RESEARCH SUMMARY

A complete diploid human genome: the T2T-HG002 benchmark and its gene annotation

By Kuan-Hao Chao

Aug 6, 2026 · 9 min read · Updated Aug 8, 2026

Abstract

The Telomere-to-Telomere Consortium's Q100 project finished both parental copies of the HG002 genome and released them as a benchmark free of detectable errors across 99.4% of 6.0 Gb, adding 701.4 Mb of autosomal sequence and both sex chromosomes that no earlier benchmark covered. Alongside the sequence came a diploid gene annotation covering 39,144 protein-coding genes across the two haplotypes. It is built on the two-pass lift-over backbone I developed for Han1 — mask the ribosomal DNA arrays, lift the gene backbone onto the masked assembly, then lift the rDNA units separately and merge — which I also ran on the v1.0.1 HG002 assembly. My contribution to the released annotation was the extra-copy search that recovered 51 gene copies the DNA lift had missed, 30 on the maternal haplotype and 21 on the paternal.

You have two genomes, not one — one from your mother, one from your father, differing at almost every position. Yet nearly every tool we use flattens them back into one: we map reads to a single reference, call the places where they disagree, and hand the result to a clinician as a list of variants.

That works until it doesn't. Reads from a duplicated or structurally polymorphic region often cannot be placed on the reference at all, so those regions produce no calls — not "no variants," but no information. The benchmarks we grade callers against were built the same way, and are blind in the same places.

The complete diploid HG002 benchmark (Hansen et al., 2026) is the way out: the Telomere-to-Telomere Consortium's Q100 project, 65 authors, delivering both haplotypes of one person, finished and annotated. My own work is on the annotation — the two-pass lift-over backbone it is built on, and the extra-copy search that recovered 51 gene copies the DNA lift had missed.

Why a genome benchmark and not another variant list

The Genome in a Bottle consortium made HG002 the most characterized sample in human genomics, and its variant benchmarks (Zook et al., 2019) have been the yardstick for a decade. But they come with a confidence mask: GIAB v4.2.1 excluded both sex chromosomes and about 12% of the autosomes — disproportionately the complex, medically interesting part, where you then cannot measure at all.

T2T-HG002 removes the mask. It is free of detectable errors across 99.4% of the diploid genome and adds 701.4 Mb of autosomal sequence plus both sex chromosomes at 216.8 Mb — 15.3% no previous benchmark assessed. On one published HG002 assembly, comparing against the complete genome found 66,544 errors — only 30,505 of which fell inside the regions the variant benchmark could assess.

How the genome was finished

HG002 is a good choice partly for reasons unrelated to sequencing: the cell line is openly available, the data fully open, and both parents were sequenced. The assembly draws on about 170× PacBio HiFi and 209× Oxford Nanopore ultra-long reads, built with Verkko (Rautiainen et al., 2023), which resolves the haplotypes using parental k-mers.

The remaining uncertainty is concentrated, not diffuse. Merqury (Rhie et al., 2020) quality rose across releases from Q63.1 to Q68.9, roughly one error per 8 million bases. What is left is 3,172 low-confidence regions covering 38.8 Mb, 86.3% of it ribosomal DNA — tandem arrays of near-identical 45 kb units, hard to reconstruct and harder to prove.

Annotating both haplotypes

A finished sequence is not yet a usable reference: somebody has to say where the genes are, and for a diploid genome, twice. The places where the haplotypes agree are the easy 99%; the value lives where they differ — in the segmental duplications and tandem arrays hardest to annotate.

The pipeline

The reference is the T2T-CHM13v2.0 annotation (Nurk et al., 2022), derived from RefSeq (O'Leary et al., 2016). The obstacle is that a complete genome actually contains the repeats a draft leaves as gaps, and an alignment-based transfer will scatter spurious copies across them.

The backbone that solves this is the one I built for Han1. Annotating that genome (Chao et al., 2023) I hit the same wall, and the answer was a two-pass lift: locate the rDNA arrays by aligning a reference 45S unit against the assembly and mask them, lift the gene backbone onto the masked sequence, then lift the rDNA units separately under a copy-aware condition and merge the two. Masking first is what stops the arrays from absorbing spurious copies; lifting them separately is what gets the real ones back. The HG002 annotation is built on that structure.

