UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES UNIDAD JURIQUILLA

GENOME-WIDE MAPS OF IDENTICAL REPEATS THAT CAN MEDIATE COMPLEX GENOMIC REARRANGEMENTS IN THREE HUMAN GENOME ASSEMBLIES

T E S I S

QUE PARA OBTENER EL GRADO DE:

LICENCIADO EN CIENCIAS GENÓMICAS

PRESENTA:

LUIS GERARDO FERNANDEZ LUNA

TUTOR:

DRA. CLAUDIA GONZAGA JAUREGUI

JURIQUILLA, QUERÉTARO, 2023

Acknowledgements

I would like to express my deepest gratitude to the individuals who have supported me throughout this journey that has now come to an end.

First and foremost, I want to thank my advisor, Claudia Gonzaga Jauregui, for accepting me into your lab. Under your guidance, expertise, and invaluable feedback, I have learned and grown tremendously. I greatly appreciate the opportunities you have given me, your incredible teaching, and the support you have provided throughout my career. I am forever grateful for your mentorship and for preparing me with the skills to tackle the next challenges in my scientific journey. I am deeply indebted to you.

I am also grateful to Chris Van Hout for always being there to listen to my questions and provide answers. Thank you for your insightful advice and for the ketchup. Your support has been invaluable.

To my committee members Alejandra Medina, Cris Van Hout, Rafael Palacios, and Cynthia, I want to express my gratitude for your time and effort in reviewing my thesis and offering valuable suggestions for improvement.

To the members of the Carvalho Lab, I want to express my gratitude for all the support and warmth I experienced during my internship. Special thanks to Michelle for helping me navigate through the project and for the daily cookies that always brightened my mornings. Your insightful advice about my academic career has been invaluable. I also want to thank the members of the Mendelian Lab, especially Tania, for their unconditional support.

To my closest friends, Valeria, Natalia, and Victor, thank you for your moral support and friendship.Your presence in my life has made a significant difference.

Last but certainly not least, to my parents, Luis and Yadira, and my sister, Nicole, for their unconditional love, effort, and support throughout this entire process.I am forever grateful for your encouragements and presence in my life.

Finally, to all these individuals and many others who have supported me along the way, I extend my deepest gratitude. Without your help, guidance, and love, I would not have been able to reach this milestone. Thank you from the bottom of my heart.

Abstract

The human genome is constantly evolving and shaping its structure through a wide diversity of mechanisms. Some major events that contribute to genome dynamics and evolution include genome-wide, chromosomal or regional duplications and structural rearrangements. At the chromosomal and regional scale, a major architectural feature that contributes to rearrangements is the presence of repetitive elements. These elements include a great variety of sequences ranging from interspersed to tandem repeats, including segmental duplications (SD) and low-copy repeats (LCRs) which are >1kb repeated elements spread throughout the genome but that share a high identity between copies and play key roles in promoting the plasticity and dynamic nature of the genome.

The presence of repetitive elements can impact the structure and function of the genome promoting genomic rearrangements, chromosomal instability and evolutionary dynamics. Genomic rearrangements can lead to the creation of copy number variants(CNVs) that can represent neutral or adaptive polymorphisms in the population but are mostly benign in function. However, some rearrangements can also lead to genomic disorders which are genetic conditions caused by alterations in the number or structure of large genomic regions containing generally dosage-sensitive genes within.

Understanding the characteristics of repeated sequences in the genome including their relative orientation, identity, size, distribution and density, is important to understand their role in the formation of structural variants. Through DNA recombination-based processes and replication-based processes, repeated genomic sequences can contribute to genomic instability and rearrangement susceptibility.

The work reported in this thesis focused on exploring and identifying repetitive sequences in the human genome that can be potentially involved in genomic rearrangements in the human genome. We performed genome-wide analyses and comparisons of direct and inverted repeats in the latest available human genome reference assemblies including GRCh37 and GRCh38 and the most recent telomere-to-telomere alternate assembly, T2T-CHM13. Through these analyses, I have produced a catalog of direct and inversely oriented repeated sequences across the currently three most widely used human genome assemblies. I explored their main characteristics and their potential contribution to human phenotypes by cross-referencing our repeats with genes across the genome. Bioinformatic analyses of these repeats and their contribution to genome architecture can reveal regions that are most susceptible to genomic instability. Overall, this work provides a genome-wide landscape of repetitive elements and their key features to understand complex genomic rearrangement formation and gain insights into the molecular mechanisms leading to genomic disorders and genome evolution.

Table of contents

List of abbreviations

LCR: Low-copy repeat

SD: Segmental duplication

NAHR: Non-allelic homologous recombination

MMBIR: Microhomology mediated break induced replication

FoSTeS: Fork stalling template switching

SINE: Short interspersed nuclear element

LINE: Long interspersed nuclear element

CNV: Copy number variant

LTR: Long terminal repeat

List of figures

represents the number of genes contained in the gene set. A) GO analysis for genes found in GRCh37. B) GO analysis for genes found in GRCh38. C) GO analysis for genes found in T2T-CHM13.

Figure 6. Genomic region 17p12 with Recurrent/Non-recurrent Rearrangement including *PMP22* Gene associated with Genomic Disorders. Displaying GRCh37, GRCh38, T2TCHM-13 respectively. Noting Highlighted Pair of Direct Repeats Flanking Dup/Del Region.

FIgure7. Percentage of inverted elements for each chromosome recovered/captured in our study that overlaps with other elements flanking inversions.

List of tables

1. Introduction

Genomic repetitive elements represent a wide range of non-unique genomic sequences that play important roles in shaping the genomes of species, influencing their evolution, and contributing to genetic and phenotypic variation(Liehr, 2021). They can affect chromosome structure, gene transcription, splicing, and may lead to structural variants. These elements are broadly classified into two types: tandem repeats and interspersed repeats.

Tandem repeats are short DNA sequences that repeat in a head-to-tail fashion, including satellites and simple repeats(Hauth & Joseph, 2002). On the other hand, interspersed repeats primarily consist of transposable elements (TEs), which can be classified based on how they spread. Class I elements, such as long interspersed elements (LINEs) and long terminal repeats (LTRs), use retrotransposition to spread and are considered autonomous as they encode their own enzymes. Nonautonomous elements, like short interspersed elements (SINEs) and composite retroelements (SVAs), rely on LINE-encoded proteins for retrotransposition. Class II elements, including Tc1-Mariner and hAT, propagate through a transposase helicase (Wells & Feschotte, 2020).

These repetitive elements can influence gene expression, regulation, and genomic rearrangements as they serve as substrates for recombination and replication-related processes, leading to the formation of new structural variants(Pappalardo & Barra, 2021; Zepeda-Mendoza et al., 2010). Additionally, segmental duplications (SDs) or low-copy repeats are duplicated DNA segments found throughout the genome. SDs share high sequence identity (around 90%) and vary in size from 1 to 400 kb(Dittwald, Gambin, Gonzaga-Jauregui, et al., 2013; Vollger et al., 2022). They contribute to the dynamic nature of the human genome.

Repetitive elements, including transposable elements, simple repeats, and segmental duplications, make up approximately 45% of the human genome, while SDs account for

6-8%(Hoyt et al., 2022; Vollger et al., 2022). These sequences possess important characteristics such as size, identity, orientation, and distribution, which make them significant contributors to mutational processes related to DNA recombination, replication, and repair. They have diverse modes of propagation, ranging from simple insertion events to facilitating non-allelic recombination while promoting genomic diversity(Emanuel & Shaikh, 2001).

In summary, the repetitive landscape of the human genome is intricate and dynamic, consisting of various repetitive elements dispersed throughout. These sequences play crucial roles in genome evolution, gene regulation, and disease mechanisms.

1.1 Repetitive Element Composition in Human Genome Assemblies

Since the completion of the initial draft of the human reference genome assembly, ongoing bioinformatic analyses have led to the discovery and improvement of genes, regulatory elements, enhancers, sequence motifs, and repeated elements. The focus of the reference assembly was primarily on the euchromatic regions of the human genome, which typically lack repeated sequences. However, despite significant progress in the past two decades, the reference assembly still contains gaps, particularly in regions that are enriched with repetitive elements such as tandem repeats, interspersed repeats, and segmental duplications(Bailey et al., 2001; Eichler, 2001). Mapping these regions accurately has proven to be challenging.

The advent of long-read genomic sequencing technologies in recent years has greatly enhanced the efficiency of whole genome sequencing and assembly. The Telomere-to-Telomere Consortium has made a recent breakthrough by adding and assembling 8% of the human genome that had previously eluded the reference assembly. This achievement has resulted in an alternate human genome assembly, known as T2T-CHM13v2 (T2T-CHM13), which is gapless(Nurk et al., 2022).

The newly characterized regions in the T2T-CHM13 assembly provide a more comprehensive understanding of the structure and organization of repetitive regions in the human genome. These regions predominantly consist of tandemly arrayed repeats,

segmental duplications, and complex repeats in pericentromeric and subtelomeric regions. Furthermore, the T2T-CHM13 assembly offers sequence and context for the ribosomal RNA gene clusters located in the short arms of acrocentric chromosomes 13, 14, 15, 21, and 22(Hoyt et al., 2022; International Human Genome Sequencing Consortium, 2004). These regions were previously unassembled in the GRCh37 and GRCh38 human genome reference assemblies, and their resolution represents a significant improvement(Lander et al., 2001).

1.2 Genomic disorders.

Genomic disorders are a group of diseases that result from DNA rearrangements in the human genome caused by the inherited genomic instability and mutability of our genome facilitated by the presence of repeat sequences, such as low copy repeats (LCRs) or segmental duplications (SDs), as well as by the presence of repetitive sequences such as short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs)(Carvalho et al., 2010; Stankiewicz & Lupski, 2002). In contrast to single base mutation or indels, genomic rearrangements result in the formation of structural variants including duplications, deletions, inversion, or even translocations.

The initial studies related to the analyses of chromosomal abnormalities that convey in severe phenotypes reveal the 17p12 as a region susceptible to rearrangements encompassing several dosage-gene and rich in both direct and inverted LCR flanking the region(Carvalho et al., 2010; Gonzaga-Jauregui & Lupski, 2021). The first genomic disorders reported in the literature are related to this region such as Charcot–Marie–Tooth disease type 1A (CMT1A), hereditary neuropathy with liability to pressure palsies (HNPP), Smith–Magenis microdeletion syndrome (SMS) and Potocki–Lupski microduplication syndrome (PTLS). The high frequency of interspersed LCR predisposes the region to genomic instability leading to rearrangements mediated mainly by NAHR(Cardoso et al., 2016). Consequently, the region undergoes structural variant formation including duplication, deletions, and inversion affecting one or more genes associated with the aforementioned genomic disorders

CMT1A is a sensorineural peripheral polyneuropathy caused by duplications encompassing the gene *PMP22*, while HNNP is a milder condition with susceptibility to neuropathy that can be caused by deletion of the gene *PMP22.* Although copy number variants (CNV) affect the dosage-sensitive gene, a point mutation in PMP22 can also convey in phenotypes related to HNPP or CMT1A(Zhang et al., 2010). Additionally, SMS is a disorder characterized by developmental delay, cognitive impairment, and behavioral abnormalities caused by genomic deletion of *RAI1,* similarly, PTLS shows overlapping clinical features including features of autism, but it is caused by genomic duplications of *RAI1.* As with *PMP22*, point mutations have also been described in *RAI1* leading to gain of function in patients with PTLS or implicating haploinsufficiency in patients with SMS(Vissers et al., 2007).

1.3 Genomic rearrangements

These sequences play major roles in shaping human genomes, promoting polymorphism, and contributing to genomic instability through genomic rearrangements leading to population diversity and genetic disease. Genomic rearrangements describe mutational changes in the genome such as duplication, deletion, insertion, inversion, and translocation(Currall et al., 2013; Stankiewicz & Lupski, 2002). Genomic rearrangements can represent polymorphisms that are neutral in function, or they can also convey phenotypes via diverse mechanisms, including changing the copy number variation (CNV)of dosage-sensitive genes, disrupting genes, creating fusion genes, or other mechanisms.

Characterization of many genomic rearrangements causative of human diseases revealed two rearrangement types that could be distinguished at a given locus: recurrent and nonrecurrent rearrangements.

Genomic rearrangements can be categorized into two major groups: recurrent and nonrecurrent rearrangements.Nonallelic homologous recombination (NAHR) between paralogous sequence repeats is the predominant mechanism underlying recurrent rearrangements with clustered breakpoints, whereas various mechanisms are

implicated in nonrecurrent rearrangements with variable breakpoints(Stankiewicz & Lupski, 2002).

