DOCUMENT SUMMARY

This is the definitive *Science* paper from the Telomere-to-Telomere (T2T) Consortium announcing the first truly complete sequence of a human genome. This document is a cornerstone for Enlitens' mission, as it provides an unassailable, peer-reviewed critique of a long-standing "gold standard"—the GRCh38 human reference genome. The paper proves that the previous standard was missing a staggering 8% of the genome, leading to "tens of thousands of spurious variants" per analysis. It powerfully demonstrates how a flawed and incomplete reference can create systemic errors, false positives, and false negatives, providing a perfect biological analogy for the failures of standardized psychological testing. Crucially, the paper concludes by arguing that even a single complete genome is insufficient and that science must move towards a "pangenome" that embraces the full diversity of human variation, dismantling the "one normal reference" model at the heart of standardized assessments.

FILENAME

NURK_2022_Complete_sequence_of_a_human_genome_the_failure_of_the_gold_standard.md

METADATA

- **Primary Category**: ASSESSMENT
- **Document Type**: research_article
- **Relevance**: Core
- **Key Topics**: assessment_critique, standardized_testing, data_bias, neurodiversity, individual_differences, pangenome
- **Tags**: #goldstandard, #critique, #genomics, #T2T, #GRCh38, #bias, #falsenegative, #falsepositive, #pangenome, #assessment, #methodology

CRITICAL QUOTES FOR ENLITENS

"Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished."

"Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion base pair (bp) sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million bp of sequence containing 1,956 gene predictions, 99 of which are predicted to be protein coding."

"The GRCh38 reference assembly contains 151 Mbp of unknown sequence distributed throughout the genome, including pericentromeric and subtelomeric regions, recent segmental duplications, ampliconic gene arrays, and ribosomal DNA (rDNA) arrays, all of which are necessary for fundamental cellular processes (Fig. 1A)."

"In addition to these apparent gaps, other regions of GRCh38 are artificial or are otherwise incorrect."

"When compared to other human genomes, GRCh38 also shows a genome-wide deletion bias that is indicative of incomplete assembly (8)."

"The resulting T2T-CHM13 reference assembly removes a 20-year-old barrier that has hidden 8% of the genome from sequence-based analysis, including all centromeric regions and the entire short arms of five human chromosomes."

"Reanalysis of 3,202 short-read datasets from the 1KGP showed that T2TCHM13 simultaneously reduces both false-negative and false-positive variant calls due to the addition of 182 Mbp of missing sequence and the exclusion of 1.2 Mbp of falsely duplicated sequence in GRCh38."

"These improvements, combined with a lower frequency of rare variants and errors in T2T-CHM13, eliminate tens of thousands of spurious variants per 1KGP sample (25)."

"When mapping HiFi reads, absence of the additional FRGI paralogs in GRCh38 causes their reads to incorrectly align to FRGIDP resulting in many false-positive variants (Fig. 5B)."

"Any variants within these paralogs, and others like them, will be overlooked when using GRCh38 as a reference."

"This 8% of the genome has not been overlooked due to its lack of importance, but rather due to technological limitations."

"Although CHM13 represents a complete human haplotype, it does not capture the full diversity of human genetic variation."

"To address this bias, the Human Pangenome Reference Consortium (HPRC) (59) has joined with the T2T Consortium to build a collection of high-quality reference haplotypes from a diverse set of samples."

"Until this goal is realized, and any human genome can be completely sequenced without error, the T2T-CHM13 assembly represents a more complete, representative, and accurate reference than GRCh38."

KEY STATISTICS & EVIDENCE

- **The Missing 8%**: The new T2T-CHM13 genome completes the 8% of the human genome that was previously unfinished.
- **New Sequence Added**: The complete genome introduces nearly 200 million base pairs of new sequence.
- **New Gene Predictions**: The new sequence contains 1,956 gene predictions, of which 99 are predicted to be protein-coding.

