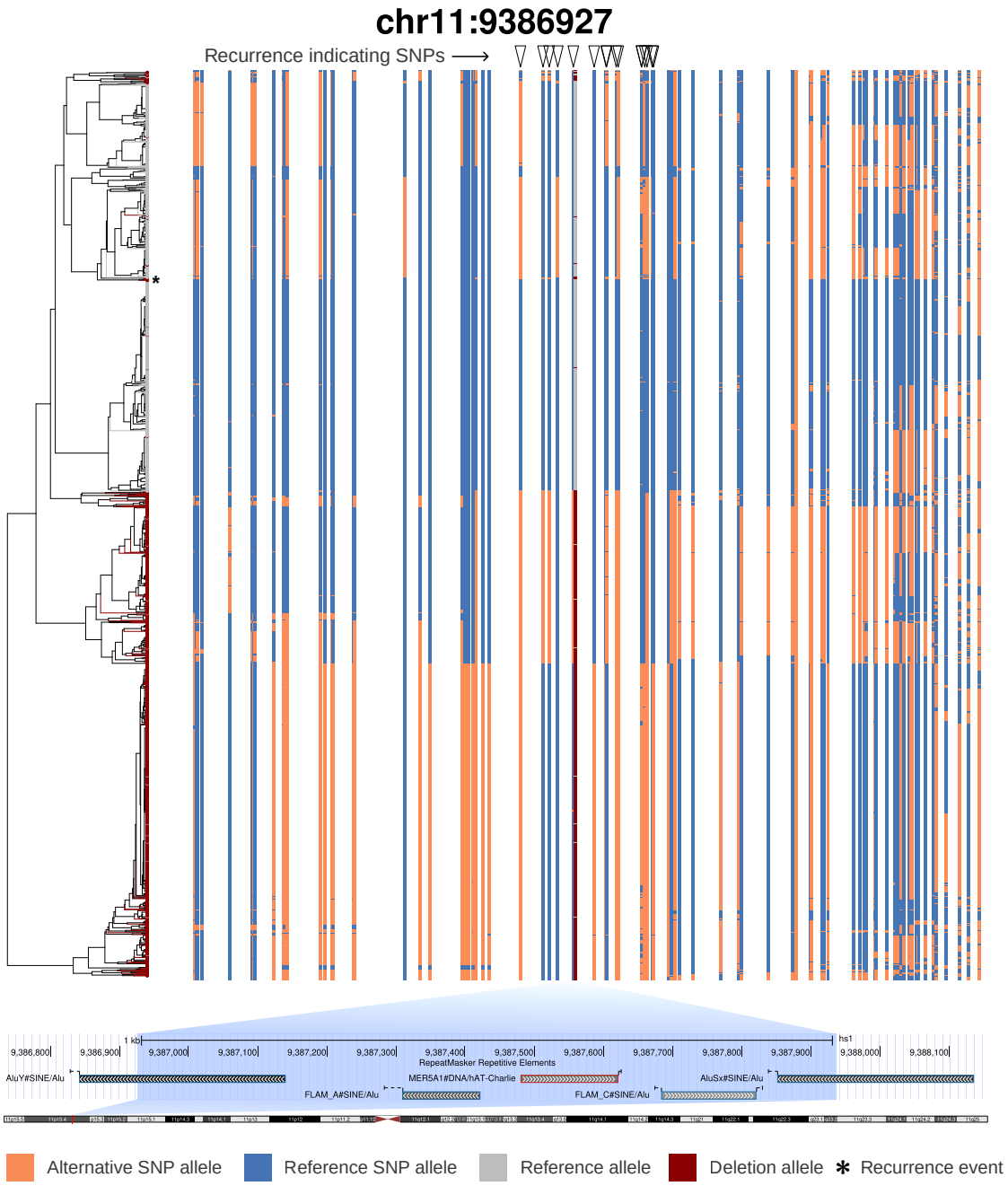


Figure B.2: Alu-mediated recurrent deletion at 11p15.4. A predicted recurrent deletion at 11p15.4 mediated by an AluY-AluSx pair. The figure depicts the variation of haplotypes in a 100 kbp window centered around the deletion and the relationship between haplotypes with (red) and without the deletion (gray). Dendrograms of haplotypes are plotted using centroid hierarchical clustering algorithm. In each haplotype, reference and alternative alleles are shown in blue and orange, respectively. Recurrence-indicating SNPs within 20 kbp region around the deletion are marked by triangles at the top. A predicted independent occurrence of the deletion event is marked by *. The UCSC Genome Browser [173] view, with the region of interest highlighted in blue, is shown at the bottom of the figure.

