RESEARCH SUMMARY

Han1: an algorithmic path to a complete annotated genome

By Kuan-Hao Chao

Mar 1, 2023 · 9 min read · Updated Jul 1, 2026

Abstract

Han1 follows a staged finished-genome workflow: long-read assembly, reference-guided chromosome scaffolding, layered gap closing, manual repeat repair, polishing, and two-pass annotation lift-over. The resulting assembly was a gap-free, reference-quality, fully annotated genome from a Southern Han Chinese individual, enabling gene-level comparison between two finished, annotated individual human genomes.

For more than twenty years, much of human genomics has relied on a single reference assembly. That reference — currently GRCh38 — is extraordinarily useful, but it has two important limitations. It is a mosaic assembled from multiple individuals rather than one person's genome, and it retains hundreds of gaps, including centromeric and acrocentric regions (the latter on chromosomes whose centromere sits very near one end). Mapping data to it therefore means comparing each sample with a composite reference that is incomplete and draws most of its sequence from people of European ancestry.

In 2022 that changed, at least in part. The Telomere-to-Telomere consortium published T2T-CHM13 (Nurk et al., 2022), the first gapless sequence of a human genome, completing centromeres and other regions missing from GRCh38. It was a landmark. But CHM13 represented one northern European female sample, and the Y chromosome used alongside that assembly came from a second individual, HG002. One complete genome is a beginning, not an end: humanity is not one genome.

That paper became the spark for my first Ph.D. project. Steven Salzberg's idea was direct: if a complete human genome could now be built, we should assemble and annotate another gapless, reference-quality individual genome from a different genetic background. I went through the available samples, kept coming back to HG00621, and stepped forward to take on the project.

Han1 (Chao et al., 2023) became our answer: the first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. Just as important, it was an algorithmic test case for how to turn one person's long-read data into a finished, annotated genome that can be compared gene by gene with another finished genome.

Why HG00621?

HG00621 was recorded as Southern Han Chinese. That mattered to me personally because it is my own ethnicity, and it made the annotation question feel concrete: what would a finished, annotated genome from this background look like, including gene copy-number and coding differences relative to T2T-CHM13?

It also made sense scientifically. The Han Chinese are the largest ethnic group on Earth — roughly 1.4 billion people — yet there was no complete, gap-free, fully annotated Han Chinese individual genome. Earlier Chinese and Han Chinese assemblies existed, including YH, HX1, NH1.0, and HJ-H1/HJ-H2, but the paper's comparison showed that they still contained gaps and none were fully annotated. HG00621 also broadened the set of genomes our group had already been assembling and annotating: it was relatively different from the Ashkenazi and Puerto Rican samples we had worked on, rather than another nearby data point.

Finally, it was practical. HG00621 had the kind of public data needed for a serious finishing attempt: high-accuracy PacBio HiFi reads, Oxford Nanopore ultralong reads, and supporting read data for checking difficult calls. No single genome can represent an entire population, and Han1 should not be read that way. The goal was narrower and more concrete: build a finished genome from HG00621, a Southern Han Chinese male from Fujian Province, so that studies of human variation would have one more complete, annotated individual genome rather than another fragmented draft.

The algorithm we built

We started from public Human Pangenome Reference Consortium data for HG00621 (Liao et al., 2023). The raw material was two complementary kinds of long reads: PacBio HiFi reads at 39.45x coverage, which are long and highly accurate, and Oxford Nanopore ultralong reads at 34.75x coverage, which span difficult repeats but are noisier. Accuracy and length each solve what the other cannot.