- **Errors in the Old Standard**: The previous reference, GRCh38, contains 151 Mbp of unknown sequence, has a genome-wide deletion bias, and includes regions that are artificial or incorrect.
- **Superiority in Variant Calling**: T2T-CHM13 is nine times more predictive of copy number in segmentally duplicated regions of the genome compared to GRCh38.
- **Comparison of GRCh38 and T2T-CHM13v1.1 assemblies**: | Summary | GRCh38 | T2T-CHM13 | % | | :--- | :--- | :--- | :--- | | **Assembled bases (Gbp)** | 2.92 | 3.05 | +4.5% | | **Gap bases (Mbp)** | 120.31 | 0 | -100.0% | | **# Contigs** | 949 | 24 | -97.5% | | **# Issues** | 230 | 46 | -80.0% | | **Issues (Mbp)** | 230.43 | 8.18 | -96.5% | | **# Genes** | 60,090 | 63,494 | +5.7% | | **# Exclusive genes** | 263 | 3,604 | | | **% SDs** | 5.00% | 6.61% | | | **Satellite bases (Mbp)** | 76.51 | 150.42 | +96.6% |

METHODOLOGY DESCRIPTIONS

Flaws in the Previous Standard's Construction

"Unlike the competing Celera effort (3) and most modern sequencing projects based on "shotgun" sequence assembly (4), the GRC assembly was constructed from sequenced bacterial artificial chromosomes (BACs) that were ordered and oriented along the human genome via radiation hybrid, genetic linkage, and fingerprint maps." "However, limitations of BAC cloning led to an underrepresentation of repetitive sequences, and the opportunistic assembly of BACs derived from multiple individuals resulted in a mosaic of haplotypes." "As a result, several GRC assembly gaps are unsolvable due to incompatible structural polymorphisms on their flanks, and many other repetitive and polymorphic regions were left unfinished or incorrectly assembled (5)."

The Need for Novel Technology

"Despite finishing efforts from both the Human Genome Project (9) and GRC (1) that improved the quality of the reference, there was limited progress towards closing the remaining gaps in the years that followed (Fig. ID)." "Long-read shotgun sequencing overcomes the limitations of BAC-based assembly and bypasses the challenges of structural polymorphism between genomes." "However, the high error rate (>5%) of these technologies posed challenges for the assembly of long, near-identical repeat arrays." "To finish the last remaining regions of the genome, we leveraged the complementary aspects of PacBio HiFi and Oxford Nanopore ultra-long read sequencing to assemble the uniformly homozygous CHM13hTERT cell line (hereafter, CHM13) (17)."

PRACTICAL APPLICATIONS

A Case Study in Diagnostic Error: The FRG1 Gene

This example provides a perfect, concrete case of how an incomplete reference standard (GRCh38) leads directly to analytical errors, analogous to a misdiagnosis. "The T2T-CHM13 assembly reveals 23 paralogs of FRGI spread across all acrocentric chromosomes as well as chromosomes 9 and 20 (Fig. 5A)." "However, only 9 FRGI paralogs are found in GRCh38, hampering sequence-based analysis." "When mapping HiFi reads, absence of the additional FRGI paralogs in GRCh38 causes their reads to incorrectly align to FRGIDP resulting in many false-positive variants (Fig. 5B)." "When mapped to CHM13, HiFi reads show the expected coverage and a typical heterozygous variation pattern for the three non-CHM13 samples (variants >20% coverage shown)." "Any variants within these paralogs, and others like them, will be overlooked when using GRCh38 as a reference." "CHM13 copy number resembles all samples from the SGDP, whereas GRCh38 underrepresents the true copy number."

The Future is a Diverse "Pangenome," Not a Single Reference

This section is a powerful argument against the concept of a single standard and for embracing diversity, which is the core of the neurodiversity paradigm. "The T2T-CHM13 assembly adds five full chromosome arms and more additional sequence than any genome reference release in the past 20 years (Fig. 1D)." "This 8% of the genome has not been overlooked due to its lack of importance, but rather due to technological limitations." "High accuracy long-read sequencing has finally removed this technological barrier, enabling comprehensive studies of genomic variation across the entire human genome, which we expect to drive future discovery in human genomic health and disease." "Such studies will necessarily require a complete and accurate human reference genome." "Although CHM13 represents a complete human haplotype, it does not capture the full diversity of human genetic variation." "To address this bias, the Human Pangenome Reference Consortium (HPRC) (59) has joined with the T2T Consortium to build a collection of high-quality reference haplotypes from a diverse set of samples." "Ideally, all genomes could be assembled at the quality achieved here, but automated T2T assembly of diploid genomes presents a difficult challenge that will require continued development." "Until this goal is realized, and any human genome can be completely sequenced without error, the T2T-CHM13 assembly represents a more complete, representative, and accurate reference than GRCh38."