Comparative Analysis of Chromosome 4 between GRCh38 and T2T-CHM13

Zwae Pyae[1]

Received April 14, 2026
Accepted August 12, 2026
Electronic access September 15, 2026

GRCh38, the previous human reference genome assembled primarily with short read sequencing, contained substantial unresolved regions, with roughly eight percent of the genome represented by gaps within highly repetitive telomeric and centromeric sequences, whereas the T2T-CHM13 assembly provides a gapless, telomere-to-telomere human reference genome. This progress was enabled by the development of long-read sequencing technology, specifically those developed by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). A comparative analysis of chromosome 4 from GRCh38 and T2T-CHM13 was performed to examine genomic regions missing from GRCh38 and characterized their nucleotide composition. This analysis shows that T2T-CHM13 closes all gaps found in GRCh38, filling in all centromeric satellite arrays and segmental duplications. T2T-CHM13 has around 3 million more base pairs than GRCh38 in chromosome 4, and T2T-CHM13 shows repetitive AT-rich telomeric and centromeric DNA not found in GRCh38. Sliding window comparison also shows that the total GC content is similar between assemblies, but T2T-CHM13 has unique AT-rich regions that were hidden by gaps in GRCh38. Although the overall nucleotide content is similar, position-specific and unusual high or low GC and AT content reveal how gaps in GRCh38 impact the assembly. With the resolution of all centromeric satellite arrays, segmental duplications, and the short arms of all five acrocentric chromosomes, T2T-CHM13 enhances the detection of structural chromosomal abnormalities. More complete reference assemblies may support future studies of structural variations by providing sequence in regions that were unresolved in earlier references.

Keywords: CHM13, GRCh38, ONT, PacBio, Reference genome.

Introduction

A genome is the complete genetic information of an organism that is encoded as a sequence of nucleic acids[1]. The size of the genomes vary depending on species and can range from millions to billions of base pairs. DNA sequencing technology is used to collect this information. DNA sequencing determines the nucleotide order of genomes and has long been a central goal of molecular biology and genomics. Early sequencing methods, known as short-read sequencing, generate relatively short DNA fragments. Because many repetitive sequences are longer than these short reads, they cannot be uniquely resolved, leaving highly repetitive regions of the genome poorly assembled or missing[2]. The advent of long-read sequencing technologies, which can generate DNA reads longer than the repetitive sequences themselves, have enabled the resolution of previously inaccessible regions, accounting for approximately 8% of the human genome[3]. Reference genomes can be used for read mapping and resequencing, gene annotations, comparative genomics and evolutionary studies. Therefore, it is important that a reference genome is complete and accurate so that it can be generalized to the entire species[4].

In assembled genomes, the character "N" denotes an unknown nucleotide at a given position. Such Ns are inserted by assembly algorithms when reads cannot be confidently assembled, often due to repetitive sequences, and mark unresolved regions in the reference genome. Earlier human reference genomes contain significant gaps and unresolved regions, commonly in highly repetitive regions such as telomeres and centromeres[5]. The presence of these unresolved regions limits the completeness of a human reference genome and prevents accurate detection of structural variations and mutations within these regions. However, recent advancements in long-read sequencing technologies have resulted in complete and accurate reference genomes. Comparing older and newer reference genome assemblies may help reveal the limitations of earlier assemblies and the benefits that newer sequencing technologies bring to the field of genomics and healthcare. This will enable detailed studies of centromere structure, improve the detection of structural variants, and improve overall accuracy of genome mapping[6].

[1] *Malden Catholic High School, USA*

Literature Review

GRCh38 Human Reference Genome

In December 2013, the GRCh38 human reference genome was released through the use of Sanger sequencing. Sanger sequencing, first introduced in 1977, uses DNA polymerase to synthesize DNA from a template while incorporating chain-terminating nucleotides, or ddNTPS, which stop DNA synthesis and produce fragments of different lengths[1]. These fragments are then separated by size to determine the original DNA sequence[7]. GRCh38 was the result of the resolution of roughly 1000 issues from previous human reference genome assemblies. These previous issues were mainly resolved through data from new genome mapping technologies that assisted in identifying and resolving larger assembly issues, such as significant gaps or collapsed segments[8]. Modifications ranging from single bases to thousands of bases, as well as the closing of gaps that were left out due to uncertainty from sequencing, resulted in GRCh38 being the most complete and accurate analysis produced of the human genome. To this day, this reference genome remains one of the most widely used in genetic research[9].

Although GRCh38 represents the culmination of years of improvement upon Sanger sequencing, it contains more than 200 million base pairs of unknown sequences, including the highly repetitive centromeric and telomeric regions[6]. Notably, some of the largest reference gaps include the short arms of all five acrocentric chromosomes, which are chromosomes with centromeres located very close to one end. In addition to these gaps, other regions of GRCh38 may be artificial or incorrect[3]. When highly repetitive regions are longer than the read length of short-read sequencing, gaps will be created in the assembly. These gaps are the result of sequenced repeats erroneously collapsing on top of one another. This may cause complex rearrangements, especially when utilizing Sanger sequencing, where a single read is only 800–900bp long[2].

Third Generation Sequencing (Long-read Sequencing)

These limitations were addressed through the development of long read sequencing technologies. Long-read sequencing produces longer DNA reads than short-read methods, resulting in fewer fragments to be assembled and reducing assembly complexity[10]. Improvements can be seen especially around highly repetitive genomic regions where sequenced repeats may collapse on top of one another. Long-read sequencing technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), are commonly used to generate more complete and accurate assemblies of the entire human genome[11]. However, the completion of telomere to telomere human genome assemblies required not only long-read sequencing, but also ultra-long read datasets, improved assembly algorithms and specialized centromere assembly methods.

Nanopore Sequencing

A nanopore is a small protein pore that acts as a sensor and is embedded in a thin membrane. The nanopore sits in an electrolyte solution, and a constant electric voltage is applied across the membrane. DNA and RNA molecules are negatively charged and are prepared with a motor protein that controls their movement through the nanopore at a constant rate[11]. As the molecule passes through the pore, it partially blocks the electric current, and these current changes are decoded by computational algorithms to determine nucleotide sequence in real time. Additionally, specifically for DNA, the motor protein attached to the DNA enables the double-stranded DNA to be unwound into a single strand to be passed through the nanopore[12]. Nanopore technology has improved over time, producing average read lengths of about 10–20kb, ultra-long reads exceeding 100kb, and some reads reaching several Mb, which is much longer than short read technologies that only produce reads hundreds of bases long. Continuous improvements in nanopore and motor-protein design increased sequencing yield, read length, and accuracy, with modern nanopores reaching up to 450 bases/second. Nanopore sequencing is useful for resolving highly repetitive or complex genomic regions that were difficult to assemble with earlier technologies[13].