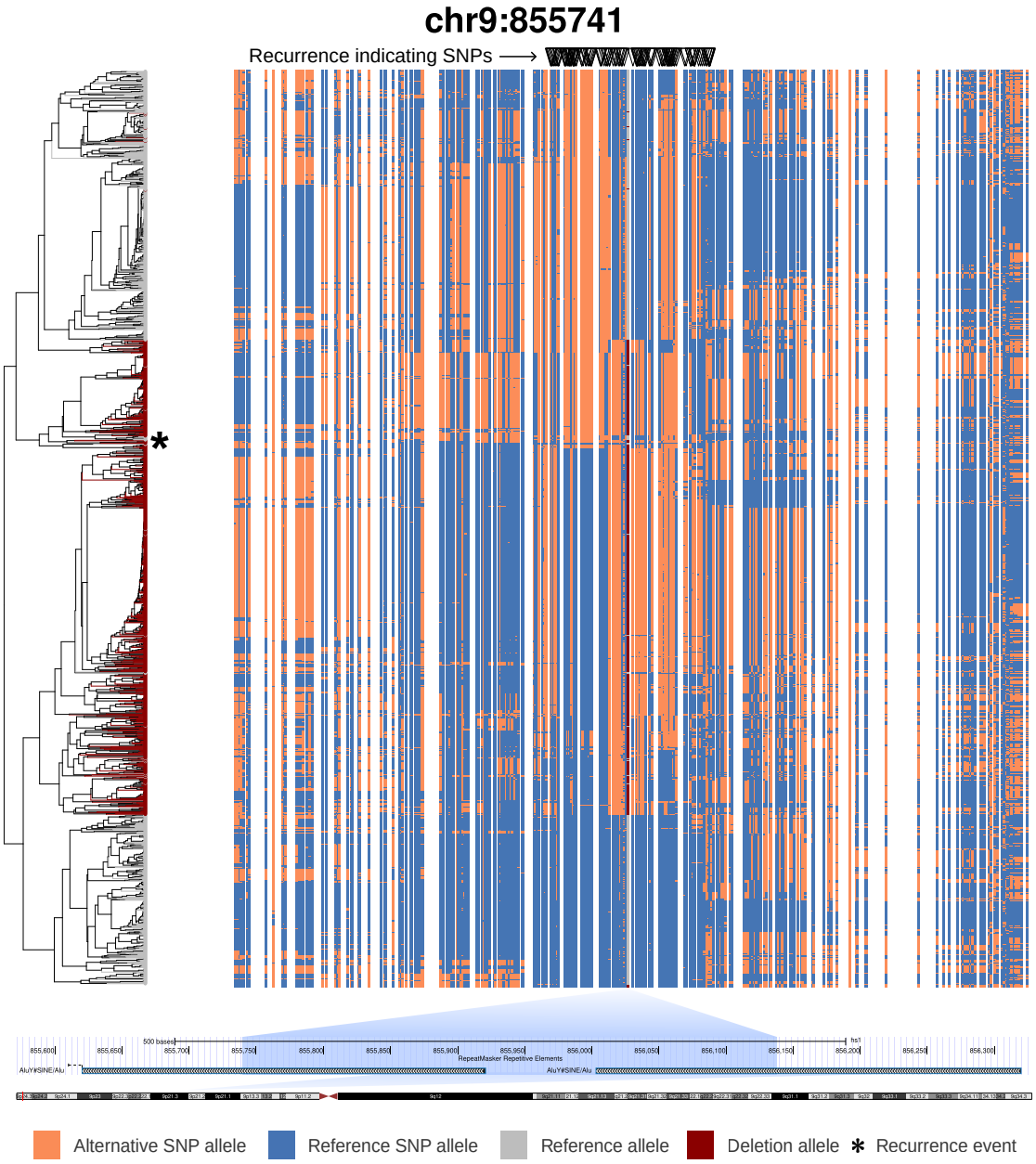


Figure B.3: Alu-mediated recurrent deletion at 9p24.3. A predicted recurrent deletion at 9p24.3 mediated by an AluY-AluY pair. The figure shows the variation of haplotypes in a 100 kbp window centered around the deletion and the relationship between haplotypes with (red) and without the deletion (grey). Dendrograms of haplotypes are plotted using centroid hierarchical clustering algorithm. In each haplotype, reference and alternative alleles are shown in blue and orange, respectively. Recurrence-indicating SNPs within 20 kbp around the deletion are marked by triangles at the top. A group of haplotypes not carrying the deletion (gray) but appearing with the haplotypes carrying it (red), hence supporting the recurrence of the respective deletion event, is marked by *. The UCSC Genome Browser [173] view, with the region of interest highlighted in blue, is shown at the bottom of the figure.