Thus, two general types of genomic rearrangement, recurrent and nonrecurrent, are observed and show intrinsically distinct features that reflect their underlying mechanisms of formation. While recurrent rearrangement involves the same size and genomic content in unrelated individuals, nonrecurrent rearrangements have a unique size and genomic content at a given locus in unrelated individuals. These rearrangements are mediated by highly identical and long repeats such as LCR, although there are cases, where rearrangements are mediated by short and less identical, repeats such as Alus(Ade et al., 2013). The nonrecurrent SV breakpoints are characterized by blunt ends or 133 bp microhomologies,

Overall, there are major mechanisms that have been proposed for genomic rearrangements in the human genome such as nonallelic homologous recombination (NAHR), microhomology-mediated break-induced replication (MMBIR) and the Fork Stalling and Template Switching (FoSTeS) each one using specific repeat sequences as the substrate for recombination based on the inheritance principle of each mechanism(Carvalho & Lupski, 2016).

For NAHR, SD and LCR serve as the ideal substrate for recombination as the properties of these sequences such as their high degree of sequence identity and the presence of large flanking repeats predispose these regions to misalignments and the subsequent crossover between can result in recurrent rearrangement resulting in deletion, duplication, or inversion of the intervening sequence. Recurrent structural variants often result from NAHR between directly-oriented or inverted LCRs that flank unique sequence genomic regions(Liu et al., 2012). Most NAHR events that are causative of genomic disorders result from crossovers between LCRs located on the same chromosome, that is, intrachromosomal NAHR, or between non-allelic LCRs located in homologous chromosomes, that is, interchromosomal NAHR. although in some instances NARH process can be driven by tandem repeats(Dittwald, Gambin, Szafranski, et al., 2013).

Even though NARH is the most common mechanism for genomic rearrangements, there are certain genomic regions prone to instability driven by this mechanism due to the presence of large and highly identical sequences such as SD and LCR predisposing these regions to recurrent rearrangements(Shaw, 2004). In addition, the frequency in which this mechanism occurs is driven by the local genomic architecture depending on the distance, identity, and size between the couple of LCR acting as the substrate for recombination.

The major mechanism underlying the recurrent rearrangements is nonallelic homologous recombination (NAHR); however, the mechanism(s) for nonrecurrent rearrangements are less well established. Nonhomologous end joining (NHEJ) is a candidate recombination-based mechanism to explain some nonrecurrent rearrangements.Although NARH is the most well know, other mechanisms play key roles in the formation of CNV through nonrecurrent rearrangements such as MMMBIR or FoSTes in which the microhomology is used as a primer to assist the rearrangement.

The crossovers between directly oriented or inverted LCRs that flank distinctive genomic regions driven by NAHR events generally result in recurrent rearrangements that are responsible for most of the well-known and characterized genomic disorders. In contrast, non-recurrent genomic rearrangements which can include complex rearrangements produced by microhomology-mediated mechanisms (MMBIR/FoSTeS) have been identified in other less common disorders.

Overall, recurrent rearrangements are dominated by genomic architectural features of LCR or SD as act as substrates for an NAHR recombination event, while the role of genomic architectural features of nonrecurrent rearrangements is still unclear as the potential mechanisms involved in the formation of nonrecurrent CNVs are caused by a variety of microhomology-mediated mechanisms such as MMEJ, FoSTeS, MMBIR, SRS, and BISRS using different sequences as the substrate for recombination including Alus, L1, and tandem repeats elements(Carvalho & Lupski, 2016). Finally, the genomic architecture might stimulate the formation of these rearrangements by increasing the susceptibility for DNA breakage or promoting replication fork stalling.

1.4 Molecular mechanisms for human genomic rearrangements

There are several mechanisms involved in the formation of structural variants, however the most common ones and best characterized are Non-Homologous End Joining (NHEJ), Non Allelic homologous recombinations (NAHR), microhomology-mediated break induced replication (MMBIR) and Fork stalling template switching (FoSTeS). NAHR, NHEJ, MMBIR, and FoSTeS probably account for the majority of genomic rearrangements in our genome and the frequency distribution of the four at a given locus may partially reflect the genomic architecture of repetitive elements(Carvalho & Lupski, 2016).

NAHR is the most common mechanism-driven recurrent rearrangement and most of these events result in genomic disorders from crossovers between LCR. This mechanism cause interchromosomal and intrachromosomal deletions, duplications, and inversions depending on variables such as size, orientation, the degree of identity, and the distance between low-copy repeats (LCRs) or other repetitive elements that serve as substrates.

The crossovers between directly oriented or inverted LCRs that flank distinctive genomic regions driven by NAHR events generally result in recurrent rearrangements that are responsible for most of the well-known and characterized genomic disorders. The crossover can occur at different levels including intrachromatid recombination between direct-oriented LCRs resulting in deletions; interchromatid recombination leading to deletions and duplications; whereas intrachromosomal inverted repeats can lead to sequence inversions.

The presence of highly identical and long flanking repeats such as LCRs or SDs predispose genomic regions to instability favoring the formation of recurrent rearrangements mediated by NAHR.

Although NAHR is commonly associated as the main mechanism related to classical genomic disorders, nonrecurrent rearrangements have also been linked to dozens of genomic disease cases.

While recurrent rearrangements are primarily generated by recombination-associated mechanisms such as NAHR, non-recurrent rearrangements are caused by a wide variety of mechanisms such as nonhomologous repair or recombination processes, including nonhomologous end joining (NHEJ) and microhomologymediated end joining (MMEJ), as well as replication-based mechanisms, including break-induced replication (BIR), microhomology-mediated break-induced replication (MMBIR), fork stalling and template switching (FoSTeS), and serial replication slippage (SRS)(Lee et al., 2007).

NHEJ can cause some basic non-recurrent rearrangements. NHEJ accounts for one of the major mechanisms responsible for the repair of double-strand breaks (DSB) that can result in small deletions or insertions. This mechanism can be divided into 4 essential steps, which are as follows: detection of DSB, followed by the molecular bridging of both broken DNA ends, then modification of the ends to make them compatible and ligatable; and finally the ligation step. For MMEJ uses small lengths of microhomology to recombine in order to repair DSBs(Weterings & Chen, 2008).

As these mechanisms do not require from homologous template for the repair of the DNA, different breakpoints have been associated with different repetitive elements such as LTR, LINE and Alu. Both NHEJ and MMEJ involve the alignment of microhomologous sequences internal to the broken ends before joining and are associated with deletions and insertions, however, NHEJ strictly requires no homology or only 1-4 bp of homology at the junction, while MMEJ need 1-6 bp of homology to the ends to align them for repair.

NHEJ can result in accurate blunt breakpoints, leading to small deletions (14 bp) or insertions of free floating DNA; whereas MMEJ invariably leads to deletion of sequences between annealed microhomologies.

Importantly, recombination based repair mechanisms, either homologous (NAHR) or non-homologous (NHEJ, MMEJ), aim to repair double-strand breaks in the DNA;

whereas replication-based mechanisms (such as MMBIR or FoSTeS) repair single-ended, double-stranded DNA (seDNA) breaks that may result from collapsed forks or at chromosome telomeres

In contrast to non-homologous recombination mechanisms, MMBIR and FoSTeS have been proposed as the major contributors to the generation of nonrecurrent rearrangements in human genome disorders coupled with structural variants associate to genomic disorder.

According to FoSTeS, this model implies template switching which refers to a change of the single-stranded DNA template during replication within the same replication fork or between distinct replication forks, consequently when the DNA replication fork stalls at one position, the lagging strand disengages from the original template, transfers and then anneals to another replication fork in physical proximity, and restarts the DNA synthesis. The invasion and annealing depend on the microhomology between the invaded site and the original site. Switching to another fork located downstream (forward invasion) would result in a deletion, whereas switching to a fork located upstream (backward invasion) results in a duplication. Depending on whether the lagging or leading strand in the new fork was invaded and copied, and the direction of the fork progression, the erroneously incorporated fragment from the new replication fork would be in a direct or inverted orientation to its original position.This procedure of disengaging, invading/ annealing and synthesis/extension could occur multiple times in series caused by the poor processivity of the involved DNA polymerase, and causing the observed complex rearrangements(Zhang et al., 2009).

In a similar manner, MMBIR acts during the replication stage, but its mechanism is based on the principle guiding break-induced replication (BIR). In BIR, when the replicative helicase encounters a nick on the template strand one arm of a replication fork breaks off promoting homologous recombination mediated by Rad51 that repairs the broken or collapsed replication forks, however, if the template used for repair involves a homologous or paralogous sequence in a different chromosome position it can produce deletion, duplication, and translocation events. While for MMBIR instead of

requiring RecA/Rad51 for reparation, uses microhomology (1-4bp) to resume a stalled or collapsed replication fork as opposed to the longer homology tracts that are used in BIR(Hastings et al., 2009).

Importantly, recombination based repair mechanisms, either homologous (NAHR) or non-homologous (NHEJ, MMEJ), aim to repair double-strand breaks in the DNA; whereas replication-based mechanisms (such as MMBIR) repair single-ended, double-stranded DNA (seDNA) breaks that may result from collapsed forks or at chromosome telomeres.

1.5 The Role of genomic architecture

Importantly, highly identical sequences such as SD and LCR provide the ideal substrates for recombination promoting genomic instability and creating hotspots for recurrent rearrangements by NAHR. The frequency at which NAHR events occur at a given locus is determined by several factors related to the structure and features of homologous sequences(Liu et al., 2011). The genome-wide frequency of NAHR is positively associated with flanking LCRs length but inversely influenced by the distance between the LCRs. In summary, NAHR is the predominant mechanism for recurrent rearrangements promoted by nearby LCRs and different pairs can be utilized. However, LCR length favorably affects the probability of using a particular LCR pairs, while the distance between repeats may have an adverse effect(Shaw, 2004; Stankiewicz & Lupski, 2002).

In contrast to the genomic architectural features underlying recurrent rearrangements, the genomic architectural landscape mediating non-recurrent rearrangements is broad as different molecular mechanisms are involved in the formation of nonrecurrent structural variants. The presence of specific genomic structures, such as repetitive sequences and repeated elements can stimulate the occurrence of template switching or promote the formation of non-B DNA structures through A-T rich palindromes, G-quadruplexes, short inverted repeats, and retrotransposable elements. These events can lead to the collapse of replication forks and, in some cases, the creation of

double-strand breaks (DSBs). DSBs can be a primary factor contributing to instability in a specific genomic region, although it is also possible that genomic instability can occur without the necessity of DSBs.Nevertheless, secondary DNA structures, such as hairpin loops formed by inverted repeats, have the potential to expose single-stranded sequences. Moreover, these hairpin structures can enhance the chances of replication fork stalling, which in turn can trigger a non-recurrent replication-based mechanism. These secondary DNA structures play significant roles in these processes(Makova & Weissensteiner, 2023).

Understanding these features of repetitive elements provides valuable insights into the mechanisms by which they contribute to genomic rearrangements. The length, abundance, sequence identity, orientation, involvement in replication-based processes, formation of non-B DNA structures, retrotransposition ability, genomic proximity, and epigenetic regulation collectively shape the impact of repetitive elements on genomic rearrangements.

2. Results

2.1 Previous research contributions

Previous studies have identified repetitive elements that act as templates for various mechanisms involved in genomic rearrangements. The initial study focused on the identical repeat backbone of the human genome, specifically exploring the minimum requirements for non-allelic homologous recombination (NAHR) events(Zepeda-Mendoza et al., 2010). NAHR requires sequences with a high level of identity and a minimum length of approximately 300 base pairs (bp). However, this study solely focused on 100% identity, thereby limiting the identification of other potential repetitive elements involved in NAHR. The dataset generated in the study comprises around 2% of the total reference human genome and includes potential recombinotion sites which overlap important functional and structural elements such as SDs, including transposon-derived repeats, processed pseudogenes, simple sequence repeats, and blocks of tandemly repeated sequences and genes. In contrast, the other study took a broader approach and investigated the distribution of Inverse and direct paralogous low-copy repeats larger than 1 kB, with over 95% sequence identity, throughout the entire genome(Dittwald, Gambin, Gonzaga-Jauregui, et al., 2013; Dittwald, Gambin, Szafranski, et al., 2013).

Unlike the previous studies that focused solely on repetitive elements that could potentially facilitate NAHR, our research takes a different approach. We introduce a novel set of parameters that allow us to investigate and identify repeat elements not only involved in NAHR but also other mechanisms like MMBIR or FoSTeS.We applied these parameters to analyze three human genome assemblies, aiming to compare the differences and compositions of the repeated elements identified in each assembly

2.2 Genome-Wide landscape of identical repeats in the human genome

In my bioinformatic analyses, I obtained datasets for direct and inverted intrachromosomal repeats for each of the current human genome assemblies, namely GRCh37 (hg19), GRCh38 (hg38), and the most recent telomere-to-telomere alternate assembly (T2T-CHM13). This selection of assemblies allowed the comparison and assess the variations in repeat sequences across different genome references.

