RESEARCH SUMMARY

LiftOn: combining DNA and protein evidence for genome annotation

By Kuan-Hao Chao

Feb 1, 2025 · 12 min read · Updated Jul 1, 2026

Abstract

LiftOn addresses a practical failure mode in annotation transfer: DNA alignments preserve gene structure near the reference, while protein alignments stay informative across larger evolutionary distances but lose noncoding context. The post explains how LiftOn reconciles Liftoff and miniprot annotations with a two-step protein-maximization algorithm and what the paper showed across human, same-species, and cross-species annotation transfers.

Genome assembly has accelerated faster than genome annotation. Public databases now contain thousands of eukaryotic assemblies, but many lack a comparably complete map of their genes and other biological features. A bare assembly is a long string of A, C, G, and T; the annotation identifies where genes begin and end, how transcripts are structured, and which regions encode proteins.

The pragmatic way to draw that map is to borrow one. If a closely related genome is already well annotated — say, the human reference — you can lift its genes over to the new assembly: line the two genomes up and carry each gene across to its matching location. Done well, annotation lift-over is fast, reproducible, and far cheaper than annotating from scratch. Done badly, it quietly fills your new genome with broken genes.

LiftOn (Chao et al., 2025) is the algorithm we built for that problem: use DNA and protein as complementary evidence, then construct the gene model that best preserves the reference protein.

Why not just lift over the DNA?

My lab already had a DNA-based lift-over tool: Liftoff (Shumate and Salzberg, 2021), written by Alaina Shumate and Steven Salzberg. Liftoff maps annotation hierarchies between assemblies and performs especially well for closely related genomes. As evolutionary distance increases, however, nucleotide alignments become harder to interpret and can yield misplaced splice sites or disrupted open reading frames.

Protein alignment provides complementary evidence. Orthologous protein sequences can remain conserved across distances where their nucleotide sequences have diverged, in part because synonymous substitutions do not change amino acids. Heng Li's miniprot (Li, 2023) aligns a reference protein directly to a target genome and can recover coding models across relatively distant species.

Protein alignment has its own limitations. It does not define untranslated regions; it can miss very small exons; it is susceptible to processed pseudogenes (intron-less, reverse-transcribed gene copies); and it can merge nearby members of tandem gene families. Neither evidence source solves the full annotation-transfer problem alone.

The insight behind LiftOn is simple: **DNA alignment and protein alignment fail in different ways.** Where one is weak, the other is often strong. So instead of choosing one source, LiftOn pairs them at the same locus, lets each contribute the coding segments it supports best, and uses the reference protein as the scoring target. I built LiftOn on that idea in Steven Salzberg's and Mihaela Pertea's groups at Johns Hopkins — the same place Liftoff was born, with Liftoff's own author on the team.