The workflow was deliberately staged:

1. Assemble contigs (continuous stretches of sequence) de novo — from scratch, without copying a reference. We assembled the HiFi reads with hifiasm (Cheng et al., 2021) and the Nanopore reads with Flye (Kolmogorov et al., 2019), then chose the hifiasm assembly as the backbone because it was much more contiguous and accurate: 182 contigs, a 95.8 Mb contig N50 (half the assembly lies in contigs at least this long), and QV 57.8 (a log-scaled base-accuracy score, higher being fewer errors).
2. Order those contigs into chromosomes. We used the MaSuRCA chromosome scaffolder (Zimin et al., 2013) with T2T-CHM13 as the guide, split 12 misassembled contigs, and produced chromosome scaffolds for chromosomes 1–22, X, and Y.
3. Close the remaining gaps in layers. HiFi reads closed 24 gaps; Nanopore-derived contigs closed 19 more; CHM13 sequence closed 45 of the remaining 58. The final unresolved cases required manual inspection of pericentromeric (near the centromere) repeat misassemblies and redundant haplotype contigs.
4. Clean up and polish. We removed an Epstein-Barr virus contaminant contig, treated the remaining unplaced contigs as redundant haplotype sequence, extracted a single mitochondrial genome, and

polished the chromosomes with JASPER.

The final Han1 assembly contains 3,099,707,698 bases in 25 sequences: chromosomes 1–22, X, Y, and the mitochondrial genome. The autosomes, sex chromosomes, and mitochondrial genome were assembled end to end with no gaps, and every chromosome is a single contig. About 120 Mb was filled with T2T-CHM13 sequence, concentrated in the hardest repeat-rich regions, so the finished assembly is mostly HG00621 sequence but not purely de novo sequence.

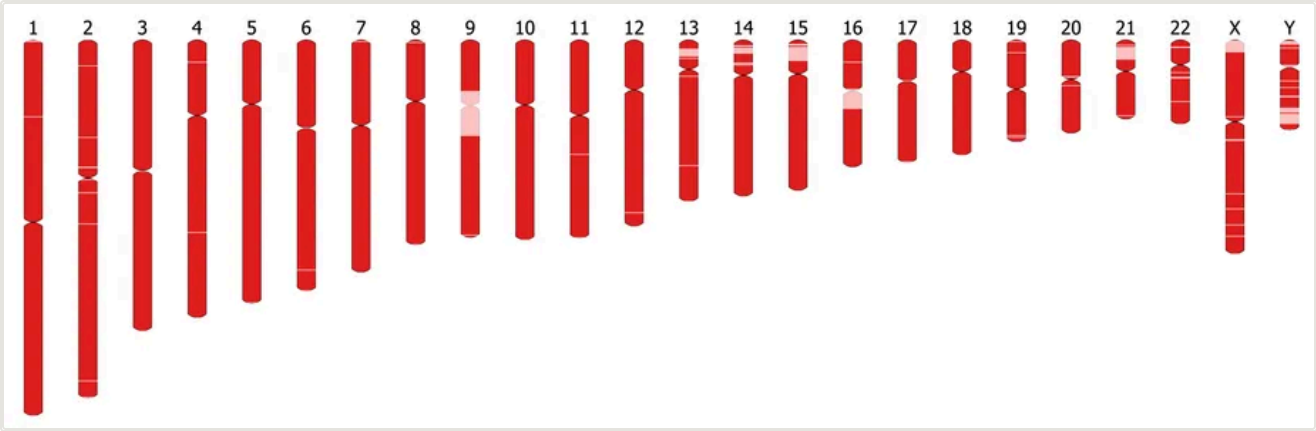

Figure 1. The Han1 nuclear chromosomes. Red marks sequence assembled from HG00621; light pink marks sequence inserted from T2T-CHM13 during gap closing. The final assembly has 25 sequences including the mitochondrial genome, while this figure shows the 24 nuclear chromosomes.

A finished sequence is not yet a finished reference. The next algorithmic step was annotation: moving the T2T-CHM13 RefSeq-derived annotation (O'Leary et al., 2016) onto Han1 with Liftoff (Shumate and Salzberg, 2021). Most genes could be lifted directly, but the ribosomal DNA arrays needed special handling because they occur in long, near-identical tandem arrays on the acrocentric chromosomes. We therefore used a two-pass process: first mask candidate rDNA arrays and map the non-rDNA genes, then lift the rDNA units separately under a stricter copy-aware condition and merge the results. In the end Han1 carried 60,708 putative genes, including 20,003 protein-coding genes.