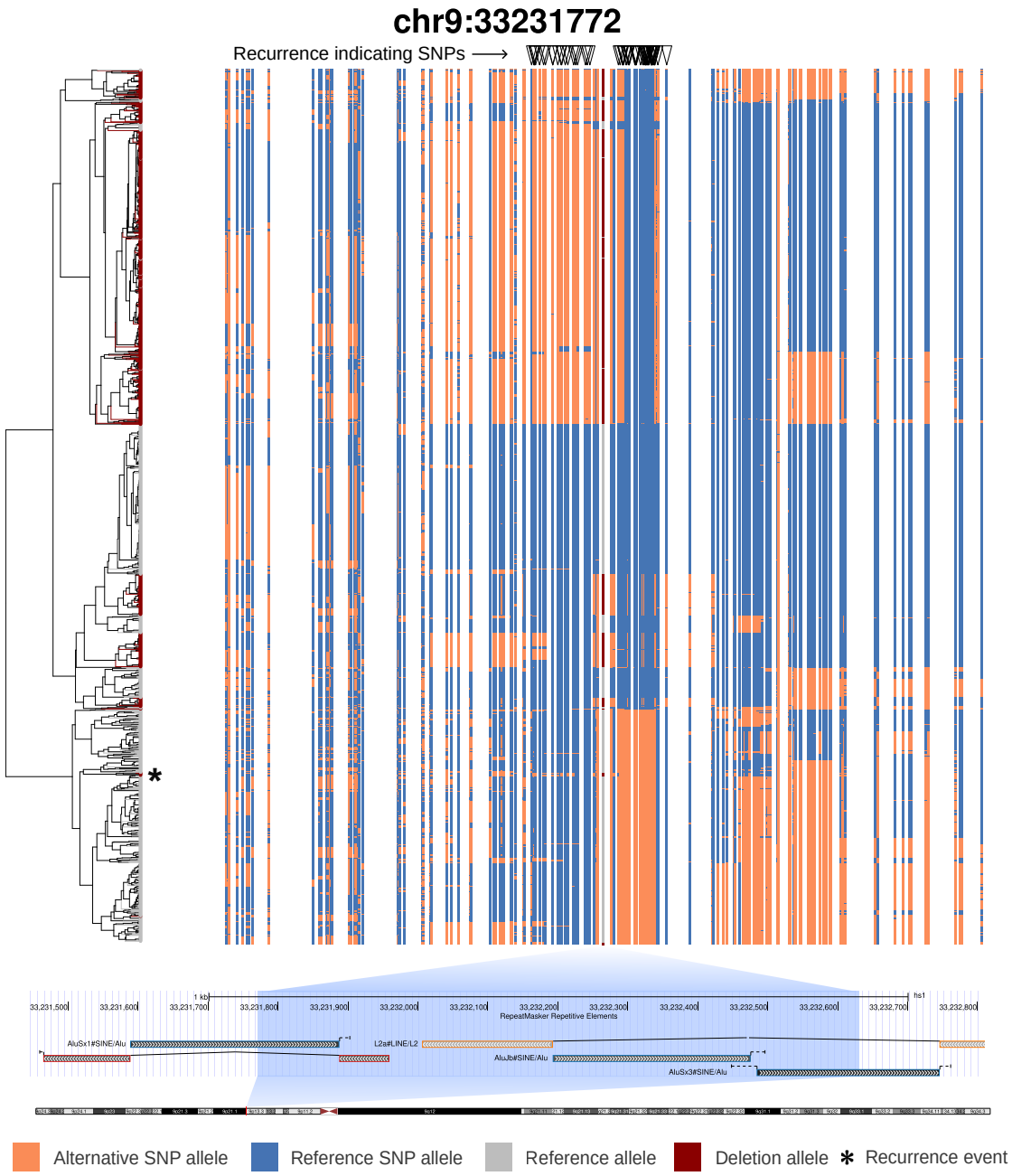


Figure B.4: Alu-mediated recurrent deletion at 9p13.3. A predicted recurrent deletion at 9p13.3 mediated by an AluSx-AluSx pair. The figure displays the variation of haplotypes in a 100 kbp window centered around the deletion and the relationship between haplotypes with (red) and without the deletion (gray). Dendrograms of haplotypes are plotted using centroid hierarchical clustering algorithm. In each haplotype, reference and alternative alleles are shown in blue and orange, respectively. Recurrence-indicating SNPs within 20 kbp around the deletion are marked by triangles at the top. A predicted independent occurrence of the deletion event is marked by *. The UCSC Genome Browser [173] view, with the region of interest highlighted in blue, is shown at the bottom of the figure.