In the GRCh37, GRCh38, and T2T-CHM13 assemblies, I identified a total of 570,829, 573,085, and 585,604 repeated sequence pairs in direct orientation, respectively. Similarly, for inverted repeats, 611,838, 612,089, and 627,791 pairs in each of the assemblies analyzed were found. These numbers provide insight into the abundance of repeated sequences and highlight the importance of studying their characteristics.

To determine the similarities between the repeat pairs, I utilized percent identity as a measure, considering pairs with a range of 80% to 100% similarity for both direct and inverted repeats across the assemblies based on the chosen identity parameters related to genomic rearrangement mediation. Additionally, I observed that the size of the identified repeats varied significantly, ranging from the minimum parameter of 200 base pairs to several million base pairs, across all three assemblies and orientations. Notably, a majority of the repeats, approximately 62-64% in the different assemblies, had a length below 1 kilobase (kb) (Table 1, Figure 1 D). However, depending on the assembly and its orientation, around 1% of the repeats exceeded a size of 6 kb or 10 kb (Table 1).

	GRCh37			GRCh38			T2T-CHM13		
Repeats	**1% repeats**	**Base pairs**	**Percent**	**1% repeats**	**Base pairs**	**Percent**	**1% repeats**	**Base pairs**	**Percent**
Direct	>6264 bp	320525555	10.35%	>6360 bp	328261336	10.63%	>10038 bp	391552384	12.56%
Inverted	>6240 bp	328497932	10.61%	>6296 bp	336565558	10.90%	>8164 bp	365728379	11.73%
Both Non-overlapping bp		406107270	13.12%		416913534	13.50%		487539841	15.64%

Table 1. **Percentage of base pairs covered by repeat elements.** It is reported the base pairs(bp) covered by direct and inverted repeats as well as the non-overlapping bp covered by both type of repeats ans its representative percentages in each one of the assemblies.

The median percent identity for all the identified repeat pairs, considering both direct and inverted orientations in the three assemblies, was approximately 84.3%. This statistic provides an overall understanding of the level of similarity observed among the repeats. Furthermore, the distance between pairs of repeat elements had a median value of 30-31 megabase pairs (Mbp) across the assemblies (Figure 1, Table 2). These findings shed light on the genomic organization of repeats and their distribution within the genome.

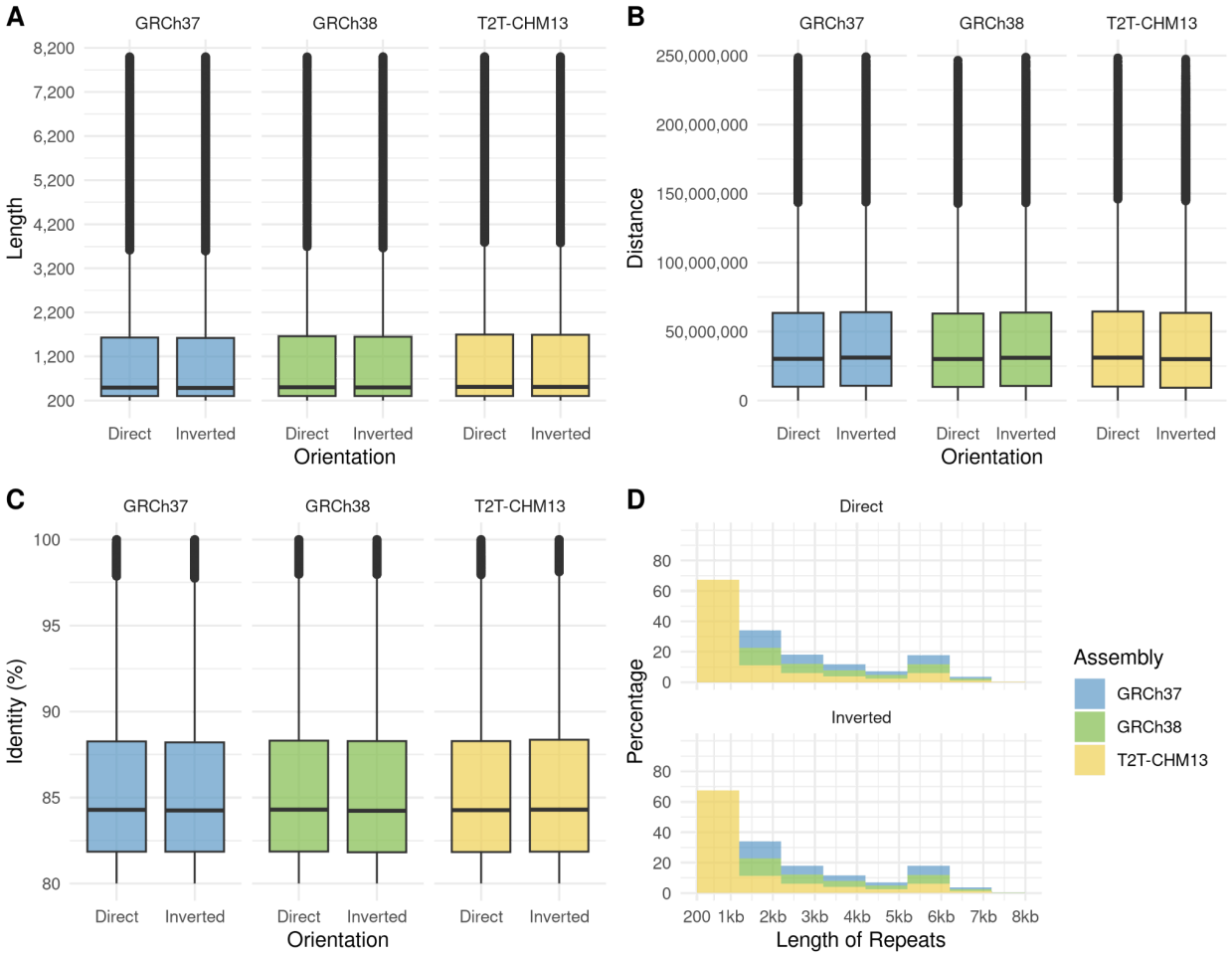


Figure 1. Statistics for identified direct and inverted repeats across the three assemblies (GRCh37, GRCh38, and T2T-CHM13). The overall statistics for the identified repeats are very similar across the three assemblies. A)Length distribution. B) Distance distribution. C) Pairwise percent identity distribution. D) Size distribution and percentage of repeats in size bins.

		Direct			Inverted		
		GRCh37	**GRCh38**	**T2T-CHM13**	**GRCh37**	**GRCh38**	**T2T-CHM13**
Total repeat pairs (N)		570,829	573,085	585,604	611,838	612,089	627,791
Total repeat pairs (bp,%)		320525555, (10.35%)	328261336, (10.63%)	391552384, (12.56%)	328497932, (10.61%)	336565558, (10.90%)	365728379, (11.73%)
Length (bp)	*Min-Max*	200-395595	200-395596	200-1663487	200-647494	200-647495	200-495481
	Median	500	509	525	469	505	520
	Mean	1450	1503	2263	1455	1483	1612
	<1kb (N, %)	367079 (64.31%)	364350 (63.58%)	367493 (62.75%)	394368 (64.46%)	390579 (63.81%)	395079 (62.93%)
Identity (%)	*Min-Max*	80-100	80-100	80-100	80-100	80-100	80-100
	Median	84.29	84.3	84.3	84.25	84.23	84.27
	Mean	85.48	85.5	85.52	85.45	85.46	85.47
Distance (bp)	*Min-Max*	-14-248756443	-8-246623862	-8-247651277	-9995-249229690	-2-248935491	163-248381454
	Median	30263091	30098648	30016196	31119448	30978896	31094072
	Mean	42739049	42431293	42353751	43275746	43075283	43149712

Table 2. Overall stats of direct and inverted repeats - The statistics for the identified direct and inverted repeats show remarkable similarities across all three assemblies.

In terms of the distribution patterns of direct and inverted repeats across chromosomes, I observed a similar genome-wide distribution for the three assemblies, with one exception being the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22) in the T2T-CHM13 assembly (Figure 2, Figure 3). This disparity can be justified by considering the availability of new genomic sequence data in the T2T-CHM13 assembly specifically for these chromosomal arms, which differs from the reference assemblies. Consequently, the improved resolution provided by the T2T-CHM13 assembly enabled the detection of repeated sequences in these regions that were previously unidentified. This finding highlights the importance of the T2T-CHM13 assembly in uncovering previously unknown repeat elements in these specific chromosomal arms.

Furthermore, we observed an increased number of repeat elements in chromosome Y of the T2T-CHM13 assembly, which can be attributed to the enhanced resolution provided by long-read sequencing techniques. The improved resolution facilitated better

characterization of the repetitive nature of chromosome Y, leading to the identification of a higher number of repeat elements compared to the reference assemblies. This highlights the advantage of utilizing long-read sequencing in capturing the intricate repetitive structures of chromosome Y.

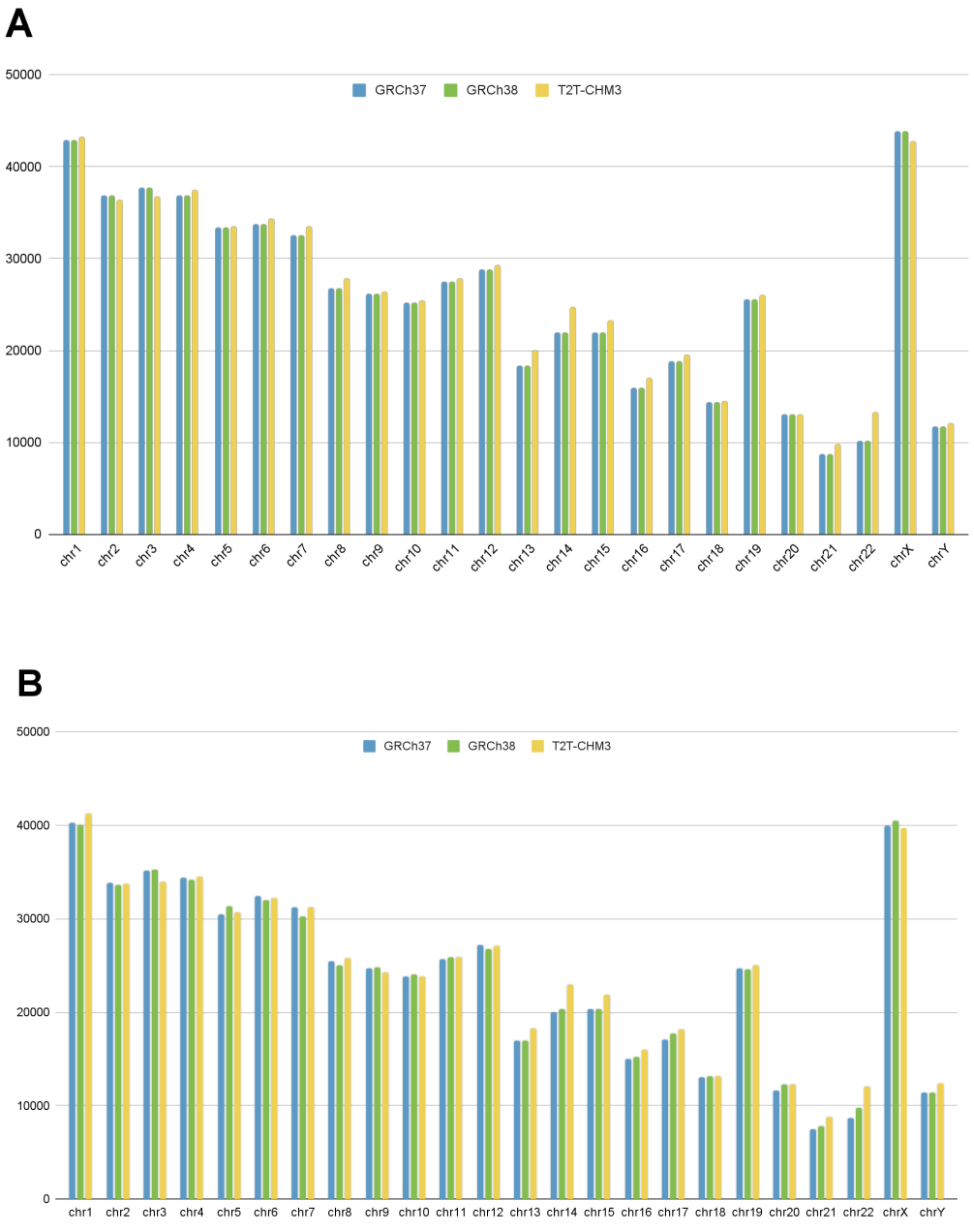
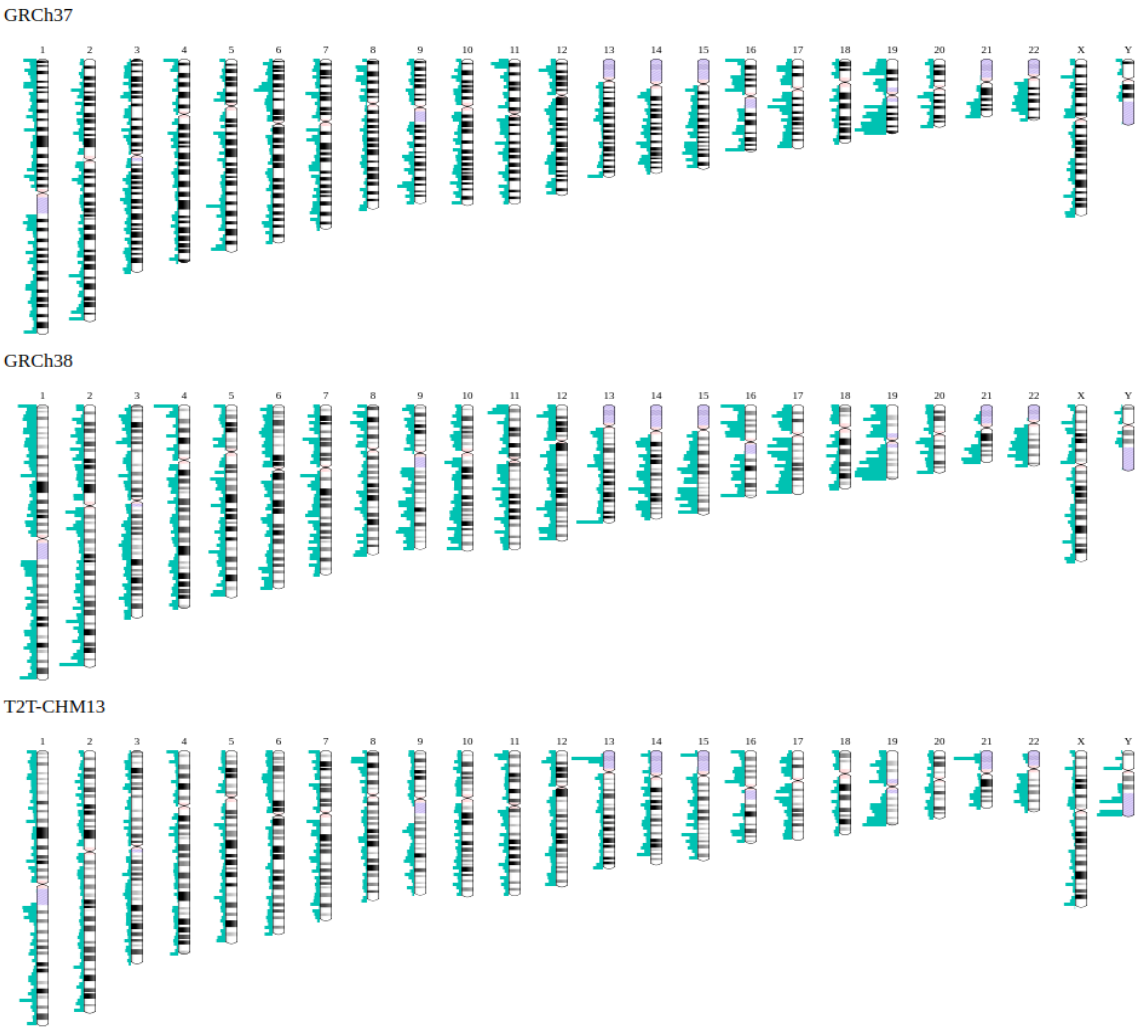