That annotation problem became one of the threads that later led me to LiftOn: finishing a genome is not only about removing gaps from the sequence, but also about carrying a trustworthy gene map onto the new assembly.

What it showed

Two finished human genomes, side by side for the first time. With Han1 complete and annotated, we could line it up against T2T-CHM13 — two finished individual human genomes compared gene by gene rather than only against a gappy composite. The two are highly collinear: chromosome for chromosome, the sequences run in near-diagonal agreement, and in the gene-order plot 59,654 of 60,746 genes, about 98.2%, had sequence identity above 95%.

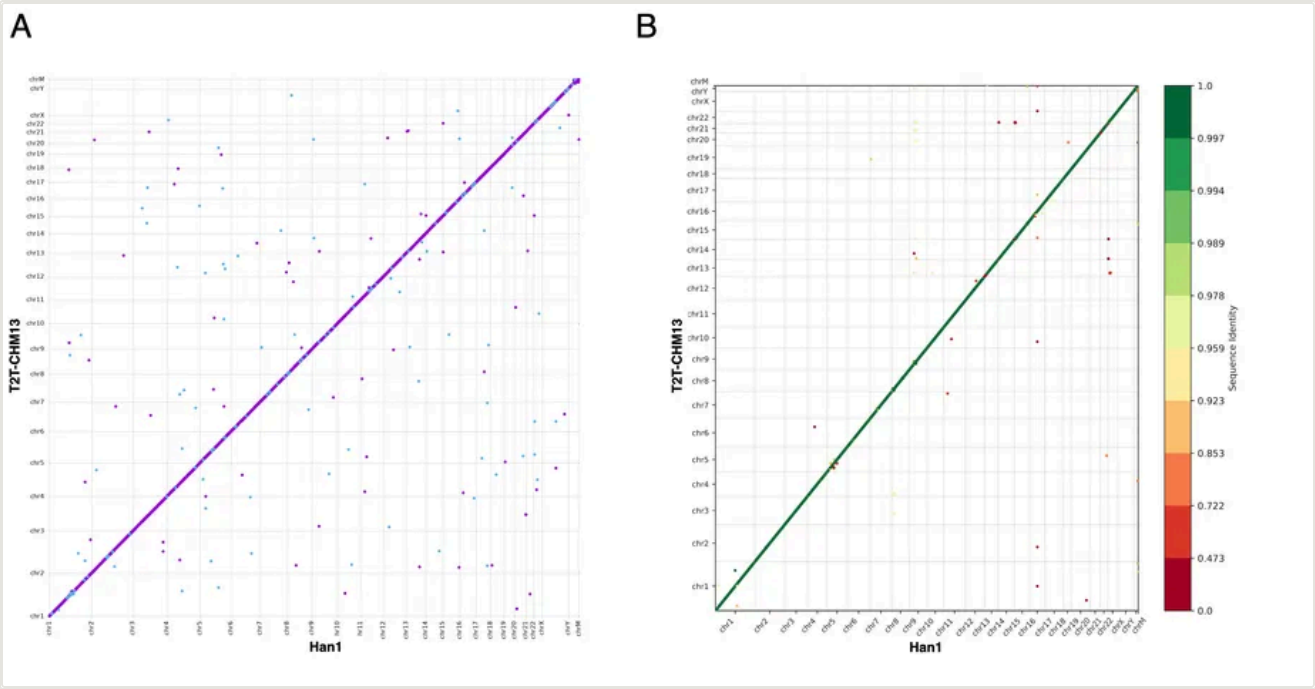

Figure 2. Han1 versus T2T-CHM13. **(A)** A whole-genome dot plot: the tight diagonal shows that the two genomes are highly collinear, with purple marking same-orientation alignments and blue marking inversions. **(B)** A gene-order plot, with genes numbered along both genomes and colored by sequence identity. Most genes retain both order and high sequence identity.