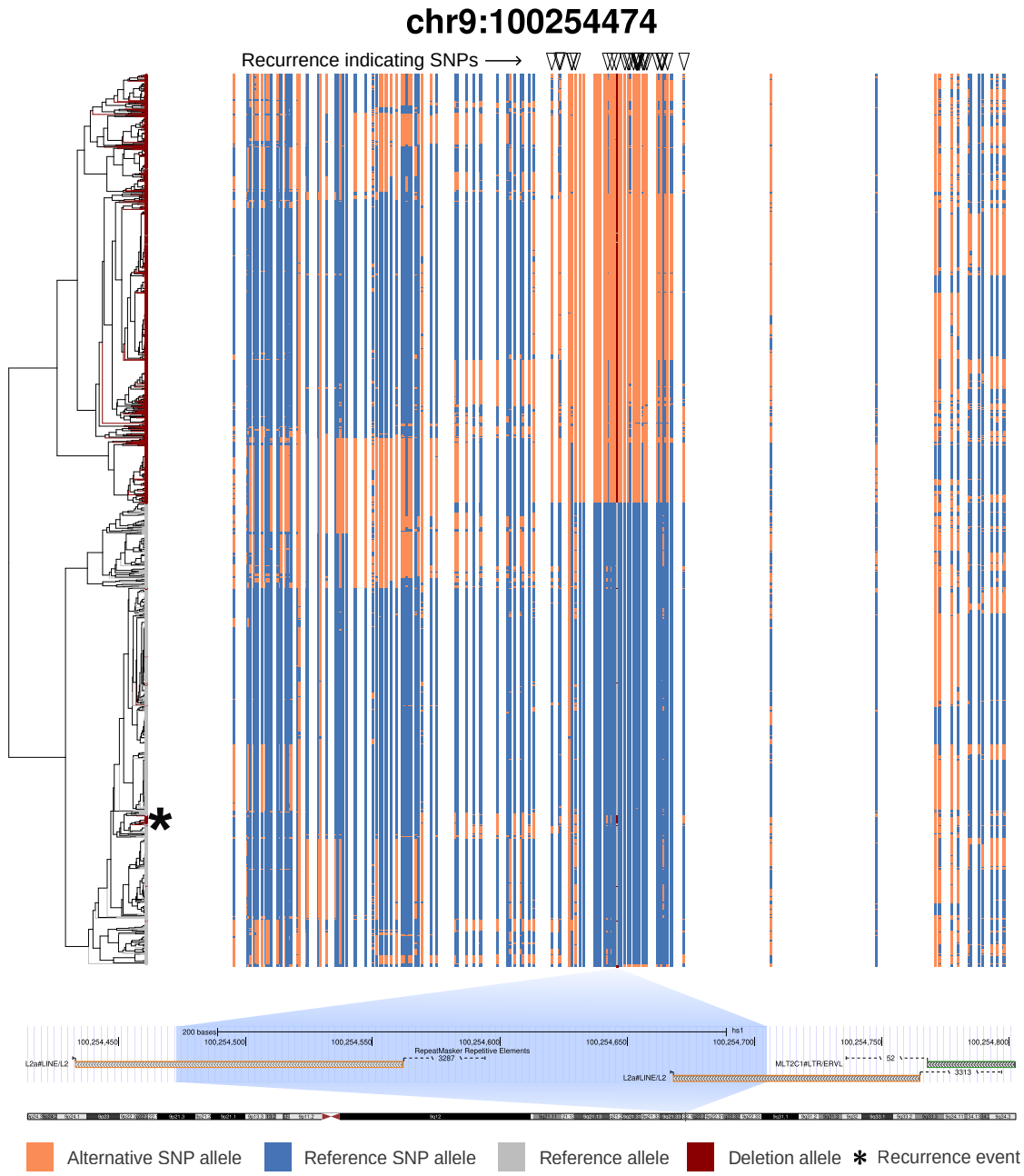


Figure B.5: L2-mediated recurrent deletion at 9q22.1. A predicted recurrent deletion at 9q22.1 mediated by an L2-L2 pair. Figure shows the variation of haplotypes in a 100 kbp window centered around the deletion and the relationship between haplotypes with (red) and without the deletion (gray). Dendrograms of haplotypes are plotted using centroid hierarchical clustering algorithm. In each haplotype, reference and alternative alleles are shown in blue and orange, respectively. Recurrence-indicating SNPs within 20 kbp around the deletion are marked by triangles at the top. A predicted independent occurrence of the deletion event is marked by *. The UCSC Genome Browser [173] view, with the region of interest highlighted in blue, is shown at the bottom of the figure.

Appendix C

Published articles underlying this thesis

C.1 Allele detection using *k*-mer-based sequencing error profiles

The manuscript "Allele detection using *k*-mer-based sequencing error profiles" [36] was published in *Bioinformatics Advances*. Author information, author contributions, license and copyright information are listed in the subsections below.

C.1.1 Authors

Hufsah Ashraf, Jana Ebler and Tobias Marschall

C.1.2 Contributions

Tobias Marschall and I designed the algorithm and the study. I developed *k*-merald and performed all the experiments. I and Tobias Marschall wrote a draft of the paper and Jana Ebler contributed edits and comments. All authors approved the final manuscript.

C.1.3 License and copyright information

C.2 Haplotype-resolved diverse human genomes and integrated analysis of structural variation

The manuscript “Haplotype-resolved diverse human genomes and integrated analysis of structural variation” [12] was published in *Science*. Author information, author contributions, license and copyright information are listed in the subsections below.

C.2.1 Authors