I also ran it on HG002 directly: in 2024, on the v1.0.1 assembly, masking 20 rDNA intervals on the maternal haplotype and 46 on the paternal and lifting the backbone across both, mapping 54,763 and 53,106 genes. That run was superseded when the assembly moved to v1.1.

Hyun Joo Ji led the released pipeline, which extends that backbone. Masking covers the V(D)J segments — TRA, TRB, TRG, IGL, IGH and IGK — as well as the 45S arrays, of which 51 copies were found on the maternal haplotype and 81 on the paternal. The first pass runs Liftoff (Shumate and Salzberg, 2021) with copy search at 95% identity; the second lifts 219 rDNA units, wrapped in synthetic parent features so that all four rRNA genes travel together, recovering 21 arrays on the maternal haplotype and 28 on the paternal. Merged, that gives 59,525 genes on the maternal haplotype and 57,882 on the paternal. She

added two stages with no equivalent in Han1: a repair pass that trims 586 and 534 CDS features back to codon structure and extends a few hundred more to their next in-frame stop, and an assembly-specific identifier scheme.

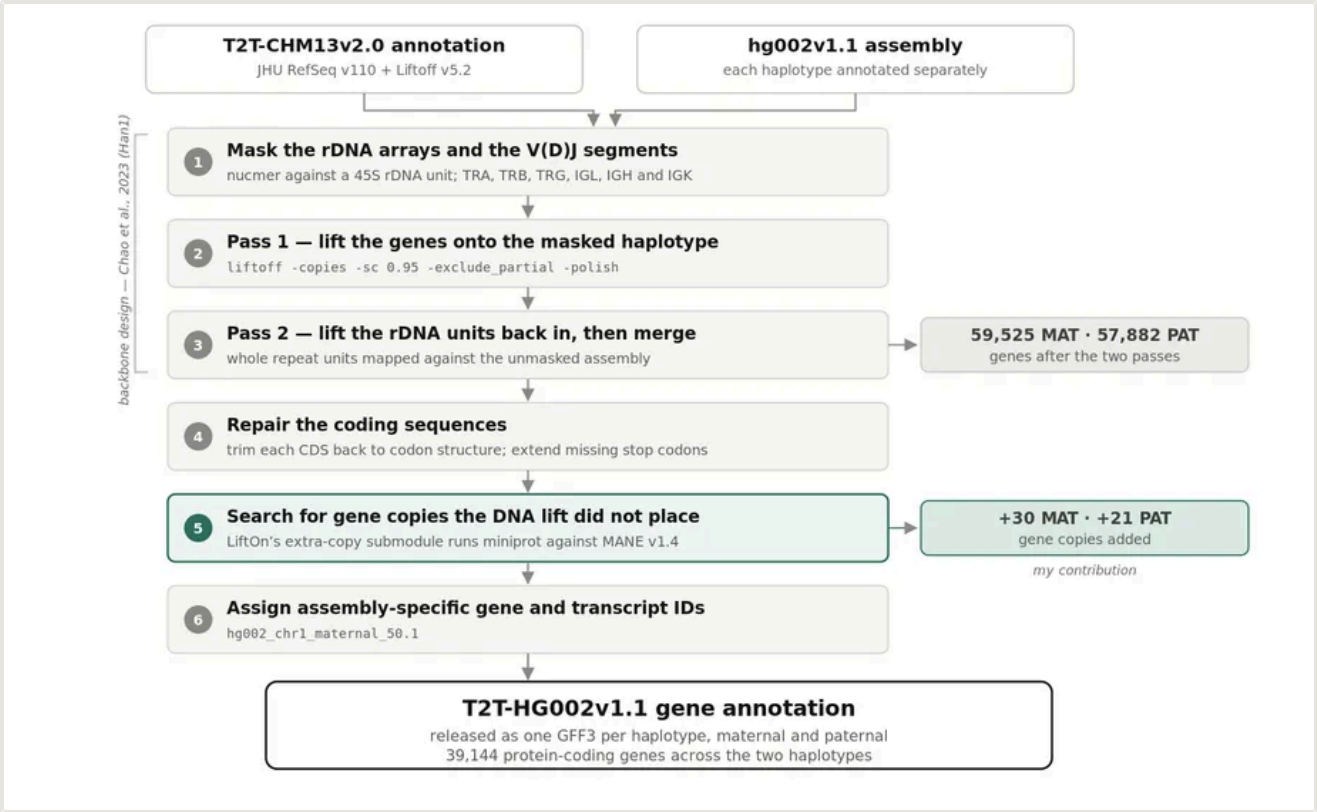


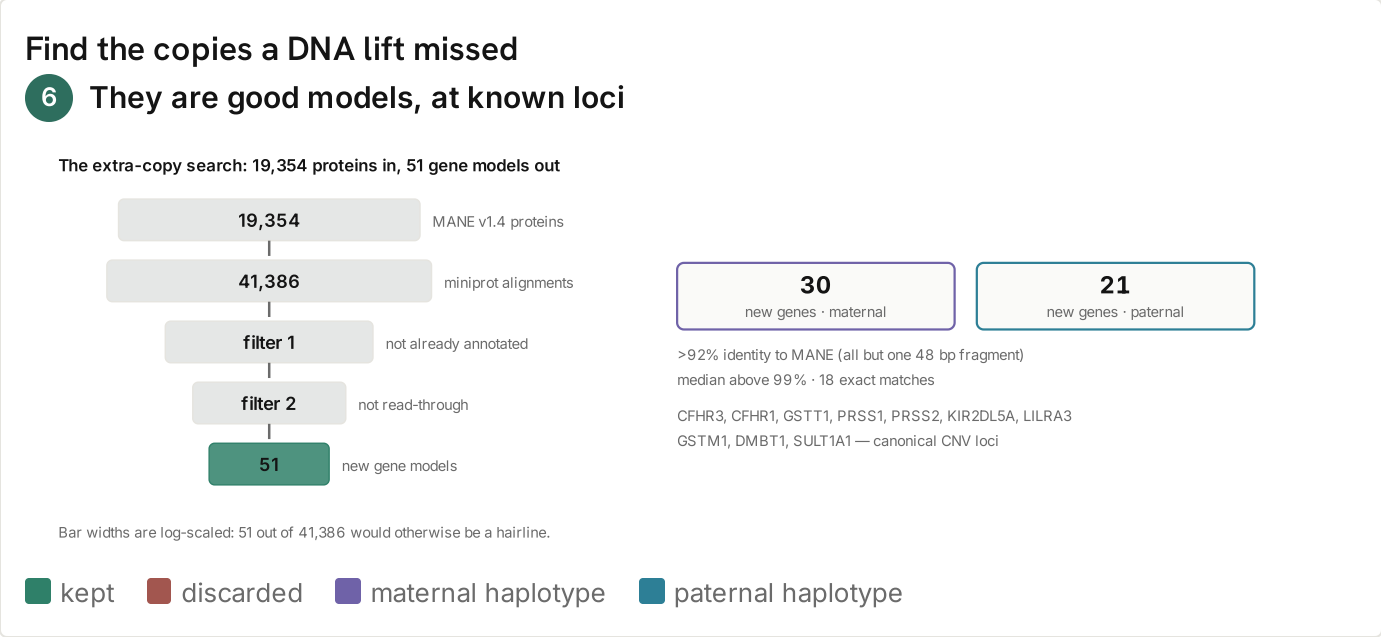
Figure 1. How each HG002 haplotype was annotated. Steps **1–3** are the two-pass backbone I developed for Han1 — mask the repeats, lift the genes, lift the rDNA units back in. Step **5**, in teal, is the extra-copy search I contributed to the released annotation.

The copies a DNA lift misses

That step was mine. DNA-based lift-over undercounts gene copies, and a complete diploid genome is exactly where the missing ones are. Liftoff does search for extra copies, but one that has diverged, or sits inside a segmental duplication where the alignment is ambiguous, can still be dropped — and this hits hardest the genes that vary in copy number between people. Protein alignment fails differently — an orthologous protein stays recognizable across far more divergence than its DNA. That is the premise behind LiftOn (Chao et al., 2025), and I used the part of it aimed at this problem, its extra-copy search submodule.

The search is deliberately blunt; the filter is where the work is. The protein set was MANE v1.4 (Morales et al., 2022), one representative transcript per protein-coding gene. Those 19,354 proteins were aligned to the whole diploid assembly in a single miniprot (Li, 2023) run, producing 41,386 alignments from 19,351 of them. Almost all land on genes we already have. An alignment was discarded if it overlapped 10% or more of an existing gene, or spanned more than two adjacent gene loci — the first finds unannotated space, the second removes read-through alignments across gene families. Survivors were rebuilt into full gene–transcript–exon records.