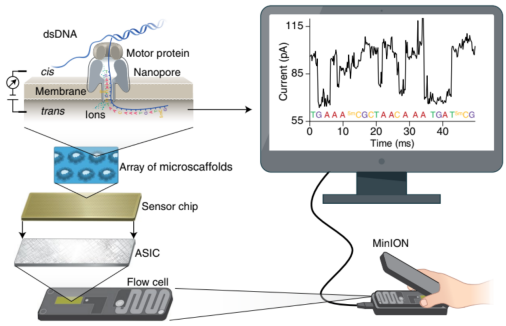


Figure. 1 Principle of Nanopore Sequencing[13]

PacBio HiFi Sequencing

In 2019, Pacific Biosciences introduced high-fidelity (HiFi) sequencing technology. PacBio HiFi Sequencing utilizes Circular Consensus Sequencing (CSS), which converts linear DNA fragments into circular DNA templates[14]. This is done by adding hairpin adapters to both ends of each DNA fragment, transforming the fragment into a circle so that the DNA polymerase can read around the molecule repeatedly. Each

circular DNA is loaded into an SMRT Cell, where a single DNA polymerase is attached at the bottom of a nanostructure called a zero-mode waveguide (ZMW). Each of the four nucleotides (A, T, C, G) carrying a distinct fluorescent dye will diffuse into the ZMW. When the correct nucleotide is incorporated, a light pulse corresponding to the dye will be emitted. The sequencer detects the color and duration of the emitted light pulse, and each color corresponds to a specific base. Circular Consensus Sequencing now achieves 99.9% accuracy for long (up to 25 kb) single-molecule reads. This technology can generate such high-accuracy base calls because the same DNA molecule is read multiple times, which are computationally compared and combined to form a HiFi read [15].

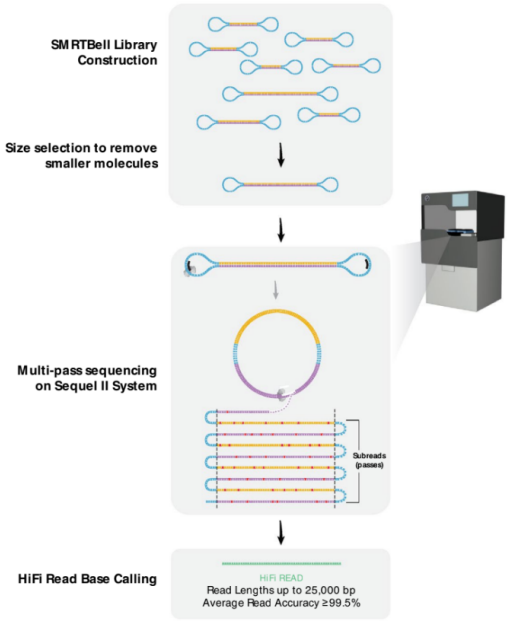


Figure. 2 Flowchart of HiFi sequence read generation [16]

CHM13 Human Reference Genome

The Telomere-to-Telomere CHM13 (T2T-CHM13) human reference genome assembly was published in 2022, and it represents the first complete human genome. CHM13 includes gapless telomere to telomere assemblies for all 22 human autosomes and Chromosome X, and it adds or corrects over 200 million base pairs of genomic sequences compared to GRCh38 [3]. GRCh38 contained unresolved gaps, especially in highly repetitive centromeric, telomeric and acrocentric short arms. CHM13 was created and refined through multiple third-generation sequencing technologies, including PacBio HiFi sequencing and Oxford Nanopore Sequencing. The completed regions from the human genome include all centromeric satellite arrays, the short arms of all five acrocentric chromosomes, and other minor structural variants [17]. The completion of T2T-CHM13 is not only an exercise in technical achievement, but it also provides the essential foundation for understanding the full spectrum of human genetic variation. With a complete and accurate map of a normal, healthy genome, it becomes possible to identify with unprecedented precision the genetic changes that lead to disease.

Chromosomal Abnormalities

The human karyotype, which is made up of 46 chromosomes, contains all the genetic information that is necessary for growth and development. Chromosome abnormalities, or chromosomal aberrations, are disorders usually caused by errors in cell division, which could result in significant consequences, such as malformations or intellectual disabilities. Chromosomal abnormalities can be divided into two categories: numerical changes or structural changes. Numerical chromosomal abnormalities occur when an individual has more or fewer chromosomes compared to the normal total of 46. Structural chromosomal abnormalities occur when the structure of a specific chromosome in an individual is altered. These may include deletions, insertions, duplications, inversions, or translocations. These structural variations may result in important genes being deleted, duplicated, or moved around, potentially disrupting the function of the gene [18].

Chromosome 4 structural abnormalities are associated with different clinical disorders, including Wolf-Hirschhorn Syndrome (WHS) and partial 4p trisomy. Wolf-Hirschhorn Syndrome is a multiple congenital anomaly syndrome and affects many parts of the body at the same time [19]. The first cases of Wolf-Hirschhorn Syndrome were published in the 1960s, and identified that the disorder was caused by the partial deletion of the terminal end of the short arm of chromosome 4 near the telomere [20]. The major clinical features of a person with WHS are poor growth before and after birth and other neurological features, such as intellectual disabilities. Common facial traits include a large forehead, hypertelorism, and the wide bridge of the nose continuing towards the forehead, which leads to the appearance of a typical 'Greek warrior helmet" [21].On the other hand, partial 4p trisomy was first introduced in the 1970s. Partial 4p trisomy is caused by structural variations and rearrangements, typically large duplications that contain two-thirds of the short arm of chromosome 4. 4p trisomy can be characterized by facial dysmorphia, poor growth before and after birth, intellectual disabilities, and other anomalies [22]. The clinical features and traits of trisomy 4p vary from individual to individual because the exact chromosomal region that is affected by the structural variation, usually duplication, differs between different cases [23].

This study evaluates the composition of chromosome 4 between an older human reference genome (GRCh38) and

a newer, complete reference genome (T2T-CHM13) to assess improvements in genome completeness and quality. By focusing on chromosome 4, the study provides a detailed chromosome-level assessment of improvements in the reference genome quality. The findings highlight the importance of accurate and complete reference genomes.