But "highly similar" is not "identical," and the differences are where annotation matters. LiftoffTools compared 181,029 transcripts between T2T-CHM13 and Han1. At the gene level, we focused on 235 protein-coding genes with major predicted coding differences: frameshifts (insertions or deletions that shift the reading frame), truncations, start-codon changes, or premature stop codons. The raw-read check then narrowed the claim. Many apparent disruptive differences were heterozygous, where the other haplotype likely carried an unaffected copy. After filtering for homozygous non-SNP (single-nucleotide polymorphism) mutations in coding regions, the paper identified 54 distinct mutation sites in 46 distinct genes.

That is a much more careful conclusion than treating all 235 target genes as fixed disruptive differences. Among the 46 genes, many were hypothetical proteins, olfactory receptors, or VDJ segments (immune-receptor gene segments), where copy number and annotation are difficult. The protein-level follow-up found two protein-coding genes that appeared severely truncated in Han1 relative to RefSeq and five genes where Han1 was better conserved than T2T-CHM13. The point is not that one genome is better than the other; the point is that finished sequence plus finished annotation lets us ask this question at gene resolution.

A structural difference you can see. The clearest example is on chromosome 8, where Han1 and T2T-CHM13 carry a 4.1 Mb segment in opposite orientations. In Han1 the interval is chr8:7,306,407–11,410,283; in T2T-CHM13 it corresponds to chr8:11,717,452–7,617,994. This is one of the largest common

inversion polymorphisms in humans and includes the β-defensin gene cluster, a structurally dynamic immune-related locus.