Figure 2. Per chromosome distribution of direct and inverted repeats across the genome in the analyzed assemblies (GRCh37, GRCh38, and T2T-CHM13). The overall distribution of direct and inverted repeats across the genome was observed to be very similar for the three assemblies we studied except for an increased number of repeats detected in the acrocentric chromosomes (13, 14, 15, 21 and 22) of the T2T-CHM13 assembly. A) Per chromosome distribution of direct repeats across assemblies. B) Per chromosome distribution of inverted repeats across assemblies.



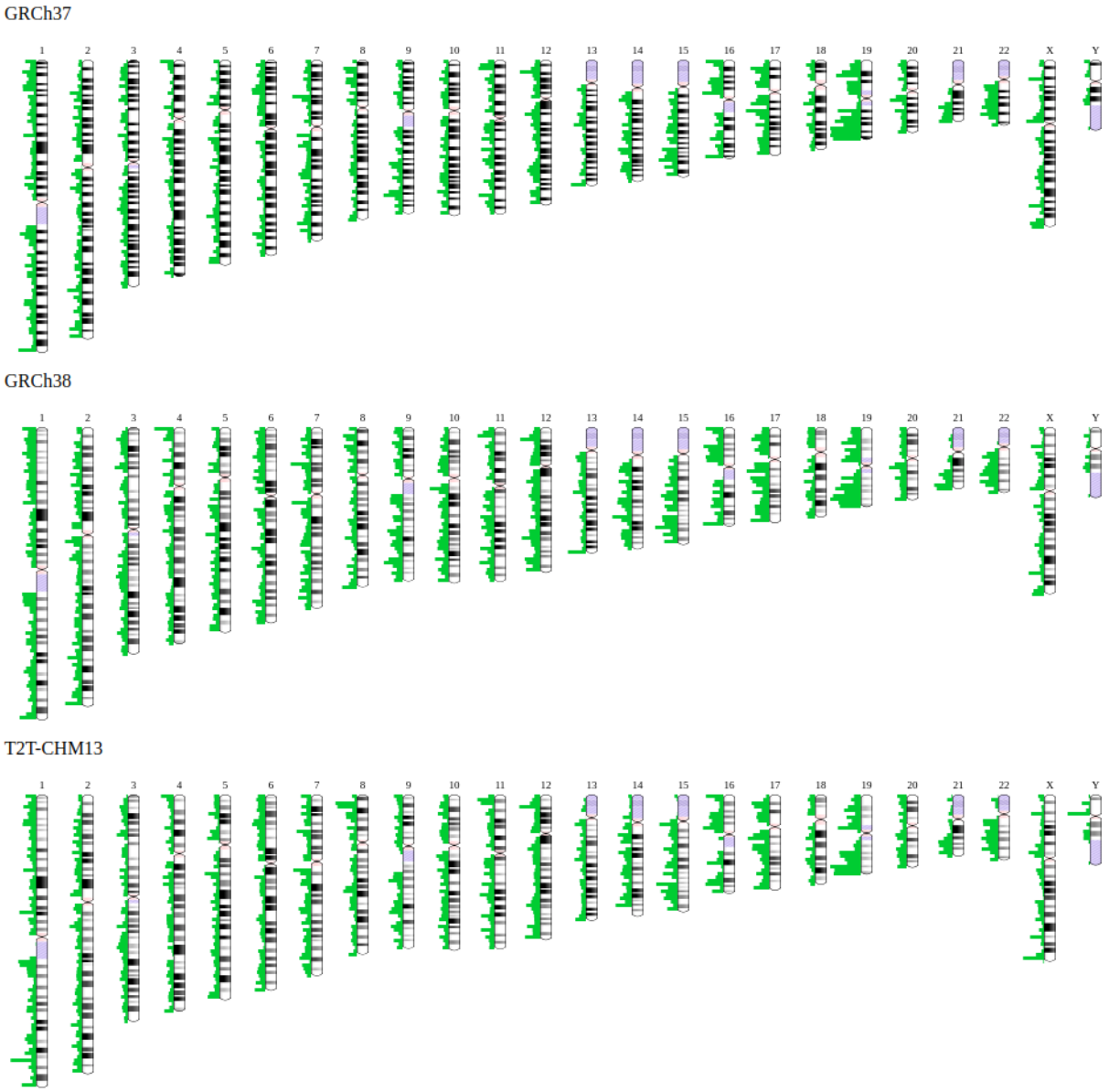


Figure 3. Ideogram of the human chromosomes showing the distribution of direct and inverted repeats identified in this study across three human genome assemblies analyzed (GRCh37, GRCh38, and T2T-CHM13). A) Genome-wide distribution of direct repeats across human chromosomes. B) Genome-wide distribution of inverted repeats across human chromosomes. Note the representation of repeat elements in the short arms of acrocentric chromosomes 13, 14, 15, 21 and 22 in the T2T-CHM13 assembly versus the two reference

assembly versions for both direct and inverted repeats, and more representation of inverted repeats in the Y chromosome of T2T-CHM13.

By presenting these comprehensive statistics and observations, highlighting the variations in repeated sequences among different human genome assemblies and emphasizing the significance of the T2T-CHM13 assembly in uncovering additional repeat elements in specific chromosomal regions.

2.3 Composition and annotation of direct and inverted repeats across the genome

It is essential to understand the genomic rearrangements and their association with human diseases as they can directly impact gene structure, dosage, and regulation. Recurrent genomic rearrangements often involve large and highly similar repetitive regions, such as low-copy repeats (LCRs) or segmental duplications (SDs), through non-allelic homologous recombination (NAHR). Smaller and more divergent repetitive elements like SINEs and LINEs can also contribute to genomic rearrangements through microhomology-mediated mechanisms (MMBIR/FoSTeS). For instance, Alu elements, a subclass of SINEs comprising 11% of the genome, have been implicated in Alu/Alu-mediated rearrangements.

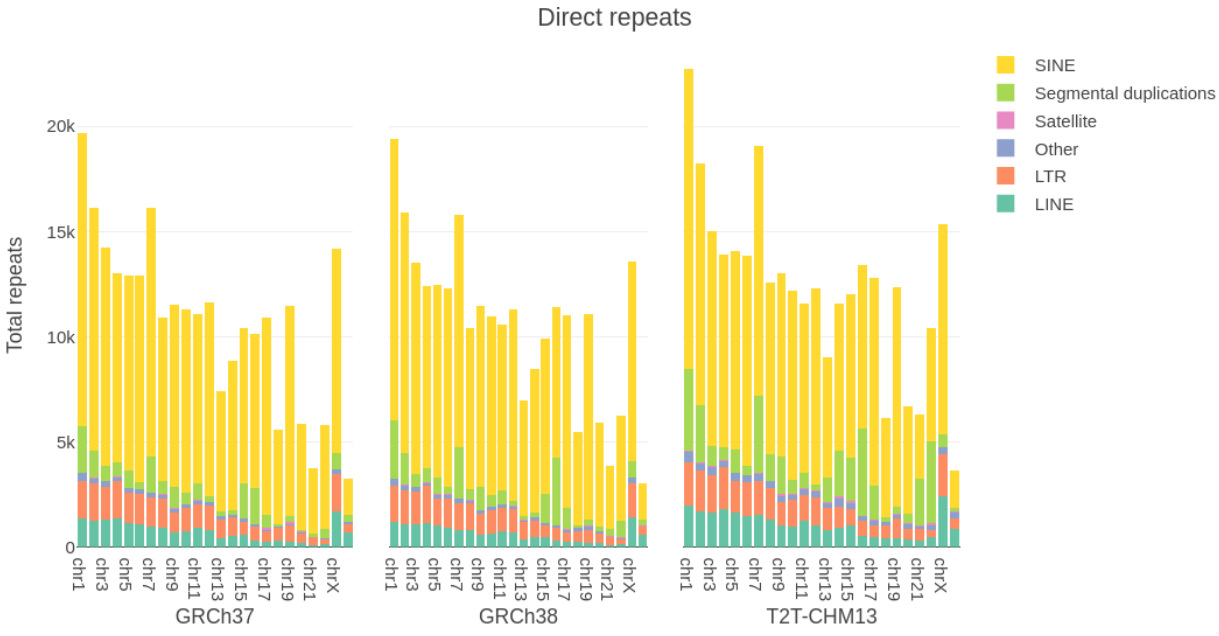
To gain further insights into the identified repeat datasets, I cross-referenced them with known repeated elements and protein-coding genes in the human genome. This allowed a better characterization of the nature of these repeats and their potential impact on genomic architecture.

In the T2T-CHM13 assembly, it was observed a larger proportion of repeat pairs (65,863 out of 83,512) overlapping segmental duplications compared to the reference assemblies (approximately 68.84% for GRCh37 and 63.79% for GRCh38). Other types of repeats, such as LINEs, SINEs, and Alus (Table 3), exhibited similar distribution patterns across the genome in all three assemblies. However, as expected, there were

variations in the newly assembled regions of T2T-CHM13 that were not present in the two reference assemblies (Figure 4).