This analysis was motivated in part by ongoing collaborative work focused on assembling a neocentromere-containing cell line. Neocentromeres are ectopic centromeres that form at non-canonical genomic locations and retain centromere function despite lacking the large α-satellite repeat arrays typically found at canonical human centromeres. Because neocentromeres can support proper kinetochore assembly and chromosome segregation without the usual underlying satellite DNA, they provide a valuable model for investigating the relative contributions of DNA sequence and epigenetic chromatin state to centromere identity. To provide a reference framework for interpreting sequence features associated with neocentromere formation, an independent analysis of chromosome 4 in the CHM13 assembly was performed. Although inspired by the neocentromere assembly project, this analysis stands alone as a characterization of the canonical chromosome 4 centromeric region in a fully resolved human genome. The CHM13 assembly is particularly well suited for this purpose because it represents the first telomere-to-telomere human reference with complete resolution of centromeric and other highly repetitive regions, enabling analyses that were not possible with earlier references such as GRCh38. By examining chromosome 4 centromeric sequence composition, including repeat content and assembly completeness, this analysis establishes a baseline for the expected architecture of a canonical centromere and provides a foundation for future comparisons with neocentromeric loci.

Methods

The study aims to perform a comparative computational analysis of human chromosome 4 between two reference genomes: GRCh38 and T2T-CHM13. The overall analysis is structured into four sections: terminal sequence window analysis, whole-chromosome summary statistics, N content and percent N distribution, and sliding window GC and AT content analysis. The analysis focused on identifying the differences in sequence completion, composition, and gap content between the two reference genome assemblies.

The FASTA sequences for the Chromosome 4 region of both GRCh38 and T2T-CHM13 reference genomes were obtained using the publicly available UCSC Genome Browser[24]. The FASTA files were then imported, and Python (version 3.12.12) was used to perform statistical analysis on the genome sequences.

To evaluate completeness of the two sequences at the chromosomal ends at the most fundamental level, a simple terminal-sequence completeness check was performed to compare whether each chromosome end was resolved in each assembly. The first 100 bases corresponding to the 5' terminal end of chromosome 4 and the final 100 bases representing the 3' terminal end of chromosome 4 for each assembly were examined. Chromosomal ends contain telomeric and subtelomeric regions, and these regions have historically been difficult to sequence and assemble, thus representing a critical point of comparison between GRCh38 and T2T-CHM13. This 100 base window analysis was used as an illustrative check of sequence resolution rather than a complete characterization of telomeric or subtelomeric structure.

For centromeric repeat comparison, the GRCh38 chromosome 4 interval chr4:49,325,000–49,525,000 was selected because it falls within the expected centromeric region of chromosome 4. To avoid comparing identical coordinate numbers across two different genome assemblies, the GRCh38 interval was converted to the corresponding T2T-CHM13 coordinate range using the UCSC LiftOver tool. LiftOver mapped this GRCh38 interval to chr4:48,659,127–56,428,020 in T2T-CHM13. These assembly specific coordinate ranges were then viewed separately in the UCSC Genome Browser. RepeatMasker tracks were used to visualize annotated repetitive elements within each interval. Additionally, RepeatMasker annotation tables for the respective intervals were downloaded and accessed for both GRCh38 and T2T-CHM13 assemblies. For each RepeatMasker entry, repeat length was calculated as the end coordinate minus the start coordinate. The repeat class was extracted from the repeat annotation label, and entries were then grouped by repeat class. For each repeat class, the class count, total annotated bp and the percent of interval were calculated. Because RepeatMasker entries can overlap, the total annotated bp was interpreted as the summed RepeatMasker annotation length instead of non-overlapping coverage.

The two independent variables are the DNA sequence of chromosome 4 for GRCh38 and T2T-CHM13. The definitions of the dependent variables, as well as the statistical analysis techniques used to test them, are mentioned below.

Chromosome Length: Measured as the total number of base pairs in chromosome 4. It is calculated by counting all the bases in each FASTA sequence using Python

Total N Count: The total number of undefined or unresolved bases in the assembly. It can be calculated by finding the number of all occurrences of the letter " N " in chromosome 4.

Percent N: Measured as the proportion of "N" bases relative to the total chromosome length of each reference genome.

GC Percent: The calculation of the percentage of nitrogenous bases in a DNA sequence that consists of Guanine (G) or Cytosine (C), and can be used to identify how GC-rich a DNA sequence is. In Python, the GC Percent can be calculated by counting the total number of G and C bases and dividing the

result by the total chromosome length.

AT Percent: The calculation of the percentage of nitrogenous bases in a DNA sequence that consists of Adenine (A) or Thymine (T). It can be used to identify how AT-rich a DNA sequence is. In Python, the AT Percent can be calculated by counting the total number of A and T bases and dividing the result by the total chromosome length.

CpG Count: the total number of CpG dinucleotides in a DNA sequence. A CpG site occurs when a cytosine nucleotide (C) is followed by a guanine nucleotide (G) on the same DNA strand.

CpG Density: Measured as the number of CpG dinucleotides per million base pairs. CpG density was calculated using the formula: CpG Density = (CpG Count / chromosome length) $\times$ 1,000,000.

Shannon Entropy: The quantitative estimate of sequence complexity.

Sliding Window Percent N / Gap Content: The percentage of "N" bases within the fixed-length windows across chromosome 4 for both reference genomes.

Sliding Window GC Content: The GC percentage within fixed-length windows across chromosome 4 for both reference genomes.

Sliding Window AT Content: The AT percentage within fixed-length windows across chromosome 4 for both reference genomes.

The GC and AT content analyses were done in Google Colab using Python. The chromosome 4 FASTA files for GRCh38 and T2T-CHM13 were loaded using Biopython, and each chromosome 4 sequence was divided into 1 million bp windows. For GC content, the code counted the number of G and C bases, and then the number of all 4 nucleotide bases total and calculated the GC percent. For AT content, the code counted the number of A and T bases and then the number of all 4 nucleotide bases to calculate the AT percent. Both GC and AT percentages were calculated only from resolved DNA bases and did not include unresolved gap characters.

All data used was obtained from publicly available human reference genomes from the UCSC genome browser. There are no personally identifiable or individual-level genetic data that were analyzed. The study involves no human subjects or experiments. There are no ethical risks, as data analysis was limited to computational sequence analysis only.

Results: Assembly Completeness and Sequence Composition

Section I: Terminal Sequence Comparison Showing Gaps at the Beginning of Chromosome 4

For the first 100 bases of Chromosome 4 in the GRCh38 human reference genome, the sequence consisted entirely of "N" characters, indicating that this telomeric region remains unresolved and is represented solely by gap characters in the assembly. The final 100 bases of chromosome 4 in GRCh38 were also examined, and the sequence also consisted entirely of "N" characters, showing that the opposite 3' terminal end is likewise unresolved.