The whole search is five stages wide at the top and two gene models wide at the bottom, and the interesting part is which alignments each filter throws away.



That procedure added 30 genes to the maternal haplotype and 21 to the paternal, and they are good models rather than marginal alignments: all but a single 48 bp fragment exceeded 92% identity to their MANE reference protein, the median was above 99%, and 18 matched exactly. They remain identifiable in the released annotation by their `miniprot` source column.

They also show up in the paper's headline count — the cleanest check I know that a contribution landed. Table 1 reports 19,117 protein-coding genes on the maternal autosomes; counting the released annotation, Liftoff accounts for 19,088, and the remaining 29 are the autosomal copies from this pass.

The list of genes is what convinced me the filter was doing real biology. CFHR3 and CFHR1, GSTT1, PRSS1 and PRSS2, KIR2DL5A, LILRA3 and several olfactory receptors on the maternal side; GSTM1, DMBT1, SULT1A1 on the paternal. That is close to a list of the canonical human copy-number-variable loci — and HG002 turns out to carry GSTT1 only on the maternal haplotype and GSTM1 only on the paternal. Those calls were made downstream on the merged annotation, so they are not independent confirmation, but they are why the copies were worth recovering.

What the annotation shows

Both haplotypes carry a nearly complete gene complement, and the interesting numbers are the differences. Adding the paternal autosomes (19,084) and the two sex chromosomes (ChrX at 838 on the maternal haplotype, ChrY at 105 on the paternal) to that maternal figure gives 39,144 protein-coding genes across the diploid genome, and 60,074 distinct genes in total — 2,755 more than T2T-CHM13.

Fourteen autosomal genes appear only on the maternal haplotype and twelve only on the paternal — including a DUSP22 copy in the maternal Chr16 pericentromere, present on neither the paternal haplotype nor GRCh38.

Where the annotation breaks, it breaks in duplications. Counting a gene broken if its MANE transcript has an invalid start or stop codon, a premature in-frame stop, or a protein under 80% identical to the reference, 129 are broken on the maternal haplotype and 121 on the paternal — and 28% of those 250 intersect a segmental duplication, against 8% of genes that map cleanly.

The bigger picture

Annotation is where a benchmark stops being a sequence and becomes a person. A genome matters because it can be read as a statement about someone's biology — which means knowing which stretches are genes, and which copy came from which parent. A diploid annotation keeps that distinction; a haploid one averages it away, hardest on the copy-number-variable loci that reference-based pipelines already handle worst.

The paper's practical verdict points the same way: assembling HG002 de novo reaches 99.94% coverage at QV 51, where reconstructing it from variant calls manages roughly 93–98% at QV 35–40. The limits are real: a region no technology reads well is one you cannot easily prove you got right.

What the project changes is the target. For twenty years the goal of human resequencing has been a good list of differences from a reference; this says the goal should be the genome itself. That is the thread running through Han1 and LiftOn and into this: carrying a trusted gene map onto a more complete genome, and keeping the differences rather than smoothing them away.

Read the paper in Cell, or get the assembly and annotation from GitHub. Hyun Joo Ji's annotation pipeline is public; my part was the backbone and the extra-copy search described above, with Steven Salzberg at Johns Hopkins.

References

1. Hansen, N. F. et al. A complete diploid human genome benchmark for personalized genomics. Cell (2026). doi:10.1016/j.cell.2026.06.016
2. Nurk, S. et al. The complete sequence of a human genome. Science (2022). doi:10.1126/science.abj6987
3. Zook, J. M. et al. An open resource for accurately benchmarking small variant and reference calls. Nature Biotechnology (2019). doi:10.1038/s41587-019-0074-6
4. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nature Biotechnology (2023). doi:10.1038/s41587-023-01662-6
5. Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology (2020). doi:10.1186/s13059-020-02134-9

6. Shumate, A. and Salzberg, S. L. Liftoff: accurate mapping of gene annotations. Bioinformatics (2021). doi:10.1093/bioinformatics/btaa1016

7. Li, H. Protein-to-genome alignment with miniprot. Bioinformatics (2023). doi:10.1093/bioinformatics/btad014

8. Chao, K.-H., Mao, A., Salzberg, S. L., and Pertea, M. Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Research (2025). doi:10.1101/gr.279620.124

9. Morales, J. et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature (2022). doi:10.1038/s41586-022-04558-8

10. 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

11. 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

SHARE X Bluesky LinkedIn Copy link Subscribe