Peter Ebert*, Peter A. Audano*, Qihui Zhu*, Bernardo Rodriguez-Martin*, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T. Chuang, Xiaofei Yang, Katherine M. Munson, Alexandra P. Lewis, Susan Fairley, Luke J. Tallon, Wayne E. Clarke, Anna O. Basile, Marta Byrska-Bishop, André Corvelo, Uday S. Evani, Tsung-Yu Lu, Mark J. P. Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M. Wenger, Maryam Ghareghani, William T. Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A. Regier, Haley J. Abel, Ira M. Hall, Paul Flicek, Oliver Stegle, Mark B. Gerstein, Jose M. C. Tubio, Zepeng Mu, Yang I. Li, Xinghua Shi, Alex R. Hastie, Kai Ye, Zechen Chong, Ashley D. Sanders, Michael C. Zody, Michael E. Talkowski, Ryan E. Mills, Scott E. Devine, Charles Lee, Jan O. Korbel, Tobias Marschall and Evan E. Eichler

* joint first authors

C.2.2 Contributions

Author contributions as stated in the manuscript [12]:

> “PacBio production sequencing: K.M.M., A.P.L., Q.Z., L.J.T., and S.E.D. Strand-seq production: A.D.S., B.R., P.H., and J.O.K. Phased genome assembly: P.E., P.A.A., D.P., Q.Z., F.Y., W.T.H., and T.M. Assembly analysis: P.E. Assembly-based variant calling: P.A.A. Variant QC, merging, and annotation: P.A.A., T.R., M.J.P.C., J.R., T.L., Z.C., Y.C., K.Y., J.L., X.Y., and J.O.K. Assembly scaffolding: F.Y., D.P., and P.E. Additional long-read callsets: P.A.A., Y.C., Z.C., W.T.H., J.R., and A.M.W. Short-read SV calling and merging: X.Z., Q.Z., H.J.A., H.B., N.T.C., W.E.C., A.C., U.S.E., S.E.D., I.M.H., W.T.H., A.A.R., M.C.Z., and M.E.T. Bionano Genomics SV discovery and analysis: F.Y., J.L., and A.R.H. Strand-seq inversion detection and genotyping: D.P., W.T.H., H.A., M.G., T.M., A.D.S., and J.O.K. MEI discovery and integration: B.R.-M., W.Z., M.S., N.T.C., J.M.C.T., J.O.K., R.E.M., and S.E.D. Variant hotspot analysis: D.P. and E.E.E. Breakpoint analysis: S.K., J.L., X.Y., M.G., K.Y., and J.O.K. PanGenie genotyping: J.E. and T.M. Illumina genotype analysis: J.E., X.Z., W.E.C., P.E., T.R., P.A.A., H.B., J.O.K., M.E.T.,