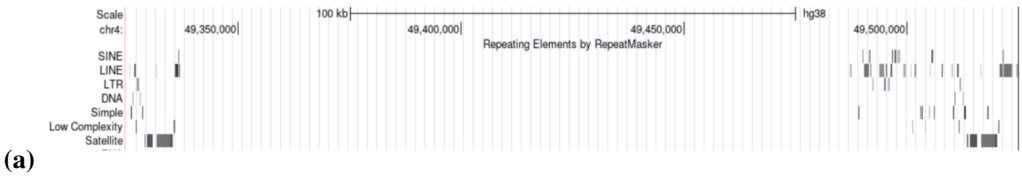
In comparison, for the first and final 100 bases of Chromosome 4 in the T2T-CHM13 human reference genome, the sequences were fully resolved with no "N" characters present. These terminal sequences consist of repetitive sequence patterns, consistent with the known sequence composition of telomeric regions. A preliminary visual inspection suggests the presence of tandem repeat motifs, though definitive characterization of the repeat structure would require dedicated repeat analysis tools (e.g., Tandem Repeats Finder) rather than manual examination of the first and final 100 bases alone. This finding supports the conclusion that CHM13 provides sequence coverage at both terminal regions where GRCh38 contains only gaps, though a comprehensive repeat analysis across the full telomeric region would be necessary to characterize the repetitive landscape accurately.

Because GRCh38 and T2T-CHM13 were derived from different biological sources and assembly strategies, this comparison should be interpreted as an assembly level comparison of terminal sequence completeness rather than a direct comparison of the same individual genome.

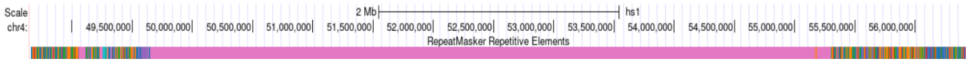
Section II : RepeatMasker Annotations and RepeatMasker Tables Revealing Additional Repetitive Elements

Repeat annotations were obtained using RepeatMasker data from the UCSC Genome browser. A visual observation of the figures indicates that a greater number of repetitive elements were identified in the CHM13 assembly (Figure 3b) compared to GRCh38 (Figure 3a). In Figure 3a, a large gap is observed within the assembly, indicating that there are missing or unresolved genome regions. However, in Figure 3b, additional repeat annotations are detected. Notably, a large proportion of the repetitive elements are shown in pink, representing satellite repeats. Satellite repeats are known to be highly concentrated in centromeric regions, and when compared to the limited satellite annotations in GRCh38, the newer assembly resolves substantially more satellite-rich centromeric DNA.

The RepeatMasker tables in Figures 4 and 5 compare the repeat composition of the chromosome 4 centromeric interval in GRCh38 and the corresponding LiftOver-derived interval in T2T-CHM13. In GRCh38, the analyzed centromeric interval contains only 70 total RepeatMasker annotations. The total annotated repeat length is around 26,500 bp across the 200,000 bp interval. The largest repeat classes in GRCh38 are Satellite repeats representing 5.4% of the interval, LINE repeats consisting of 5.01% of the interval, LTR repeats con-



(a)



Color	Repeat Class
	SINE - Short Interspersed Nuclear Element
	LINE - Long Interspersed Nuclear Element
	LTR - Long Terminal Repeat
	DNA - DNA Transposon
	Simple - Single Nucleotide Stretches and Tandem Repeats
	Low_complexity - Low Complexity DNA
	Satellite - Satellite Repeats
	RNA - RNA Repeats (including RNA, tRNA, rRNA, snRNA, scRNA, srpRNA)
	Other - Other Repeats (including class RC - Rolling Circle)
	Unknown - Unknown Classification

(b)

Figure. 3 UCSC Genome Browser RepeatMasker view for GRCh38 and T2T CHM13
(a) Track displaying UCSC Genome Browser view for GRCh38 within chromosome 4, regions 49,325,000bp - 49,525,000bp. Black color bars represent repetitive elements.
(b) Track displaying UCSC Genome Browser view for CHM13 within chromosome 4, regions 48,659,127bp - 56,428,020bp. Color legend for the UCSC RepeatMasker track showing the repeat classes represented by each color in the genome browser visualization.

Repeat Class	Class Count	Total Covered (bp)	Centromeric Interval (%)
Satellite	9	10,830	5.41
LINE	24	10,024	5.01
LTR	6	2,027	1.01
SINE	9	1,966	0.98
Low complexity	6	777	0.39
Simple repeat	12	652	0.33
DNA	4	280	0.14

Figure. 4 Repeat composition of CHM13 centromeric intervals (RepeatMasker annotation)

Repeat Class	Class Count	Total Covered (bp)	Centromeric Interval (%)
Satellite	77	5,808,643	74.768
LINE	717	3,883,497	49.99
LTR	487	641,466	8.257
SINE	1017	319,805	4.116
Low_complexity	57	4,546	0.059
Simple_repeat	450	122,522	1.577
DNA	282	231,126	2.975
Retroposon	11	16,234	0.209
scRNA	2	222	0.0029
snRNA	5	618	0.008
srpRNA	2	1078	0.0139
tRNA	1	76	0.00098
Unknown	50	87446	1.126

Figure. 5 Repeat composition of CHM13 centromeric intervals (RepeatMasker annotation). Centromeric intervals may include overlapping repeat annotations.

sisting of 1.01% of the interval, and SINE repeats representing 0.98% of the interval. The low centromeric interval percentages of the repeats observed in GRCh38 underrepresents the repetitive-rich sequence expected in the chromosome 4 centromere.

In T2T-CHM13, the corresponding centromeric interval contains over 3000 RepeatMasker annotations, which is much more than GRCh38. The repeat classes are also more diverse in CHM13, with additional categories such as Retroposon, scRNA, snRNA, etc… appearing in the table. Satellite repeats cover the greatest amount of this centromeric interval, with 74.768% coverage, whereas in GRCh38, satellite repeats only account for 5.41% of the centromeric interval. This is evidence that CHM13 resolves a large satellite-rich centromeric

region that isn't well represented in GRCh38.