	GRCh37		GRCh38		T2T-CHM13	
Repeat type	**Direct**	**Inverted**	**Direct**	**Inverted**	**Direct**	**Inverted**
SINE	194063	205439	189389	200233	198184	199981
LINE	18316	18193	15688	15566	26694	15505
LTR	25000	25579	23757	24377	28431	23977
SDs	17470	17046	20874	20099	37099	28764
Satellite	115	100	94	92	480	288
Other	4273	4302	3852	3815	7404	4418

Table 3. **Distribution Patterns of Repeat Elements across the Assemblies -** The presence of repetitive elements does not show a remarkable difference across the three assemblies, except for SD. In SD, a clear increase in frequency is observed in each of the assemblies.



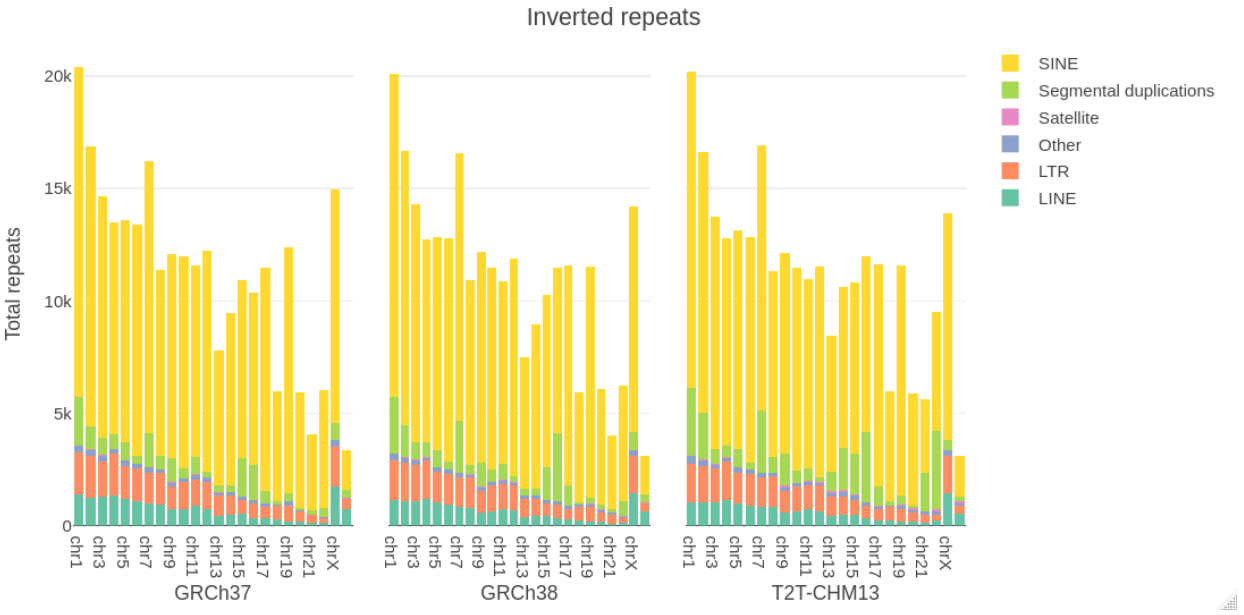


Figure 4. Overlap of identified repeated sequences with known repeat elements across human genome assemblies. A large fraction of repeat pairs was observed overlapping with segmental duplications, LINEs, and satellite repeats in T2T compared to the reference assemblies. Other types of repeats were similarly distributed across the genome in the different assemblies.

2.4 Gene overlap.

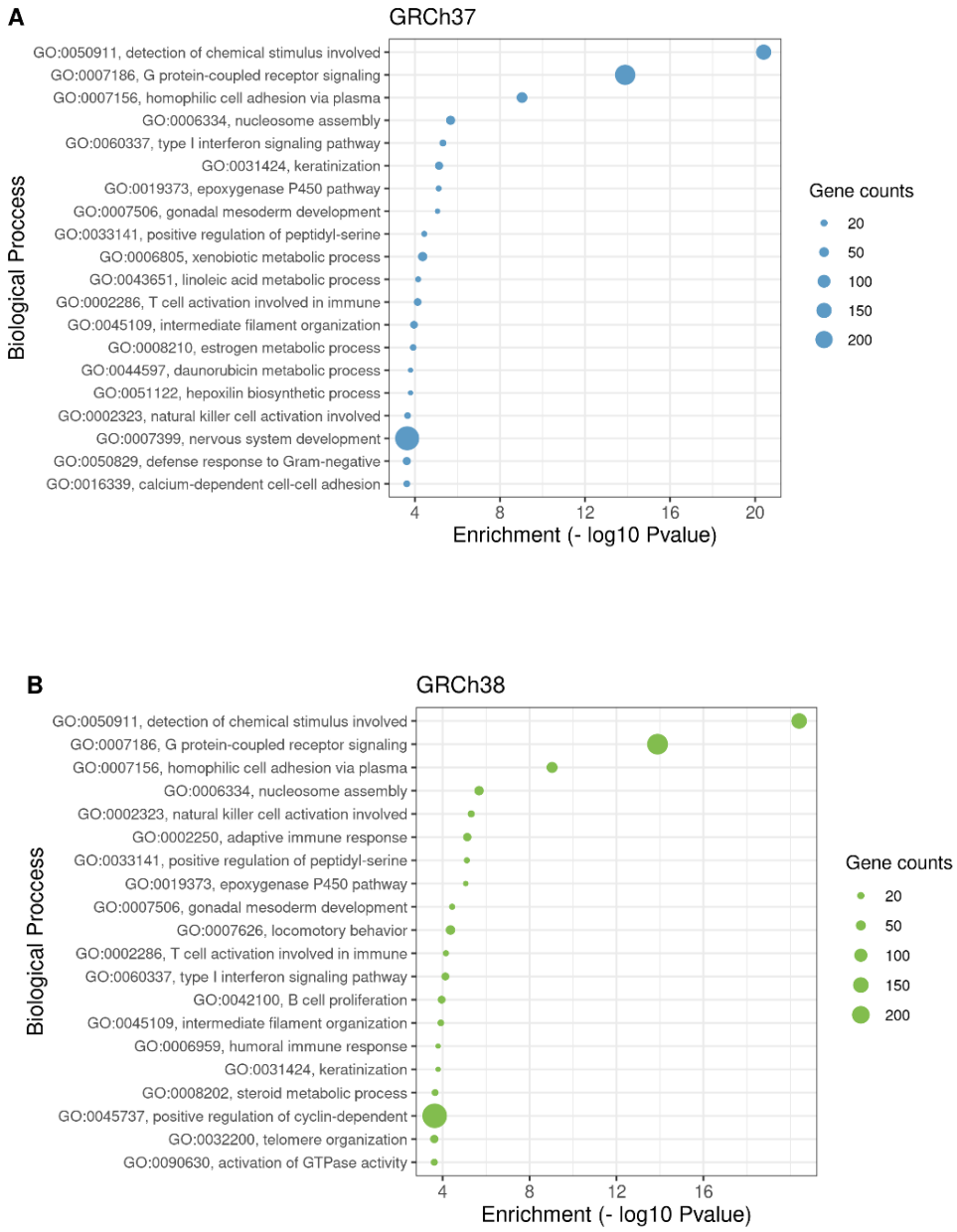
I observed that in the GRCh37, GRCh38, and T2T-CHM13 assemblies, there were 933, 970, and 914 protein-coding genes, respectively, that overlapped with direct repeats. Furthermore, there were 834, 872, and 847 genes, respectively, that overlapped with inverted repeats (Table 4). Similarly, within a range of 100kb upstream or downstream in the corresponding assemblies, there were 1872, 1803, and 1796 genes, respectively, that were flanked by direct repeat pairs. In addition, there were 1413, 1574, and 1428 genes, respectively, that were flanked by inverted repeat pairs (Table 4). Overall, a total of 3663, 3774, and 3652 nonredundant genes have the

potential to be affected by rearrangements of the repeats we have identified in the three assemblies.

	GRCh37		GRCh38		T2T-CHM13	
Protein coding genes	**Direct**	**Inverted**	**Direct**	**Inverted**	**Direct**	**Inverted**
Overlapped	933	834	970	872	914	847
Flanked	1872	1413	1803	1574	1796	1428
OMIM_genes						
Overlapped	46	33	49	36	47	35
Flanked	371	259	347	289	348	262

Table 4. **Annotation of Identified Genes -** This section presents the annotation of protein-coding genes, including those associated with diseases, which are found to be overlapped and flanked by direct and inverted repeats

To analyze the functional implications of the genes potentially affected by genomic rearrangements, I conducted gene ontology (GO) enrichment analysis. This analysis revealed that across all three assemblies, the main enriched classes were related to the olfactory system, including the detection of chemical stimulus (GO:0050911), sensory perception of smell (GO:0007608), and GPCR signaling (GO:0007186) (Figure 5). This finding aligns with the well-known genetic and functional variability observed in the olfactory receptor (OR) gene family and the role of copy number variations (CNVs) and genomic arrangements in inter-individual and cross-population variation. Additionally, besides the olfactory system, I identified enrichment for genes associated with immune response and metabolic processes among the top 10 GO terms across the three assemblies (Figure 5). Among the genes identified as potentially impacted by this repeat analysis, 709, 721, and 691 genes in GRCh37, GRCh38, and T2T-CHM13, respectively, are annotated as diseases associated in OMIM (Table 3). Some of these genes fulfill dosage-sensitive criteria and have been previously associated with genomic disorders, such as Bartter syndrome, Hajdu-Cheney syndrome, Ehlers-Danlos syndrome, and Usher syndrome.


A
GRCh37
GO:0050911, detection of chemical stimulus involved
GO:0007186, G protein-coupled receptor signaling
GO:0007156, homophilic cell adhesion via plasma
GO:0006334, nucleosome assembly
GO:0060337, type I interferon signaling pathway
GO:0031424, keratinization
GO:0019373, epoxygenase P450 pathway
GO:0007506, gonadal mesoderm development
GO:0033141, positive regulation of peptidyl-serine
GO:0006805, xenobiotic metabolic process
GO:0043651, linoleic acid metabolic process
GO:0002286, T cell activation involved in immune
GO:0045109, intermediate filament organization
GO:0008210, estrogen metabolic process
GO:0044597, daunorubicin metabolic process
GO:0051122, hepoxilin biosynthetic process
GO:0002323, natural killer cell activation involved
GO:0007399, nervous system development
GO:0050829, defense response to Gram-negative
GO:0016339, calcium-dependent cell-cell adhesion
Biological Proccess
4
8
12
16
20
Enrichment (- log10 Pvalue)
Gene counts
20
50
100
150
200

B
GRCh38
GO:0050911, detection of chemical stimulus involved
GO:0007186, G protein-coupled receptor signaling
GO:0007156, homophilic cell adhesion via plasma
GO:0006334, nucleosome assembly
GO:0002323, natural killer cell activation involved
GO:0002250, adaptive immune response
GO:0033141, positive regulation of peptidyl-serine
GO:0019373, epoxygenase P450 pathway
GO:0007506, gonadal mesoderm development
GO:0007626, locomotory behavior
GO:0002286, T cell activation involved in immune
GO:0060337, type I interferon signaling pathway
GO:0042100, B cell proliferation
GO:0045109, intermediate filament organization
GO:0006959, humoral immune response
GO:0031424, keratinization
GO:0008202, steroid metabolic process
GO:0045737, positive regulation of cyclin-dependent
GO:0032200, telomere organization
GO:0090630, activation of GTPase activity
Biological Proccess
4
8
12
16
Enrichment (- log10 Pvalue)
Gene counts
20
50
100
150
200

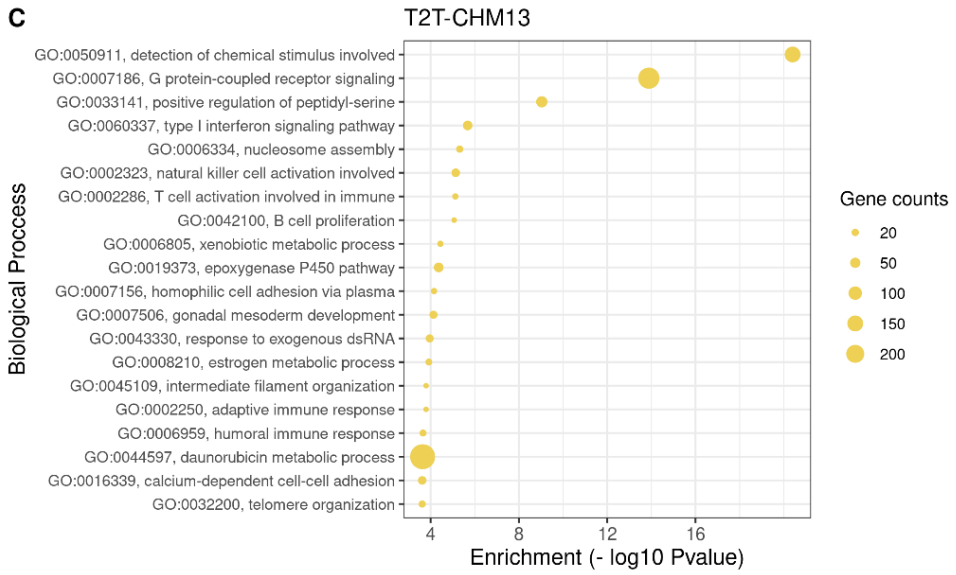


Figure 5. Gene ontology enrichment analysis. Our analysis showed that genes related to the olfactory system, G protein-coupled receptor signaling, and a few immune and metabolic processes were enriched in regions overlapped or flanked by our identified repeated sequences in the three genome assemblies analyzed. The size of the dot represents the number of genes contained in the gene set. A) GO analysis for genes found in GRCh37. B) GO analysis for genes found in GRCh38. C) GO analysis for genes found in T2T-CHM13.