> M.C.Z., and T.M. RNA-seq and QTL analysis: M.J.B., A.S., Z.M., J.C., C.L., M.B.-B., A.O.B., O.S., Y.I.L., X.S., M.C.Z., and J.O.K. Ancestry and population genetic analyses: PH.H., R.S.M., P.A.A., T.M., and E.E.E. Data archiving: S.F., P.A.A., K.M.M., and P.F. Organization of supplementary materials: Q.Z. and C.L. Display items: P.A.A., P.E., J.E., A.R.H., PH.H., R.S.M., T.M., D.P., T.R., B.R.-M., M.S., F.Y., X.Z., and W.Z. Manuscript writing: P.A.A., P.E., B.R.-M., A.S., D.P., PH.H., Q.Z., F.Y., A.R.H., J.L., M.E.T., M.J.B., X.S., S.E.D., J.O.K., T.M., and E.E.E. HGSVC Co-chairs: C.L., J.O.K., and E.E.E."

I, in collaboration with Wolfram Höps, contributed towards Strand-Seq based inversion genotyping using ArbiGent and developing the integrated inversion callset.

C.2.3 License and copyright information

According to the AAAS Rights & Permissions office the following holds:

> "After publication of a manuscript in an AAAS journal, the author may reprint their full manuscript or portions of the manuscript in a thesis or dissertation written by the author as part of a course of study at an educational institution in print & electronic formats. Credit must be given to the first appearance of the material in the appropriate issue of the AAAS journal."

C.3 Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders

The manuscript "Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders" [13] was published in *Cell*. Author information, author contributions, license and copyright information are listed in the subsections below.

C.3.1 Authors

David Porubsky*, Wolfram Höps*, Hufsah Ashraf*, PingHsun Hsieh, Bernardo Rodriguez-Martin, Feyza Yilmaz, Jana Ebler, Pille Hallast, Flavia Angela Maria Maggiolini, William T. Harvey, Barbara Henning, Peter A. Audano, David S. Gordon, Peter Ebert, Patrick Hasenfeld, Eva Benito, Qihui Zhu, Charles Lee, Francesca Antonacci, Matthias Steinrücken, Christine R. Beck, Ashley D. Sanders, Tobias Marschall, Evan E. Eichler and Jan O. Korbel

* joint first authors

C.3.2 Contributions

Author contributions as stated in the manuscript [13]:

> "Conceptualization, D.P., A.D.S., T.M., E.E.E., and J.O.K.; methodology & software, D.P., W.H., H.A., P. Hsieh, B.R.-M., and M.S.; formal analysis, D.P., W.H., H.A., P. Hsieh, B.R.-M., F.Y., J.E., and P. Hallast; investigation, D.P., W.H., H.A., P. Hsieh, B.R.-M., A.D.S., M.S., and C.R.B.; resources, HGSVC, Q.Z., C.L., P. Hasenfeld, A.D.S., T.M., E.E.E., and J.O.K.; computational support, W.T.H., P.A.A., B.H., and D.S.G.; validation, F.A.M.M., P.E., E.B., and F.A.; writing, D.P., W.H., H.A., B.R.-M., T.M., E.E.E., and J.O.K., with input from all authors."

In this study, I worked along side Wolfram Höps to refine ArbiGent's performance, perform validation experiments, Strand-Seq based inversion genotyping and integrating the inversion calls from different technologies into one comprehensive inversion callset. Additionally, I developed the "toggling-indicating SNPs (tiSNPs)-based approach" for detecting recurrent inversion polymorphisms in human genomes and performed several downstream analyses using its findings in order to understand the mechanism driving the recurrence phenomenon. In collaboration with PingHsun Hsieh and Matthias Steinrücken, I contributed towards providing a comprehensive list of recurrent inversions, including novel candidates.