Section III : Whole Chromosome Statistics Showing Greater Length and Complete N Removal in T2T-CHM13

Reference Genome	Length	Total N Count	Percent N	GC Percent	AT Percent	CpG Count	Shannon Entropy
GRCh38	190,214,555	461888	0.242825	38.150604	61.606572	1,503,429	1.958483
CHM13	193,574,945	0	0.000000	38.130899	61.869101	1,546,133	1.958955

Figure. 6 Comparison of summary statistics for chromosome 4 between GRCh38 and CHM13

Total N Count

When observing and comparing the lengths of both human reference genomes, there is an observable difference in length between the two reference genomes. GRCh38 has a total of 190,214,555 bp, while CHM13 has a total of 193,574,945 bp, indicating that the length of CHM13 is longer than the length of GRCh38. A possible explanation for this 3.36 million bp difference between the two reference genomes could be the result of the inability to sequence repetitive regions and missing sequences or gaps during the assembly of the GRCh38 reference genome.

Percent N

The Percent N in chromosome 4 for the reference genome GRCh38 is 0.243, while the Percent N in chromosome 4 for the reference genome CHM13 is 0. This shows that while 0.243% of GRCh38's chromosome 4 seems to have "N" instead of a nucleotide, there are no "N"s in CHM13's chromosome 4. This indicates that GRCh38 has many gaps and unresolved regions that are being replaced by "N", whereas CHM13, due to the effective long-read sequencing used, has no gaps or unresolved regions in the sequenced chromosome 4.

GC Percent

The GC content of chromosome 4 in GRCh38 is 38.151%, while in T2T-CHM13 it is 38.131%. The difference of 0.02 percentage points is minimal, indicating that the overall base composition of chromosome 4 is highly consistent between the two assemblies.

This similarity is consistent with the genomic landscape of human chromosomes. GC content is strongly associated with gene-rich euchromatic regions, which were already well-represented in GRCh38. The newly completed sequences in T2T-CHM13 consist primarily of heterochromatic regions—including centromeric satellite arrays, telomeric repeats, and acrocentric short arms—which are known to be AT-rich. The addition of these AT-rich sequences marginally dilutes the overall GC percentage, explaining the slightly lower value observed in T2T-CHM13. The near-identity of the two values confirms that the euchromatic gene-containing regions were accurately assembled in GRCh38, while T2T-CHM13 has successfully added the AT-rich heterochromatic components without substantially altering the overall compositional statistics.

AT Percent

The AT% in chromosome 4 for the reference genome GRCh38 is 61. 607%, while the AT% in chromosome 4 for the reference genome CHM13 is 61.869%. CHM13 has a higher AT% than GRCh38. This is different from the results acquired for GC percent.

CpG Count

The chromosome 4 of GRCh38 has a CpG Count of 1,503,429, while chromosome 4 of CHM13 has a CpG Count of 1,546,133. CHM13 has around 43,000 more CpGs than GRCh38. T2T-CHM13 showed a higher CpG count when compared to GRCh38. The higher CpG count means that T2T-CHM13 contains more total CpG dinucleotide sites. This suggests that some regions that were missing or unresolved in GRCh38 contained CpG sites that were recovered in the T2T-CHM13 assembly.

CpG Count Density

GRCh38 has a CpG density of 7903.86 CpGs/Mb per million bases, while CHM13 has a CpG density of 7987.26 CpGs/Mb per million bases. CHM13 has a higher CpG density per million bases than GRCh38.

hannon Entropy

T2T-CHM13 shows slightly higher sequence complexity with 1.959, whereas GRCh38 has slightly lower sequence complexity with 1.958. This difference is negligible and suggests that the overall nucleotide-level sequence complexity of chromosome 4 is similar between the two reference genomes. Thus, the main improvement of T2T-CHM13 is the completion of regions that were unresolved or represented by "N", not a major change in the diversity of nucleotides.

Section IV: Sliding Window N Content Analysis Identifies GRCh38 Assembly Gaps Near Repetitive Regions

Figure 7 shows the sliding window percent N across chromosome 4 for both GRCh38 and T2T-CHM13. Figure 7 does not directly visualize telomeric or centromeric DNA. Instead, it shows the proportion of unresolved "N" bases across chromosome 4 in sliding windows. Telomeric regions and centromeric regions are likely repetitive and gap-rich because historically, they have been difficult to resolve in human reference genome assemblies. In Section I, printing out the first 100 bases of both GRCh38 (red) and CHM13 (blue) showed

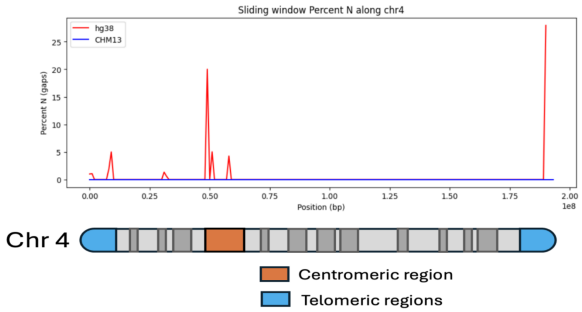


Figure. 7 Sliding Window Percent N along chromosome 4 for GRCh38 (hg38) and CHM13. Chromosome 4 ideogram displaying centromeric and telomeric regions.

that GRCh38 consisted entirely of ”N”s over the first 100 bases, while the corresponding region in T2T-CHM13 contained resolved nucleotide bases. These “N” characters indicate that the terminal sequence at the beginning of GRCh38 chromosome 4 is unresolved in the assembly. In contrast, the absence of “N” characters in the corresponding T2T-CHM13 sequence shows improved terminal sequence resolution.

Observing the visualization for GRCh38, there are moderate spikes around 50 million bp. The UCSC Genome Browser GRCh38 Cytoband Ideogram identifies the chromosome 4 centromeric region near 50 million bp[24]. Therefore, the spikes are consistent with the expected centromeric region of chromosome 4 where repetitive DNA can make assembly difficult. Additionally, another large spike can be seen around 185 million bp. Telomeres are located at the terminal ends of linear chromosomes and are composed of repetitive arrays[25]. Therefore, the spike in Figure 7 near 185 million bp is consistent with an unresolved telomeric region in GRCh38. Because repetitive telomeric sequences are difficult to assemble accurately using short read sequencing, this terminal increase in N content is consistent with the expected challenge of assembling chromosome end sequences.

In contrast, the absence of spikes in CHM13 indicates that the assembled chromosome 4 contained no “N” characters. This shows that previously gap-rich regions have been resolved due to the improvement in sequencing technology. Overall, the sliding window visualization highlights that T2T-CHM13 offers a more complete and reliable chromosome 4 assembly compared to GRCh38 by resolving all the unresolved gaps that GRCh38 contained

Figure 8 is a visualization of the gap content (N Content) comparison for chromosome 4 between GRCh38 and CHM13. Percent N: 0.24% GRCh38 N’s and 0% N’s in CHM13. Therefore, the red bar for GRCh38 (hg38) shows that a small but measurable portion of chromosome 4 remains unresolved, while T2T-CHM13 contains no unresolved N bases in the analysis.