2.5 Repeat overlap with genomic disorder regions and other reported structural variants

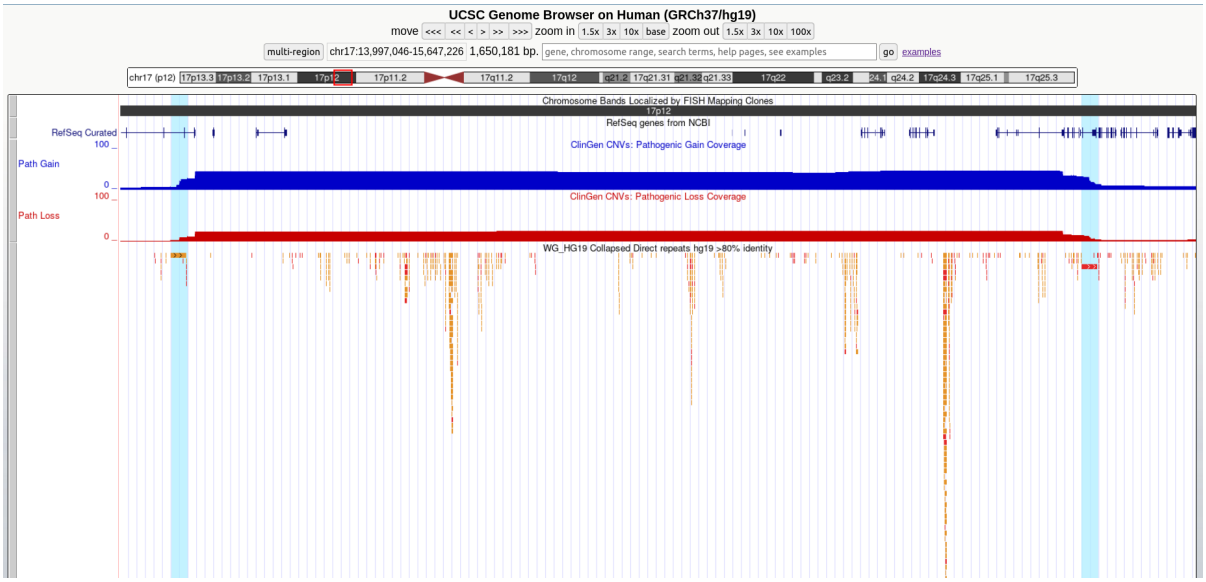
Genomic disorders have been studied over more than two decades to understand the role of repetitive DNA sequences in remodeling the genome through recombination events that can result in clinically recognizable human disorders. These events can be facilitated by the presence of repeated sequences, including low copy repeats (LCRs), segmental duplications (SDs), short interspersed nuclear elements (SINEs), and long interspersed nuclear elements (LINEs). Common chromosome deletion/duplication syndromes often involve rearranged genomic segments flanked by large LCR or SD structures that serve as recombination substrates. Overall, the instability and mutability of the genome are influenced by the presence of these repetitive sequences.

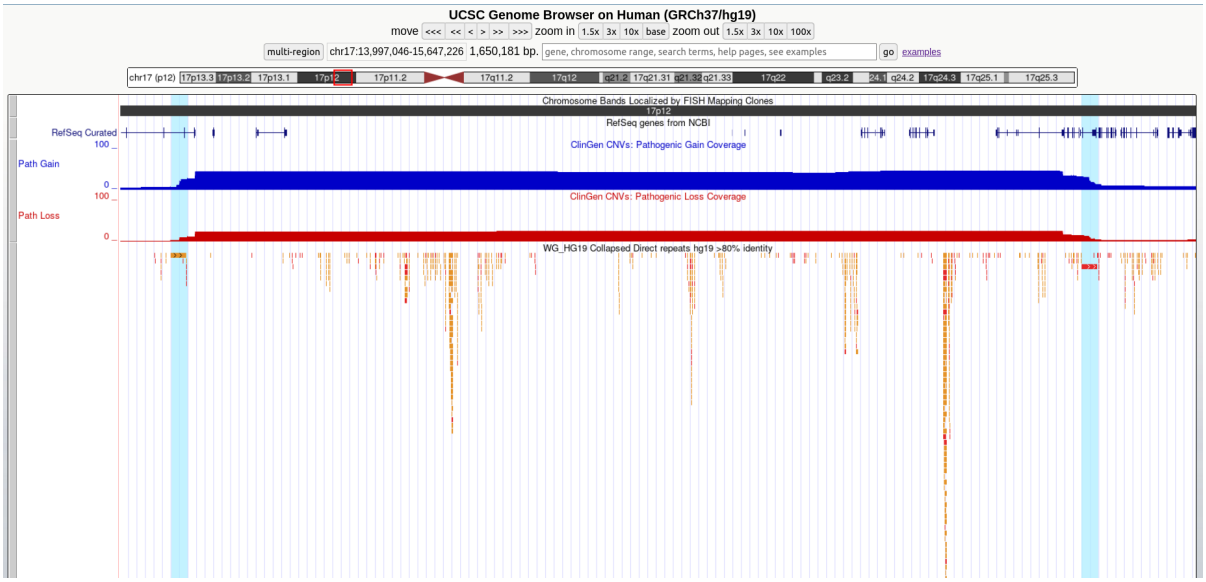
In order to evaluate the utility of the identified repeat pairs for the study of more common genomic disorders, I cross-referenced the repeats with regions known to be involved in the generation of recurrent and non-recurrent rearrangements. I observed clustering of repeats and the presence of larger repeat pairs flanking the reported deletion/duplication CNVs, consistent with known LCRs mediating recombination events in these regions (**Table 5, Figure 6**). Additionally, while looking more broadly at the potential contribution of the repeats to the generation of reported CNVs and structural variants I observed a higher frequency of deletions flanked by direct repeats compared to duplications in all three assemblies (**Table 6**) while for inversion I detected a few of them flanked by inverted repeats. I also investigated CNVs flanked by repeats on only one side, as examples have been reported of some non-recurrent rearrangement breakpoints clustering within individual repeated elements.

For inverted repeats, I looked at the overlap of our repeated elements with previously reported inversions mediated by mobile element insertions (MEIs) including LINE1 and Alu elements and SDs. Of the 65 reported MEI mediated inversions, we found 49 overlapped by our repeats, whilst 194 of 207 of the reported SD mediated inversions overlapped (**Figure 7**).

CNV	Flanking both sides			Flanking one side			Overlapping the CNV		
	GCRh37	GCRh38	T2T-CHM13	GCRh37	GCRh38	T2T-CHM13	GCRh37	GCRh38	T2T-CHM13
Deletions	4294	6779	4082	56023	55515	34460	10296	9817	7748
Duplication	1520	2430	1047	16009	15745	6948	3301	2880	1537
Inversions	373	410	219	1812	1461	907	268	246	270

Table 5. **Copy Number Variants Detected Overlapping or Flanked by Repeats -** The analysis reveals that duplications and deletions exhibit a higher frequency of overlapping or flanking with direct or inverted repeats compared to inversions.


UCSC Genome Browser on Human (GRCh37/hg19)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region chr17:13,997,046-15,647,226 1,650,181 bp. gene, chromosome range, search terms, help pages, see examples go examples
chr17 (p12) 17p13.3 17p13.2 17p13.1 17p12 17p11.2 17q11.2 17q12 q21.2 17q21.31 q21.32 q21.33 17q22 q23.2 q24.1 q24.2 17q24.3 17q25.1 17q25.3
Chromosome Bands Localized by FISH Mapping Clones
17p12
RefSeq genes from NCBI
RefSeq Curated
100
ClinGen CNVs: Pathogenic Gain Coverage
Path Gain
0
100
ClinGen CNVs: Pathogenic Loss Coverage
Path Loss
0
WG_HG19 Collapsed Direct repeats hg19 >80% identity


UCSC Genome Browser on Human (GRCh37/hg19)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region chr17:13,997,046-15,647,226 1,650,181 bp. gene, chromosome range, search terms, help pages, see examples go examples
chr17 (p12) 17p13.3 17p13.2 17p13.1 17p12 17p11.2 17q11.2 17q12 q21.2 17q21.31 q21.32 q21.33 17q22 q23.2 q24.1 q24.2 17q24.3 17q25.1 17q25.3
Chromosome Bands Localized by FISH Mapping Clones
17p12
RefSeq genes from NCBI
RefSeq Curated
100
ClinGen CNVs: Pathogenic Gain Coverage
Path Gain
0
100
ClinGen CNVs: Pathogenic Loss Coverage
Path Loss
0
WG_HG19 Collapsed Direct repeats hg19 >80% identity

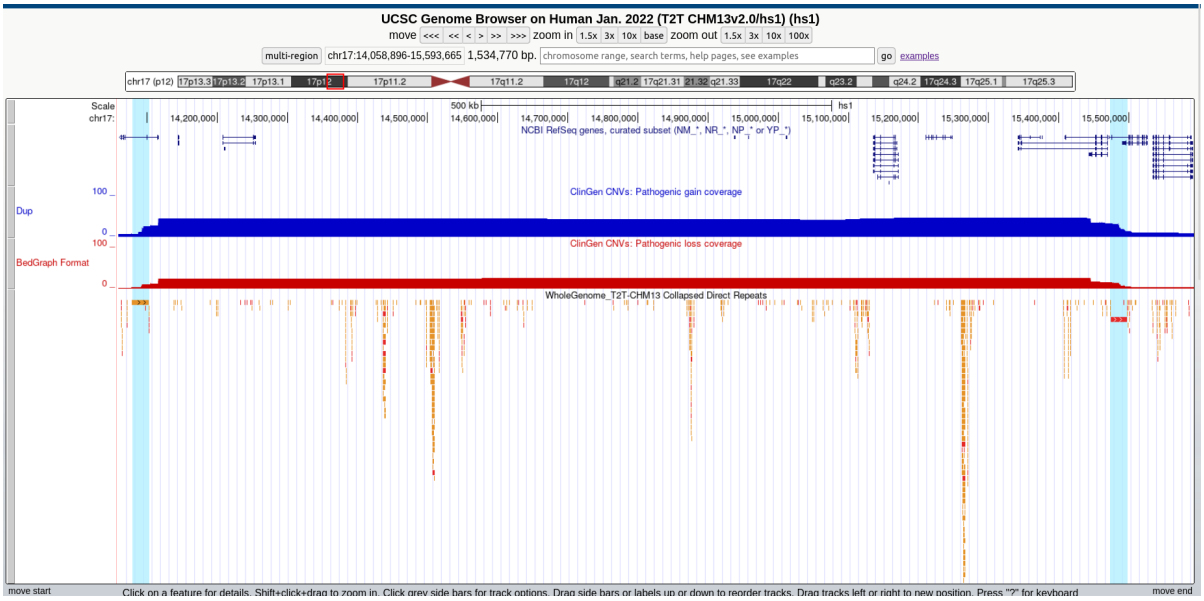


Figure 6. Genomic region 17p12 with Recurrent/Non-recurrent Rearrangement including *PMP22* Gene associated with Genomic Disorders. Displaying GRCh37, GRCh38, T2TCHM-13 respectively. Noting Highlighted Pair of Direct Repeats Flanking Dup/Del Region.