C.3.3 License and copyright information

C.4 Inversion polymorphism in a complete human genome assembly

The manuscript "Inversion polymorphism in a complete human genome assembly" [14] was published in *Genome Biology*. Author information, author contributions, license and copyright information are listed in the subsections below.

C.4.1 Authors

David Porubsky, William T. Harvey, Allison N. Rozanski, Jana Ebler, Wolfram Höps, Hufsah Ashraf, Patrick Hasenfeld, Human Pangenome Reference Consortium (HPRC), Human Genome Structural Variation Consortium (HGSVC), Benedict Paten, Ashley D. Sanders, Tobias Marschall, Jan O. Korbel and Evan E. Eichler

C.4.2 Contributions

Author contributions as stated in the manuscript [14]:

> Conceptualization, D.P., E.E.E.; Formal analysis, D.P.; Investigation, D.P., E.E.E.; Inversion genotyping, W.H., H.A., J.E., T.M.; Strand-seq data generation, A.D.S., P.H., J.O.K.; Computational support, W.T.H., A.N.R.; Assembly resources, B.P.; Writing, D.P., E.E.E., with input from all authors. The authors read and approved the final manuscript.

Wolfram Höps and I contributed towards Strand-Seq based inversion genotyping using ArbiGent and developing the integrated inversion callset for this study.

C.4.3 License and copyright information

C.5 Complex genetic variation in nearly complete human genomes

The manuscript "Complex genetic variation in nearly complete human genomes" [15] is publicly available as a *bioRxiv* preprint and has been provisionally accepted for publication in *Nature*. Author information, author contributions, license and copyright information are listed in the subsections below.

C.5.1 Authors

Glennis A. Logsdon*, Peter Ebert*, Peter A. Audano*, Mark Loftus, David Porubsky, Jana Ebler, Feyza Yilmaz, Pille Hallast, Timofey Prodanov, DongAhn Yoo, Carolyn A. Paisie, William T. Harvey, Xuefang Zhao, Gianni V. Martino, Mir Henglin, Katherine M. Munson, Keon Rabbani, Chen-Shan Chin, Bida Gu, Hufsah Ashraf, Olanrewaju Austine-Orimoloye, Parithi Balachandran, Marc Jan Bonder, Haoyu Cheng, Zechen Chong, Jonathan Crabtree,

Mark Gerstein, Lisbeth A. Guethlein, Patrick Hasenfeld, Glenn Hickey, Kendra Hoekzema, Sarah E. Hunt, Matthew Jensen, Yunzhe Jiang, Sergey Koren, Youngjun Kwon, Chong Li, Heng Li, Jiaqi Li, Paul J. Norman, Keisuke K. Oshima, Benedict Paten, Adam M. Phillippy, Nicholas R Pollock, Tobias Rausch, Mikko Rautiainen, Stephan Scholz, Yuwei Song, Arda Söylev, Arvis Sulovari, Likhitha Surapaneni, Vasiliki Tsapalou, Weichen Zhou, Ying Zhou, Qihui Zhu, Michael C. Zody, Ryan E. Mills, Scott E. Devine, Xinghua Shi, Mike E. Talkowski, Mark J.P. Chaisson, Alexander T. Dilthey, Miriam K. Konkel, Jan O. Korbel, Charles Lee, Christine R. Beck, Evan E. Eichler, Tobias Marschall

* joint first authors

C.5.2 Contributions

Author contributions as stated in the manuscript [15]:

> Sample Selection: P.H., K.M.M., T.R., A.SV., C.Lee., and E.E.E.; Data Production: K.M.M., P.H., P.H.F., K.H., Q.Z., S.E.D.; Data Management: P.E., P.A.A., S.E.H., P.H., F.Y., K.M.M., Y.K., O.A., and L.S.; Assembly Production and Quality Control: P.E., W.T.H., M.H., Z.C., M.R., S.K., Y.K., H.C., A.M.P., Y.S., E.E.E., and T.M.; Variant Discovery: P.A.A, C.A.P., and C.R.B.; Mobile Elements: M.L., W.Z., P.B., R.E.M., J.C., S.E.D., C.R.B., and M.K.K.; Inversions: H.A., V.T., D.P., T.R., J.O.K., and T.M.; Segmental Duplications: D.Y., K.R., M.J.P.C., and E.E.E.; STR and VNTR Annotation: B.G. and M.J.P.C.; Chromosome Y: P.H., P.E., M.L., M.K.K., and C.Lee.; Iso-Seq Phasing: G.V.M., M.L., and M.K.K.; SV Impact on Genes: X.Z., G.V.M., M.L., M.E.T., and M.K.K.; Transcriptional Effects of SVs: G.V.M., M.L., M.J., Y.J., J.L., M.G., and M.K.K.; Hi-C and Additional Functional Analysis: C.LI., M.J.B., and X.S.; Genotyping: J.E., T.P., G.H., B.P., and T.M.; Integrated Reference Panel: J.E., T.R., M.C.Z. and T.M.; Major Histocompatibility Complex: M.L., S.T.S., C.S., Y.Z., N.R.P., P.J.N., L.A.G., P.A.A., P.E., A.S., T.P., C.R.B., H.L., T.M., M.K.K., and A.T.D.; Complex Structural Polymorphisms: P.A.A., D.P., F.Y., M.L., M.K.K., C.R.B., C.Lee., and E.E.E.; Centromeres: G.A.L., K.K.O., M.L., M.K.K., and E.E.E.; Manuscript Writing: G.A.L., P.E., P.A.A., M.L., D.P., J.E., F.Y., P.H., T.P., D.Y., X.Z., G.V.M., C.S., H.A., M.J., C.LI., X.S., M.E.T., M.J.P.C., A.T.D., M.K.K., J.O.K., C.Lee., C.R.B., E.E.E., and T.M. All authors read and approved the final manuscript. HGSVC co-chairs: J.O.K., C.Lee., E.E.E. and T.M.

In this study, I performed Strand-Seq based inversion genotyping using ArbiGent for both GRCh38 and T2T-CHM13 based inversion callsets and evaluated the genotyping results.

C.6 Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project

The manuscript “Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project” [38] is publicly available as a *bioRxiv* preprint and has been provisionally accepted for publication in *Nature*. Author information and author contributions are listed in the subsections below.

C.6.1 Authors

Siegfried Schloissnig*, Samarendra Pani*, Bernardo Rodriguez-Martin, Jana Ebler, Carsten Hain, Vasiliki Tsapalou, Arda Söylev, Patrick Hüther, Hufsah Ashraf, Timofey Prodanov, Mila Asparuhova, Sarah Hunt, Tobias Rausch, Tobias Marschall, Jan O. Korbel

* joint first authors

C.6.2 Contributions

The author contribution statement is taken from the manuscript currently under review.

> M.A., S.S. data acquisition; P.H., T.R. base calling, read alignment, and primary SV discovery vs. linear references; S.P. performed read-based haplotype phasing, haplotype-tagging of reads and SV genotyping; T.R. pursued statistical phasing of the SV genotypes; A.S. ran graph-based SV discovery with SVarp; J.E. implemented and ran the graph augmentation pipeline; S.P., T.R. constructed the final SV callset, and led SV callset benchmarking; S.S. analysis of geographic stratification of SV alleles; C.H. analyzed breakpoint homology lengths; B.R.M. and J.E.S.F. annotated SV classes and analyzed mobile elements; H.A. analyzed deletion recurrence; T.P. ran complex locus genotyping; T.F., W.S., and E.B. performed explorative analyses of DNA methylation; V.T., R.M.P and M.C. analyzed inversions; S.H. provided data management assistance; T.E.W. and F.J.P.-L. performed PacBio sequencing of rare disease patients; H.M., W.H., S.S., D.W., C.G., and T.M. analyzed rare disease patient genomes; T.R., T.M., J.O.K. jointly supervised the work; H.A, C.H., S.P, B.R.M., T.M, T.R, S.S., V.T. and J.E.S.F. prepared main figures.; H.A., J.E., C.H., J.O.K., T.M., S.P., T.P., B.R.M., A.S., S.S., V.T., T.R. wrote the manuscript, with input from all authors.

I performed the homology-mediated deletion recurrence detection and analysis included in this study.