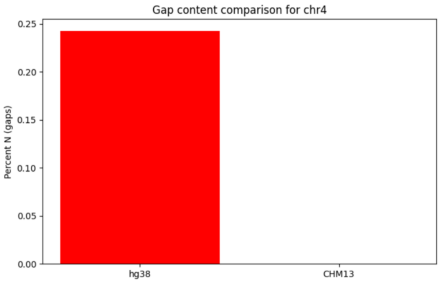


Figure. 8 Bar chart comparison of the gap content for chromosome 4

Section V: Sliding Window GC / AT Content Analysis Shows Local Composition Differences Between Assemblies

Figure. 9 Sliding window GC content across chromosome 4 for GRCh38 (hg38) and CHM13

Figure 9 is a sliding window visualization of the GC Content across Chromosome 4 for both reference genomes. The position over chromosome 4 (bp) and the GC%

From 0 to 10 million bp, there is high GC content. This region is consistent between GRCh38 and CHM13. From 47 to 52 million bp, there is low GC content. This region doesn’t seem to be consistent, as only CHM13 shows a low GC content around this region and not GRCh38. Finally, from 187 to 193 million bp, there is a high GC content for both reference genomes. This region is somewhat consistent with both, as both assemblies have this high GC content, but there is an observable positional misalignment. Due to technology constraints, GRCh38 contains many unresolved gaps in the assembly labeled “N”, which reduces the number of resolved bases. Because GC% is calculated relative to the length of the sequenced region, the lack of “G” and “C” bases in the gaps results in GRCh38 being positionally shifted when compared to CHM13.

Around the same areas where there is high or low GC content in either GRCh38 or CHM13, there also seems to be a correlation to gap content (N%), as similar positions in the chromosome can be seen having gaps as well as having large GC content. This could be an indication of a correlation be-

tween GC content and gap content, showing that the gap content increases when there seems to be unusually high or low GC content.

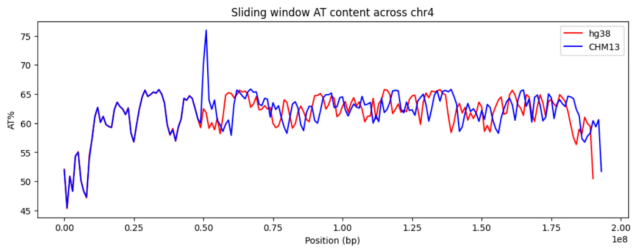


Figure. 10 Sliding window AT content across chromosome 4 for GRCh38 (hg38) and CHM13

Figure 10 is a sliding window visualization of the AT Content across Chromosome 4 for both reference genomes.

Looking at the visualization, both GRCh38 and CHM13 have similar AT content along the entirety of 0 to 50 million bp. The AT% gradually increases from 0 to 13 million bp in CHM13. However, from 50 to 55 million bp, there is a large spike in AT%. Similarly to Figure 9, there is an observable positional misalignment after the spike from 50 to 55 million bp. Due to technology constraints, GRCh38 contains many unresolved gaps in the assembly labeled "N", which reduces the number of resolved bases. Because AT% is calculated relative to the length of the sequenced region, the lack of "A" and "T" bases in the gaps results in GRCh38 being positionally shifted when compared to CHM13. Near the end, an unusually low AT content can be observed for both reference genomes.

Comparing Figure 9 and Figure 10, the average AT% is higher than the average GC%. In human chromosomes, large portions of the chromosome mainly consist of non-coding regions, and these regions tend to be more AT-rich. Moreover, GC-rich regions are often associated with gene-dense regions, and because large portions of the chromosome are mainly non-coding regions, the overall GC% will be much lower than the AT%.

Discussion

Chromosome 4 in the GRCh38 reference genome contains a large number of unresolved bases, while T2T-CHM13 contains no gaps, indicating a fully resolved assembly. Additionally, the first 100 bases of chromosome 4 in GRCh38 consist entirely of "N"s, while CHM13 shows repetitive sequences that are AT-rich, indicating the telomeric region. Moreover, the sliding window visualization also suggests that in GRCh38, the large spikes in percent N occur around the start, 50 million bp region, and the end, which are areas consisting of telomeric and centromeric regions.

A rough comparison of the two genome assemblies shows that CHM13 contains a longer chromosome 4 sequence, indicating that GRCh38 had missing genomic regions. GC content is similar across both reference genomes, but CHM13 has higher AT content, CpG count, and slightly higher CpG density. CpG Density is important for gene regulation, as CpG dinucleotides are often found in promoter regions and CpG islands. They are also important for evolutionary conservation, as CpG-rich regions tend to be highly conserved across species, which reflects their importance in regulating essential genes in organisms.

Sliding window analysis reveals that overall GC content for chromosome 4 is similar across both reference genomes. High GC content regions in both assemblies are observed near the beginning and end of the chromosome, and a low GC content region is observed around 47 to 52 million bp, which is observed only in CHM13. After this low GC content region, GC content patterns differ slightly between the two assemblies due to the positional misalignment between the two caused by gaps in GRCh38. The high or low GC content region correlates with increased gap content.

Sliding window analysis reveals that overall AT content for chromosome 4 is higher in CHM13 than GRCh38. AT content is similar across the first half of chromosome 4, but there is a large spike of AT content around 50 to 55 million bp present only in CHM13. After this spike, AT content patterns differ slightly between the two assemblies due to positional misalignment between the two caused by gaps in GRCh38. Average AT content exceeds average GC content.

Study Limitations

This study only focuses on chromosome 4 due to computational storage constraints, so the findings may not fully represent all the trends or patterns across other chromosomes or the entire genome. Additionally, CHM13 is derived from a completely homozygous human cell that only contains X chromosomes to make it an ideal, gap-free human reference genome. As a result, Y chromosome sequences are absent in CHM13, meaning that CHM13 cannot be considered a fully complete human reference genome for both sexes. Moreover, CHM13 is from an individual of European ancestry, and this reference genome may not be the most accurate reference for other ethnic groups or races[3]. For example, individuals of African ancestry typically have a higher genomic diversity. Thus, comparison to only this reference genome may be confounding.