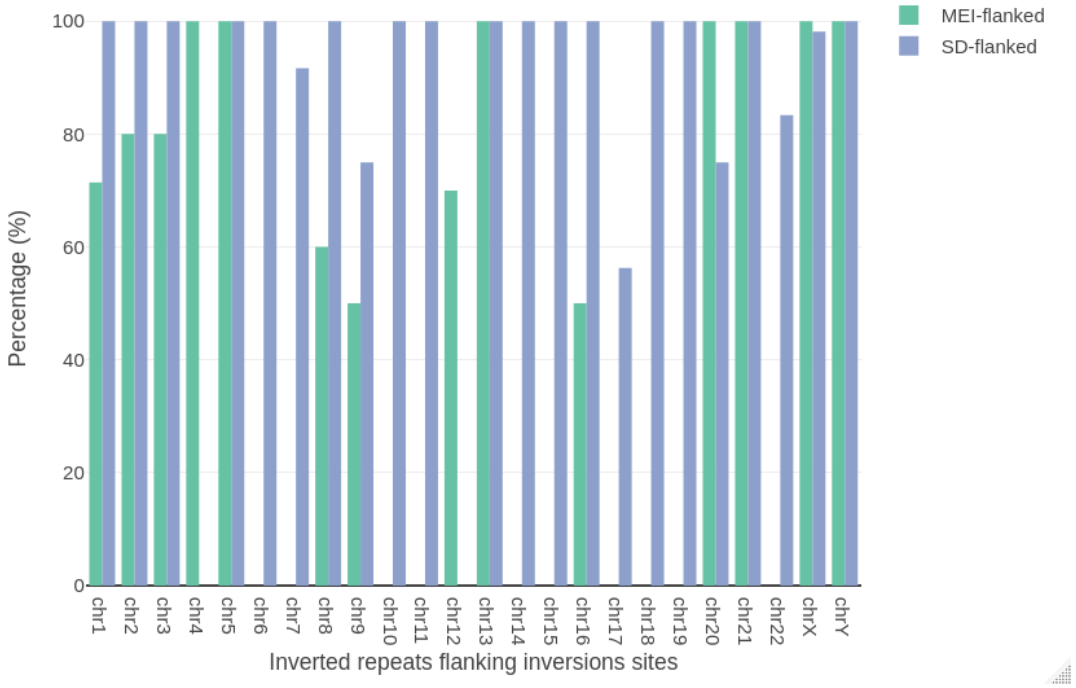


Figure 7. Percentage of inverted elements for each chromosome recovered/captured in our study that overlaps with other elements flanking inversions.

Locus	Genomic disorder	SV type	Genomic rearrangement ocurrance	Recombination substrates ID		
				GCRh37	GCRh38	T2T-CHM13
1pterp36.31	1p36 deletion syndrome	Deletion	Nonrecurrent	DR-205/206/211 a/b	DR-248/249/255 a/b	DR-121/122/132 a/b
	1p36 duplication syndrome	Duplication	Nonrecurrent			
5q35.2-q35.3	Sotos syndrome	Deletion	Recurrent/Nonrecurrent	DR-30271/30292/30289 a/b	DR-31118/31136/31132 a/b	DR-30465/30482/30479 a/b
	5q35 duplication syndrome	Duplication	Recurrent/Nonrecurrent			
7q11.23	Williams-Beuren syndrome	Deletion	Recurrent/Nonrecurrent	DR-21496/21408/21486 a/b	DR-20841/20836/20691 a/b	DR-21233/21132/21229 a/b
	7q11.23 duplication syndrome	Duplication	Recurrent/Nonrecurrent			
8q12.2	CHARGE syndrome	Deletion	Nonrecurrent	DR-16215/16243/16191 a/b	DR-15870/15869/15866 a/b	DR-16626/16426/16498 a/b
	8q12 duplication syndrome	Duplication	Nonrecurrent			
15q11-q13	Prader-Willi syndrome	Deletion (paternal)	Recurrent	DR-292/801 a/b	DR-1429/543 a/b	DR-3968/4790/4217 a/b
	Angelman syndrome	Deletion (maternal)	Recurrent			

	15q11-q13 duplication syndrome	Deletion (maternal)	Recurrent/Nonrecurrent			
16p11.2	16p11.2 deletion syndrome	Deletion	Recurrent/Nonrecurrent	DR-8345/8427/8240 a/b	DR-7848/8013/8043	DR-8305/8348/7933 a/b
	16p11.2 duplication syndrome	Duplication	Recurrent/Nonrecurrent			
17p11.2	Smith-Magenis syndrome	Deletion	Recurrent/Nonrecurrent	DR-6621/6645/6100/6072 a/b	DR-6329/6722/6749/6719 a/b	DR-7028/7025/6567/6500 a/b
	Potocki-Lupski syndrome	Duplication	Recurrent/Nonrecurrent			
17p12	Hereditary neuropathy with liability to pressure palsies (HNPP)	Deletion	Recurrent/Nonrecurrent	DR-5399 a/b	DR-5752 a/b	DR-5919 a/b
	Charcot-MarieTooth disease type 1A	Duplication	Recurrent/Nonrecurrent			
22q11.2	DiGeorge/ Velocardiofacial syndrome	Deletion	Recurrent/Nonrecurrent	DR-1925/1964/1923 a/b	DR-3698/3864/3921 a/b	DR-6710/6577/6515 a/b
	22q11.2 duplication syndrome	Duplication	Recurrent/Nonrecurrent			

Table 6. Repeats Flanking Genomic Disorder Regions - This analysis focuses on identifying direct repeat IDs that flank regions associated with the occurrence of recurrent and nonrecurrent rearrangements linked to the onset of genomic disorders.

3. Materials & Methods

3.1 Identification and collapsing of identical direct and inverted repeat pairs in the human genome

To study and identify repeat sequences in direct and inverted orientations, we conducted bioinformatic analyses using the LastZ 1.04.22 algorithm (2007). The goal was to identify repeat pairs with a minimum pairwise identity of 80% and a minimum length of 200bp. I performed self-alignment of each chromosome in three human genome assemblies: GRCh37 (hg19), GRCh38 (hg38), and T2T-CHM13v2.

LastZ 1.04.22. was run to self-align each chromosome of the three human genome assemblies referenced before, using the following command:

lastz_32 chrN.fa [unmask,softmask=centromere_coordinates] --self --nomirror --step=10 --maxwordcount=1 --masking=100 --strand=minus --seed=match15 --twins=-10..15 --gfextend --ydrop=5000 --interpolation=7000 --filter=identity:80 --filter=nmatch:160 --allocate:traceback=1.99G --outputmasking=chrN_coordinates_masking --format=general:name1,start1,end1,strand1,length1,name2,start2+,end2+,strand2,length2,number,identity,score > repeats_chrN.txt.

The choice of parameters for minimum sequence identity and length was based on experimental data and observations from the literature, both from our own research and that of others. These parameters are known to be relevant for repeat elements that can mediate genomic rearrangements through recombination mechanisms like non-allelic homologous recombination (NAHR), or replication-based processes such as microhomology-mediated break-induced replication (MMBIR) or fork stalling and template switching (FoSTeS).

For each analyzed assembly, I obtained all pairs of repeats on the positive and negative strands of each chromosome. This allowed me to distinguish between repeats on the same strand (direct repeats) and repeats on opposite strands (inverted repeats). I then used an R code for further analysis of these results.

To handle complex regions where repeats may overlap or be located nearby, we developed an algorithm to collapse these repeats while still maintaining a minimum homology of 80%. The LastZ algorithm reports direct or inverted pairs when the alignment exceeds a predetermined minimum alignment score. However, it does not consider the possibility of extending the

alignment further to maintain the required homology. Therefore, the collapsing algorithm is essential for merging the remaining alignments and ensuring accurate results.The resulting dataset, obtained after the collapse of repeats, was used for downstream analyses, annotations, and comparisons. This dataset provides valuable information for understanding the characteristics and distribution of repeat sequences in the human genome.

3.2 Collapsing algorithm development

Once obtained all the repeats across the + and - strand of each chromosome representing the repeats on the same strand (direct repeats) or repeats on different strands (inverted repeats) I piped the results to a R code. In general, this code would take some useful information from the LASTZ results and would build a dataframe the the format with the following general format: repeatID000A, start (initial coordinate of sequence A), end (end coordinate of sequence A), length (distance between these coordinates start A and end A), and repeatID000B, start_1 (initial coordinate of sequence B), end_1 (coordinate of sequence B), length (distance between these coordinates start B and end B) and homology_identity (percentage). This information would allow us to compare each pair of repeats, i.e. pair A1-B1 against pair A2-B2. This dataframe is ordered from left-most coordinate to right-most coordinate by comparing the An coordinates.

Criteria to collapse

The algorithm proceeds to take a repeat, for example, the first repeat (first row) in the dataframe, and would see if the coordinates that delineate An and Bn have an overlap with the coordinates An+1 and Bn+1 or if they are between certain distance:

1) The first overlap can be to the right, that is, that the sequence An+1 or Bn+1 has the initial coordinates somewhere in between sequence An or Bn, correspondingly.
2) The second overlap can be to the right, that is, that the sequence An or Bn has the initial coordinates somewhere in between sequence An+1 or Bn+1, correspondingly.
3) If the sequence An+1 (or Bn+1) is not further than a certain distance from An (or Bn+1, correspondingly). This distance was found to be between 2kb and 3kb, since we ran a histogram of the distance between sequence An and sequence An+1 in the original dataframe to find out what is the mean distance between different repeats. For example, if we find that the mean distance was 2kb, then we would see if sequences An and An+1 are separated by these length or less to fulfill this criteria.

If any of these criteria are met for both sequence A and B, then the algorithm takes the outer-most coordinates that represent the whole overlap for each the part A and B and creates 2 new sequences. Afterwards, we read the fasta file that corresponds to the chromosome in which these repeats were found both for the sequence A and sequence B of that repeat, so that it can use the new coordinates to retrieve the exact sequence represented by them and store it in a temporal vector, which will be our basis of comparison to decide whether the sequences can be collapsed or not. Then, it will compare if the new A sequence has a homology with new sequence B above 80% using the Levenshtein distance metric, and if so, they will be collapsed.

Until now, I explained the basic mechanism on how to collapse sequences, but the datasets are thousands of repeats, so I proceed to do the same exact core algorithm in an iterative way. If 2 repeats are collapsed, for example, repeats 1 and 2, then we store temporarily this new collapse and compare it with the following repeats, in this case repeat 3. If I can collapse them, then I update the new collapse as the sequences that represent the collapses between sequences 1, 2 and 3 and so on. From here two things can happen:

This is done until the last repeat to get the new dataframe with the collapsed repeats and the repeats that could not be collapsed, and this is done both for the direct repeats and the inverted repeats. To find the homology between the sequences A and B in the inverted repeats I need to take the reverse complementary of the coordinates of Bn so that I am able to compare and collapse, since the dataframes are arranged so that sequences An represent the sequence found in the + strand and the sequences Bn are found in the - strand.

3.3 Repeats annotation and assembly comparisons

I used the collapsed datasets for direct and inverted repeats of each of the three genome assemblies to cross-reference the coordinates with known and relevant genomic features such as segmental duplications, repeat elements, and protein-coding genes.
I used CrossMap (Zhao, H et al., 2013) to perform coordinate liftover between the repeats obtained in each of the three genome assemblies analyzed for comparisons. This allowed us to check for sequence overlap between elements in the different assemblies and compare the genome-wide distributions of repeats between and across human genome assemblies.
We looked for overlap between our identified repeats and known repeat elements, such as SINEs, LINEs, LTRs, segmental duplications, and satellites. Therefore, I compared our direct

and inverted repeat datasets for all three assemblies with the RepeatMaskerViz dataset to identify the type of repeat elements that overlapped our repeats in each assembly. Additionally, I cross-referenced the repeats datasets with RefSeq protein-coding genes to identify genes overlapped or flanked by repeated elements, and OMIM annotations for genes associated with human diseases. For overlap, we looked at any repeats overlapping genes by at least 30%, whereas for genes flanked by pairs of repeats we focused on pairs of repeats with features compatible with potential for NAHR, mainly >90% sequence identity and up to a distance of 100 Kb upstream or downstream.

3.4 Ontology analysis

The TopGo package in R was utilized for functional enrichment analysis, with the org.Hs.eg.db annotation package providing Gene Ontology (GO) terms. The gene list used in the analysis was obtained from the downstream analysis, identifying overlapped and flanked genes. GO enrichment analysis of the input gene lists was performed using the runTest function in Gene Ontology, employing Fisher's exact test to determine the significance of gene set overrepresentation in specific GO terms. The "topGO" package's weight algorithm assigned weights to GO terms based on their specificity, facilitating the determination of the number of genes annotated to each term

3.5 Overlap with experimentally validated reported rearrangements

To evaluate the utility of the bioinformatically identified repeats, I looked at the overlap or flanking of experimentally validated structural variants and genomic rearrangements with my datasets. I cross-referenced the coordinates of the repeats in the different assemblies with known genomic disorders, inversions reported by Korbel et al (Porubsky et al., 2022), and the gnomAD CNVs (Collins et al., 2020) .
The corresponding available datasets were obtained and I used bedtools to intersect the coordinates of our direct and inverted repeats datasets with the corresponding regions. The majority of reported datasets were given in GRCh37 coordinates, so we used CrossMap to perform coordinate liftover to the GRCh38 and T2T-CHM13 assemblies.