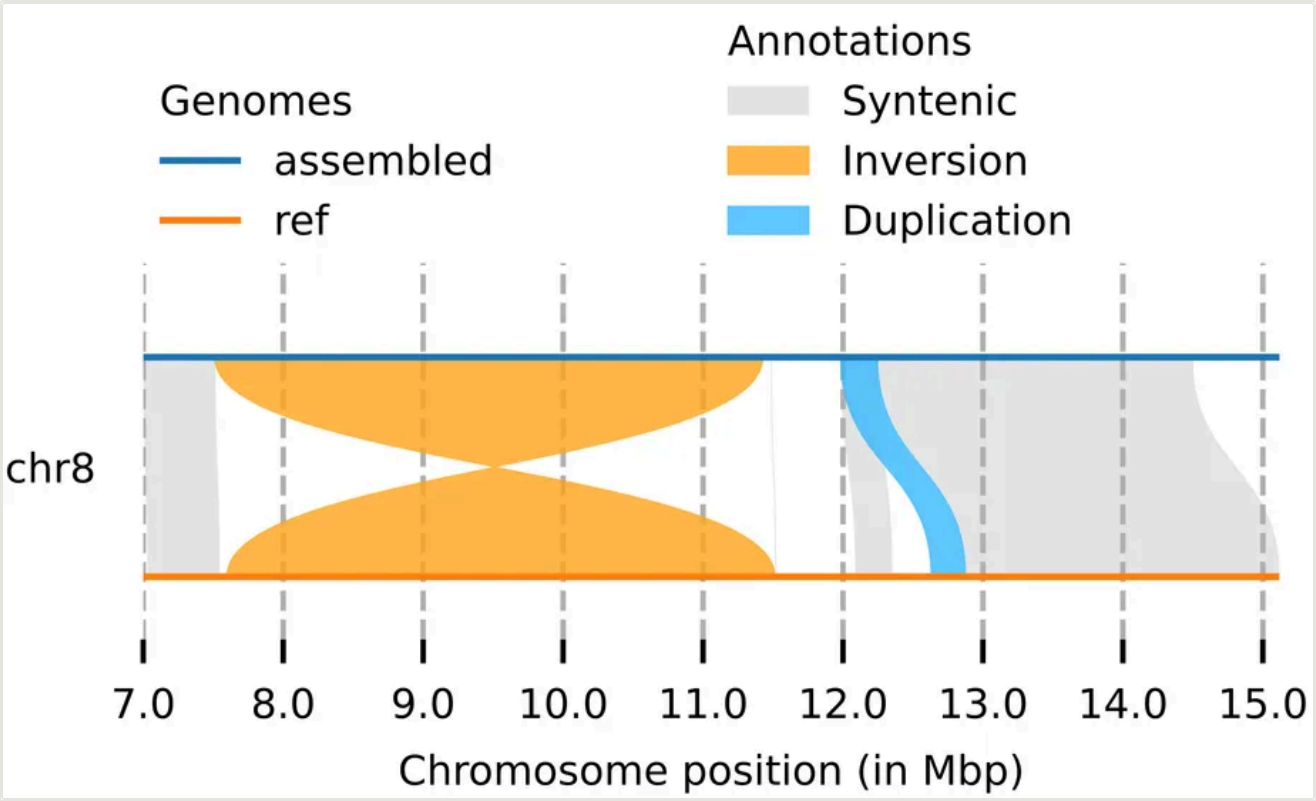

Figure 3. A 4.1 Mb inversion on chromosome 8. Across the β-defensin gene cluster, Han1 (top) and T2T-CHM13 (bottom) are inverted relative to each other (orange), flanked by syntenic sequence (gray). Han1's orientation matches the older GRCh38 reference — the two complete genomes simply happen to carry opposite versions of a common human polymorphism.

Han1's orientation in this region matches GRCh38, while T2T-CHM13 carries the other arrangement. Neither orientation is the universal answer; they are alternative forms present in human genomes. Han1 also surfaced smaller differences: 94 near-identical gene clusters showed copy-number changes relative to T2T-CHM13, and the assembly contained two nuclear mitochondrial sequences (NUMTs — fragments of mitochondrial DNA embedded in the nuclear genome). One NUMT on chromosome 13 was present in T2T-CHM13 but absent from GRCh38; a longer NUMT on chromosome 20 was unique to Han1 and supported by Nanopore reads.

The bigger picture

No single genome can represent a species of eight billion people, and no single Han Chinese genome can represent 1.4 billion Han Chinese people. T2T-CHM13 proved that a human genome could be finished. Han1 showed how the same standard could be extended to a Southern Han Chinese individual: not just a long-read assembly, but a gap-free, polished, annotated genome with enough structure to support gene-level comparison.

That is the algorithmic lesson I take from the project. The scientific value came from the whole chain: assemble, scaffold, close, curate, polish, annotate, and compare. If any step is weak, the final biological claims become weaker too. When every step is explicit and checked, a new individual genome becomes a usable reference resource and a basis for asking where two finished human genomes truly differ.

Han1 is freely available, assembly and annotation both, for anyone to build on.

Read the paper in G3: Genes|Genomes|Genetics, or get the assembly and annotation from GitHub (also on GenBank, accession JANJEX000000000). Han1 was built with Aleksey Zimin, Mihaela Pertea, and Steven Salzberg at Johns Hopkins.

References

1. Nurk, S. et al. The complete sequence of a human genome. Science (2022). doi:10.1126/science.abj6987
2. Chao, K.-H., Zimin, A. V., Pertea, M., and Salzberg, S. L. The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. G3: Genes|Genomes|Genetics (2023). doi:10.1093/g3journal/jkac321
3. Liao, W.-W. et al. A draft human pangenome reference. Nature (2023). doi:10.1038/s41586-023-05896-x
4. Cheng, H. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods (2021). doi:10.1038/s41592-020-01056-5
5. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. Nature Biotechnology (2019). doi:10.1038/s41587-019-0072-8
6. Zimin, A. V. et al. The MaSuRCA genome assembler. Bioinformatics (2013). doi:10.1093/bioinformatics/btt476
7. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research (2016). doi:10.1093/nar/gkv1189
8. Shumate, A. and Salzberg, S. L. Liftoff: accurate mapping of gene annotations. Bioinformatics (2021). doi:10.1093/bioinformatics/btaa1016