Conclusion

Advances in genome sequencing and assembly technologies have greatly improved the ability to generate complete human

genome assemblies. Future research could focus on prioritizing accuracy and speed while expanding to include genomes for a broader range of individuals. Additionally, future studies should go beyond the analysis of individual chromosomes to understand the genome-wide impact of long-read sequencing assemblies over short-read sequencing assemblies. For example, the Human Pangenome Reference Consortium is a representation of a pangenome reference, which consists of genomes from various individuals from diverse ancestral and geographic backgrounds[26]. Unlike the traditional reference genome, a pangenome combines multiple complete assemblies. Similar to the work being done by the Human Pangenome Reference Consortium, transitioning from a single reference genome to a pangenome reference will allow for a more comprehensive representation of human genetic diversity and improve the accuracy of detecting variants in the genome across multiple populations.

Additionally, the creation of more complete reference genomes will allow for better detection of structural variations and mutations that may be missed in outdated assemblies that have many unresolved gaps. Moreover, being able to accurately identify these structural variations and mutations, especially those that are related to diseases, improves the ability to connect these abnormalities to inherited disorders and diseases. Thus, the diagnostic accuracy of diseases will be improved, risks can be better assessed, and treatment strategies can be amended for the individual. Ultimately, using complete reference genomes increases the reliability of genomic research and supports its application in healthcare.

More complete reference genomes, such as T2T-CHM13, could significantly improve the identification and understanding of chromosomal abnormalities. This could provide a stronger reference framework for future structural variant studies particularly in repetitive, terminal and centromeric regions. Furthermore, improvements in reference genomes open the door to the future for gene editing technologies such as CRISPR to accurately identify and modify disease causing genetic mutations. However, clinical applications such as disease detection, risk prediction or treatment development would require additional analyses beyond the scope of this study.

References

1 A. M. Giani, G. R. Gallo, L. Gianfranceschi, G. Formenti. Long walk to genomics: history and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal. Vol. 18, pg. 9–19, 2020, https://doi.org/10.1016/j.csbj.2019.11.002.

2 T. J. Treangen, S. L. Salzberg. Repetitive dna and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. Vol. 13, pg. 36–46, 2012, https://doi.org/10.1038/nrg3117.

3 S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N.-C. Chen, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, R. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, J. Kremitzki, H. Li, V. V. Maduro, T. Marschall, J. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy. The complete sequence of a human genome. Science. Vol. 376, pg. 44–53, 2022, https://doi.org/10.1126/science.abj6987.

4 S. Ballouz, A. Dobin, J. A. Gillis. Is it time to change the reference genome? Genome Biology. Vol. 20, Article 159, 2019, https://doi.org/10.1186/s13059-019-1774-4.

5 D. Paulino, R. L. Warren, B. P. Vandervalk, A. Raymond, S. D. Jackman, I. Birol. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics. Vol. 16, Article 230, 2015, https://doi.org/10.1186/s12859-015-0663-4.

6 S. Aganezov, S. M. Yan, D. C. Soto, M. Kirsche, S. Zarate, P. Avdeyev, D. J. Taylor, K. Shafin, A. Shumate, C. Xiao, J. Wagner, J. McDaniel, N. D. Olson, M. E. G. Sauria, M. R. Vollger, A. Rhie, M. Meredith, S. Martin, J. Lee, S. Koren, J. A. Rosenfeld, B. Paten, R. Layer, C.-S. Chin, F. J. Sedlazeck, N. F. Hansen, D. E. Miller, A. M. Phillippy, K. H. Miga, R. C. McCoy, M. Y. Dennis, J. M. Zook, M. C. Schatz. A complete reference genome improves analysis of human genetic variation. Science. Vol. 376, pg. eabl3533, 2022, https://doi.org/10.1126/science.abl3533.

7 F. Sanger, S. Nicklen, A. R. Coulson. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. Vol. 74, pg. 5463–5467, 1977, https://doi.org/10.1073/pnas.74.12.5463.

8 V. A. Schneider, T. Graves-Lindsay, K. Howe, N. Bouk, H.-C. Chen, P. A. Kitts, T. D. Murphy, K. D. Pruitt, F. Thibaud-Nissen, D. Albracht, R. S. Fulton, M. Kremitzki, V. Magrini, C. Markovic, S. McGrath, K. M. Steinberg, K. Auger, W. Chow, J. Collins, G. Harden, T. Hubbard, S. Pelan, J. T. Simpson, G. Threadgold, J. Torrance, J. M. Wood, L. Clarke, S. Koren, M. Boitano, P. Peluso, H. Li, C.-S. Chin, A. M. Phillippy, R. Durbin, R. K. Wilson, P. Flicek, E. E. Eichler, D. M. Church. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Research. Vol. 27, pg. 849–864, 2017, https://doi.org/10.1101/gr.213611.116.

9 Y. Guo, Y. Dai, H. Yu, S. Zhao, D. C. Samuels, Y. Shyr. Improvements and impacts of grch38 human reference on high throughput sequencing data analysis. Genomics. Vol. 109, pg. 83–90, 2017, https://doi.org/10.1016/j.ygeno.2017.01.005.

10 C. Kim, M. Pongpanich, T. Porntaveetus. Unraveling metagenomics through long-read sequencing: a comprehensive review. Journal of Translational Medicine. Vol. 22, pg. 111, 2024, https://doi.org/10.1186/s12967-024-04917-1.

11 G. A. Logsdon, M. R. Vollger, E. E. Eichler. Long-read human genome sequencing and its applications. Nature Reviews Genetics. Vol. 21, pg. 597–614, 2020, https://doi.org/10.1038/s41576-020-0236-x.

12 T. Xiao, W. Zhou. The third generation sequencing: the advanced approach to genetic diseases. Translational Pediatrics. Vol. 9, pg. 163–173,

2020, https://doi.org/10.21037/tp.2020.03.06.

13 Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au. Nanopore sequencing technology, bioinformatics and applications. Nature Biotechnology. Vol. 39, pg. 1348–1365, 2021, https://doi.org/10.1038/s41587-021-01108-x.

14 B. Wang, P. Jia, S. Gao, H. Zhao, G. Zheng, L. Xu, K. Ye. Long and accurate: how HiFi sequencing is transforming genomics. Genomics, Proteomics & Bioinformatics. Vol. 23, No. 1, pg. 1–14, 2025, https://doi.org/10.1093/gpbjnl/qzaf003.