For flanking regions, coordinates were obtained using bedtools flank, with a maximum distance of 100 kb. The flanking regions were overlapped with the set of direct repeats provided in the

study, and downstream analyses were performed to filter and keep pairs of direct repeats with a homology above 80% flanking on both sides. Additionally, repeats that were flanking only one side of the CNV and those overlapping the whole CNV were also identified.

To visualize repeats and features in genomic regions of interest, we uploaded our generated direct and inverted repeat tracks and looked at other tracks of interest such as pathogenic deletion and duplications from ClinVar, OMIM genes, Repeat Masker, etc.

4. Perspectives and discussion.

Genomic architectural features, such as highly similar repeated sequences, play a crucial role in genome remodeling and long-term evolution. However, they can also contribute to the short-term burden of diseases. Analyzing these genomic features and utilizing in vivo and experimental observations can help predict regions of genomic instability that may lead to genomic rearrangements. Previous studies focusing on potential substrates for non-allelic homologous recombination (NAHR) events have successfully demonstrated that predicted unstable regions do undergo rearrangements in human individuals.

This study aims to address the limitations of previous research by investigating the impact of genomic repeats on protein-coding genes across multiple genome assemblies. By including three different assemblies (GRCh37, GRCh38, and T2T-CHM13), a comprehensive analysis and comparison of repeat-mediated gene rearrangements can be conducted. The study provides specific gene counts and their associations with both direct and inverted repeats, offering a detailed understanding of the potential influence of repeats on gene structure and function.

Moreover, the current work extends the analysis by examining the association of disease-related genes with the identified repeats. By identifying a significant number of genes annotated as diseases associated in OMIM, including those fulfilling dosage-sensitive criteria and known to be involved in genomic disorders, the study establishes the clinical relevance of the identified repeats. This expands our understanding of the potential implications of repeat-mediated rearrangements on disease-associated genes.

In this study, I aimed to provide a genome-wide catalog of direct and inverted repeat sequences in the human genome that can potentially act as substrates for genomic rearrangements through different known mechanisms. This study investigates the distribution and genomic features of repeat pairs with high identity (80-100%) and with a minimum length of 200bp. We analyzed their characteristics and distribution in three

available human genome assemblies to evaluate the effect and impact of variable assemblies in the landscape of these repeats. Our findings show that a large portion of the human genome is potentially susceptible to genomic instability mediated by direct and inverted repeats. No major differences were observed among the different assemblies, except for the newly resolved regions in the T2T-CHM13 assembly not present in the reference sequence. These regions had been difficult to resolve using the conventional sequencing methodologies utilized by the Human Genome Project due to their highly similar and repeat-rich architecture. They are mainly composed of segmental duplications, ribosomal rRNA gene arrays, and satellite arrays that harbor unidentified sets of direct or inverted repeats previously overseen. Therefore, it is unsurprising, yet reassuring, that our analyses were able to identify a high density of repeat pairs and potentially unstable sequences within these regions.

The majority of the repeated pairs identified through these analyses overlap repeat elements in the human genome. Although previous analyses have looked at the distribution of highly identical repeats in the genome that may be substrates for NAHR , given the parameters that we used, we now recover repeat tracts formed by smaller repeated elements such as *Alus* that have been shown to be involved in microhomology-mediated genomic rearrangements (Alu-Alu mediated rearrangements (AAMR)). Further, we identified a significant fraction of protein-coding genes that are overlapped or flanked by our identified repeats. These are interesting because they represent “at risk” genes for potentially rearranging and leading to genomic disorders.

These analyses also showed enrichment of repeats overlapping or flanking genes associated with sensory perception and immune response. CNVs have been previously reported to contribute to the genetic variation in human olfactory receptor repertoire. These findings support the observations that the high abundance of repetitive elements in OR gene clusters contribute to genomic variation. CNVs are known to impact genome evolution and adaptability by facilitating the expansion or contraction of gene families. These findings contribute to the understanding of the genetic potential impact of CNVs and repetitive elements on genome evolution in the context of olfactory receptors and other biological processes.

Although the overall distribution of identified repeat elements across human genome assemblies was very similar, it will be interesting to see if this holds similarly for other genomes as we start obtaining complete human genomes from individuals of diverse ancestries. The availability of long-read sequenced human genomes assembled *de novo* in a reference-free manner in the years to come offers the possibility to expand the landscape of human repeat variation and architecture. Analyses like this will serve as good references to compare the genome-wide landscape and characteristics of these repeats across many human genomes in the near future.

Overall, the results of this study provide a genome-wide map of potential sequences and sites that may serve as substrates for different recombination or replicative-associated mechanisms. These new datasets of direct and inverted repeats in the three currently used human assemblies could help identify elements mediating novel copy-number variants and structural rearrangements that may have functional implications. These data may help uncover new disease-gene associations, facilitate molecular diagnosis, and offer further insights into genomic unstable regions and molecular mechanisms contributing to genome rearrangements.

5. References

Ade, C., Roy-Engel, A. M., & Deininger, P. L. (2013). Alu elements: An intrinsic source of human genome instability. *Current Opinion in Virology*, *3*(6), 639–645. https://doi.org/10.1016/j.coviro.2013.09.002

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Research*, *11*(6), 1005–1017. https://doi.org/10.1101/gr.gr-1871r

Cardoso, A. R., Oliveira, M., Amorim, A., & Azevedo, L. (2016). Major influence of repetitive elements on disease-associated copy number variants (CNVs). *Human Genomics*, *10*(1), 30. https://doi.org/10.1186/s40246-016-0088-9

Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, *17*(4), 224–238. https://doi.org/10.1038/nrg.2015.25

Carvalho, C. M. B., Zhang, F., & Lupski, J. R. (2010). Genomic disorders: A window into human gene and genome evolution. *Proceedings of the National Academy of Sciences*, *107*(suppl_1), 1765–1771. https://doi.org/10.1073/pnas.0906222107

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Currall, B. B., Chiangmai, C., Talkowski, M. E., & Morton, C. C. (2013). Mechanisms for Structural Variation in the Human Genome. *Current Genetic Medicine Reports*, *1*(2), 81–90. https://doi.org/10.1007/s40142-013-0012-8

Dittwald, P., Gambin, T., Gonzaga-Jauregui, C., Carvalho, C. M. B., Lupski, J. R., Stankiewicz,

P., & Gambin, A. (2013). Inverted Low-Copy Repeats and Genome Instability-A Genome-Wide Analysis: HUMAN MUTATION. *Human Mutation*, *34*(1), 210–220. https://doi.org/10.1002/humu.22217

Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M. Y., Rodríguez Rojas, L. X., Elton, L. E., Scott, D. A., Schaaf, C. P., Torres-Martinez, W., Stevens, A. K., Rosenfeld, J. A., Agadi, S., Francis, D., Kang, S.-H. L., Breman, A., Lalani, S. R., Bacino, C. A., … Stankiewicz, P. (2013). NAHR-mediated copy-number variants in a clinical population: Mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Research*, *23*(9), 1395–1409. https://doi.org/10.1101/gr.152454.112

Eichler, E. E. (2001). Segmental duplications: What's missing, misassigned, and misassembled--and should we care? *Genome Research*, *11*(5), 653–656. https://doi.org/10.1101/gr.188901

Emanuel, B. S., & Shaikh, T. H. (2001). Segmental duplications: An "expanding" role in genomic instability and disease. *Nature Reviews Genetics*, *2*(10), 791–800. https://doi.org/10.1038/35093500

Gonzaga-Jauregui, C., & Lupski, J. R. (Eds.). (2021). *Genomics of rare diseases*. Elsevier.

Hastings, P. J., Ira, G., & Lupski, J. R. (2009). A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation. *PLoS Genetics*, *5*(1), e1000327. https://doi.org/10.1371/journal.pgen.1000327

Hauth, A. M., & Joseph, D. A. (2002). Beyond tandem repeats: Complex pattern structures and distant regions of similarity. *Bioinformatics (Oxford, England)*, *18 Suppl 1*, S31-37. https://doi.org/10.1093/bioinformatics/18.suppl_1.s31

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., Limouse, C., Halabian, R., Wojenski, L., Rodriguez, M., Altemose, N., Rhie, A., Core, L. J., Gerton, J. L., Makalowski, W., Olson, D., Rosen, J., Smit, A. F. A., Straight, A. F., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human

repeat elements. *Science*, *376*(6588), eabk3112. https://doi.org/10.1126/science.abk3112

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, *131*(7), 1235–1247. https://doi.org/10.1016/j.cell.2007.11.037

Liehr, T. (2021). Repetitive Elements in Humans. *International Journal of Molecular Sciences*, *22*(4), 2072. https://doi.org/10.3390/ijms22042072

Liu, P., Carvalho, C. M., Hastings, P., & Lupski, J. R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. *Current Opinion in Genetics & Development*, *22*(3), 211–220. https://doi.org/10.1016/j.gde.2012.02.012

Liu, P., Lacaria, M., Zhang, F., Withers, M., Hastings, P. J., & Lupski, J. R. (2011). 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. https://doi.org/10.1016/j.ajhg.2011.09.009

Makova, K. D., & Weissensteiner, M. H. (2023). Noncanonical DNA structures are drivers of genome evolution. *Trends in Genetics*, *39*(2), 109–124. https://doi.org/10.1016/j.tig.2022.11.005

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R.,

Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science (New York, N.Y.)*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Pappalardo, X. G., & Barra, V. (2021). Losing DNA methylation at repetitive elements and breaking bad. *Epigenetics & Chromatin*, *14*(1), 25. https://doi.org/10.1186/s13072-021-00400-z

Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F. A., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Lee, C., Antonacci, F., … Korbel, J. O. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, *185*(11), 1986-2005.e26. https://doi.org/10.1016/j.cell.2022.04.017

Robert S. Harris. (2007). *Improved pairwise alignment of genomic DNA.* The Pennsylvania State University.

Shaw, C. J. (2004). Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease. *Human Molecular Genetics*, *13*(90001), 57R – 64. https://doi.org/10.1093/hmg/ddh073

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, *18*(2), 74–82. https://doi.org/10.1016/S0168-9525(02)02592-1

Vissers, L. E. L. M., Stankiewicz, P., Yatsenko, S. A., Crawford, E., Creswick, H., Proud, V. K., de Vries, B. B. A., Pfundt, R., Marcelis, C. L. M., Zackowski, J., Bi, W., van Kessel, A. G., Lupski, J. R., & Veltman, J. A. (2007). Complex chromosome 17p rearrangements associated with low-copy repeats in two patients with congenital anomalies. *Human Genetics*, *121*(6), 697–709. https://doi.org/10.1007/s00439-007-0359-6

Vollger, M. R., Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K. M., Lewis, A. P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K. H., Phillippy, A. M., Timp, W., Ventura, M., & Eichler, E. E. (2022). Segmental duplications and their variation in a complete human genome. *Science*, *376*(6588), eabj6965. https://doi.org/10.1126/science.abj6965

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*(1), 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Weterings, E., & Chen, D. J. (2008). The endless tale of non-homologous end-joining. *Cell Research*, *18*(1), 114–124. https://doi.org/10.1038/cr.2008.3

Zepeda-Mendoza, C. J., Lemus, T., Yáñez, O., García, D., Valle-García, D., Meza-Sosa, K. F., Gutiérrez-Arcelus, M., Márquez-Ortiz, Y., Domínguez-Vidaña, R., Gonzaga-Jauregui, C., Flores, M., & Palacios, R. (2010). Identical repeated backbone of the human genome. *BMC Genomics*, *11*(1), 60. https://doi.org/10.1186/1471-2164-11-60

Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., & Lupski, J. R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, *41*(7), 849–853. https://doi.org/10.1038/ng.399

Zhang, F., Seeman, P., Liu, P., Weterman, M. A. J., Gonzaga-Jauregui, C., Towne, C. F., Batish, S. D., De Vriendt, E., De Jonghe, P., Rautenstrauss, B., Krause, K.-H., Khajavi, M., Posadka, J., Vandenberghe, A., Palau, F., Van Maldergem, L., Baas, F., Timmerman, V., & Lupski, J. R. (2010). Mechanisms for Nonrecurrent Genomic Rearrangements Associated with CMT1A or HNPP: Rare CNVs as a Cause for Missing Heritability. *The American Journal of Human Genetics*, *86*(6), 892–903. https://doi.org/10.1016/j.ajhg.2010.05.001

Zhao, H, Sun, Z, Wang, J, Huang, H, Kocher, J-P, & Wang, L. (2013). *CrossMap: A versatile tool*

for coordinate conversion between genome assemblies.

https://crossmap.sourceforge.net/