15 A. M. Wenger, P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall, G. T. Concepcion, J. Ebler, A. Fungtammasan, A. Kolesnikov, N. D. Olson, A. Töpfer, M. Alonge, M. Mahmoud, Y. Qian, C.-S. Chin, A. M. Phillippy, M. C. Schatz, G. Myers, M. A. DePristo, J. Ruan, T. Marschall, F. J. Sedlazeck, J. M. Zook, H. Li, S. Koren, A. Carroll, D. R. Rank, M. W. Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology. Vol. 37, pg. 1155–1162, 2019, https://doi.org/10.1038/s41587-019-0217-9.

16 T. Hon, K. Mars, G. Young, Y.-C. Tsai, J. W. Karalius, J. M. Landolin, N. Maurer, D. Kudrna, M. A. Hardigan, C. C. Steiner, S. J. Knapp, D. Ware, B. Shapiro, P. Peluso, D. R. Rank. Highly accurate long-read hifi sequencing data for five complex genomes. Scientific Data. Vol. 7, pg. 399, 2020, https://doi.org/10.1038/s41597-020-00743-4.

17 K. H. Miga, S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. ana Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, A. M. Phillippy. Telomere-to-telomere assembly of a complete human x chromosome. Nature. Vol. 585, pg. 79–84, 2020, https://doi.org/10.1038/s41586-020-2547-7.

18 D. A. Queremel Milani, P. Tadi. Genetics, chromosome abnormalities. in StatPearls StatPearls Publishing, Treasure Island (FL), 2025.

19 E.-C. Gavril, A. C. Luca, A.-S. Curpan, R. Popescu, I. Resmerita, M. C. Panzaru, L. I. Butnariu, E. V. Gorduza, M. Gramescu, C. Rusu. Wolf-hirschhorn syndrome: clinical and genetic study of 7 new cases, and mini review. Children. Vol. 8, pg. 751, 2021, https://doi.org/10.3390/children8090751.

20 A. Battaglia, J. C. Carey, S. T. South. Wolf-hirschhorn syndrome: a review and update. American Journal of Medical Genetics. Part C, Seminars in Medical Genetics. Vol. 169, pg. 216–223, 2015, https://doi.org/10.1002/ajmg.c.31449.

21 A. M. Paradowska-Stolarz. Wolf-hirschhorn syndrome (WHS) - literature review on the features of the syndrome. Advances in Clinical and Experimental Medicine: Official Organ Wroclaw Medical University. Vol. 23, pg. 485–489, 2014, https://doi.org/10.17219/acem/24111.

22 K. Nasri, N. Ben Jamaa, I. Ouertani, N. Boujelben. Partial trisomy 4p syndrome diagnosed prenatally. Fetal and Pediatric Pathology. Vol. 43, pg. 188–195, 2024, https://doi.org/10.1080/15513815.2023.2279138.

23 X. Zhang, H. Lu, H. Yang, Y. Ji, H. Liu, W. Liu, J. Li, Z. Yang, W. Sun. Genotype-phenotype correlation of deletions and duplications of 4p: case reports and literature review. Frontiers in Genetics. Vol. 14, pg. 1174314, 2023, https://doi.org/10.3389/fgene.2023.1174314.

24 J. Casper, M. L. Speir, B. J. Raney, G. Perez, L. R. Nassar, C. M. Lee, A. S. Hinrichs, J. N. Gonzalez, C. Fischer, M. Diekhans, H. Clawson, A. Benet-Pages, G. P. Barber, C. J. Vaske, M. J. van Baren, K. Wang, Y. J. P. Rodriguez, J. A. Jenkins-Kiefer, M. Chalamala, D. Haussler, W. J. Kent, M. Haeussler. The ucsc genome browser database: 2026 update. Nucleic Acids Research. Vol. 54, pg. D1331–D1335, 2026, https://doi.org/10.1093/nar/gkaf1250.

25 T. T. Schmidt, C. Tyer, P. Rughani, C. Haggblom, J. R. Jones, X. Dai, K. A. Frazer, F. H. Gage, S. Juul, S. Hickey, J. Karlseder. High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. Nature Communications. Vol. 15, Article 5149, 2024, https://doi.org/10.1038/s41467-024-48917-7.

26 W. W. Liao, M. Asri, J. Ebler, D. Doerr, M. Haukness, G. Hickey, S. Lu, J. K. Lucas, J. Monlong, G. Abel, S. Buonaiuto, X. H. Chang, H. Cheng, J. Chu, V. Colonna, J. M. Eizenga, X. Feng, C. Fischer, R. S. Fulton, S. Garg, C. Groza, A. Guarracino, W. T. Harvey, S. Heumos, K. Howe, M. Jain, T.-Y. Lu, C. Markello, F. J. Martin, M. W. Mitchell, K. M. Munson, M. N. Mwaniki, A. M. Novak, H. E. Olsen, T. Pesout, D. Porubsky, P. Prins, J. A. Sibbesen, J. Sirén, C. Tomlinson, F. Villani, M. R. Vollger, L. L. Antonacci-Fulton, G. Baid, C. A. Baker, A. Belyaeva, K. Billis, A. Carroll, P.-C. Chang, S. Cody, D. E. Cook, R. M. Cook-Deegan, O. E. Cornejo, M. Diekhans, P. Ebert, S. Fairley, O. Fedrigo, A. L. Felsenfeld, G. Formenti, A. Frankish, Y. Gao, N. A. Garrison, C. G. Giron, R. E. Green, L. Haggerty, K. Hoekzema, T. Hourlier, H. P. Ji, E. E. Kenny, B. A. Koenig, A. Kolesnikov, J. O. Korbel, J. Kordosky, S. Koren, H. Lee, A. P. Lewis, H. Magalhães, S. Marco-Sola, S. Marijon, A. McCartney, J. McDaniel, J. Mountcastle, M. Nattestad, S. Nurk, N. D. Olson, A. B. Popejoy, D. Puiu, M. Rautiainen, A. A. Regier, A. Rhie, S. Sacco, A. D. Sanders, V. A. Schneider, B. I. Schultz, K. Shafin, M. W. Smith, H. J. Sofia, A. N. Abou Tayoun, F. Thibaud-Nissen, F. F. Tricomi, J. Wagner, B. Walenz, J. M. D. Wood, A. V. Zimin, G. Bourque, M. J. P. Chaisson, P. Flicek, A. M. Phillippy, J. M. Zook, E. E. Eichler, D. Haussler, T. Wang, E. D. Jarvis, K. H. Miga, E. Garrison, T. Marschall, I. M. Hall, H. Li, B. Paten. A draft human pangenome reference. Nature. Vol. 617, pg. 312–324, 2023, https://doi.org/10.1038/s41586-023-05896-x.