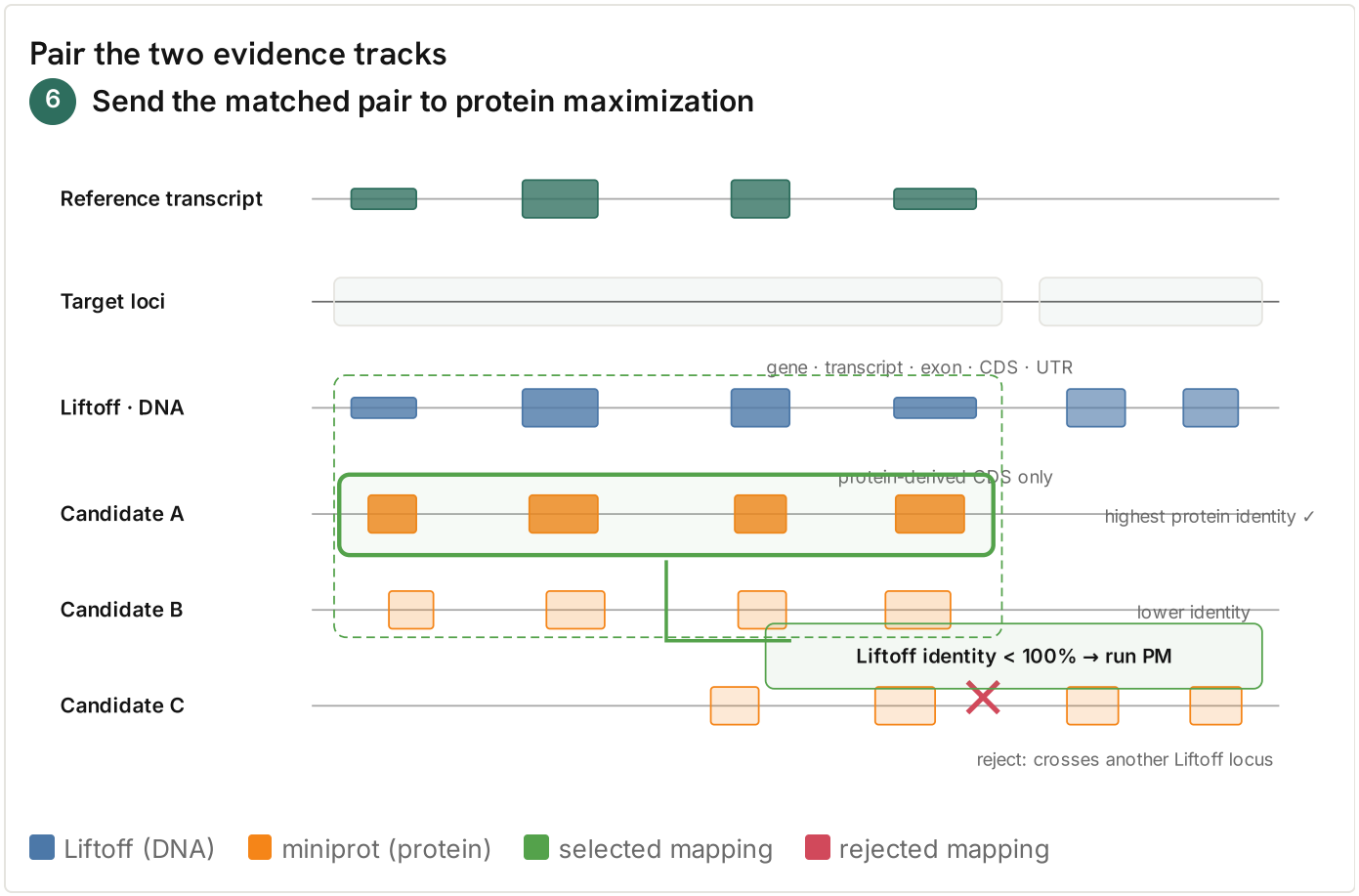
How LiftOn combines the evidence

LiftOn is a homology-based annotation lift-over algorithm, implemented as open-source software, that runs Liftoff and miniprot and reconciles their annotations around one explicit objective: preserve as much of the reference protein as possible without discarding the broader structure supplied by DNA alignment. The workflow is not a vote between two finished annotations. It pairs corresponding models, compares them in reference-protein coordinates, constructs a new CDS (coding sequence, the exon segments that translate into protein) chain, checks alternative open reading frames, and then controls where additional copies may be placed.

1. Map and pair the two annotations

Liftoff first maps the reference annotation hierarchy through DNA alignment, including genes, transcripts, exons, CDS features, and UTR context. miniprot independently maps each reference protein to the target genome and produces CDS-only transcript models. LiftOn pairs a miniprot transcript with a Liftoff transcript when their loci overlap and the protein mapping does not cross an unrelated Liftoff gene. Read-through mappings that span multiple loci are rejected; when several protein mappings remain for one transcript, LiftOn selects the one with the highest reference-protein identity.

This pairing step is also an important boundary on what protein maximization does. It operates on matched **protein-coding transcripts**. Noncoding genes and other features have no reference protein to optimize, so they remain on the DNA-based transfer path. If the Liftoff translation is already identical to the reference protein, the published algorithm has no coding discrepancy to repair.

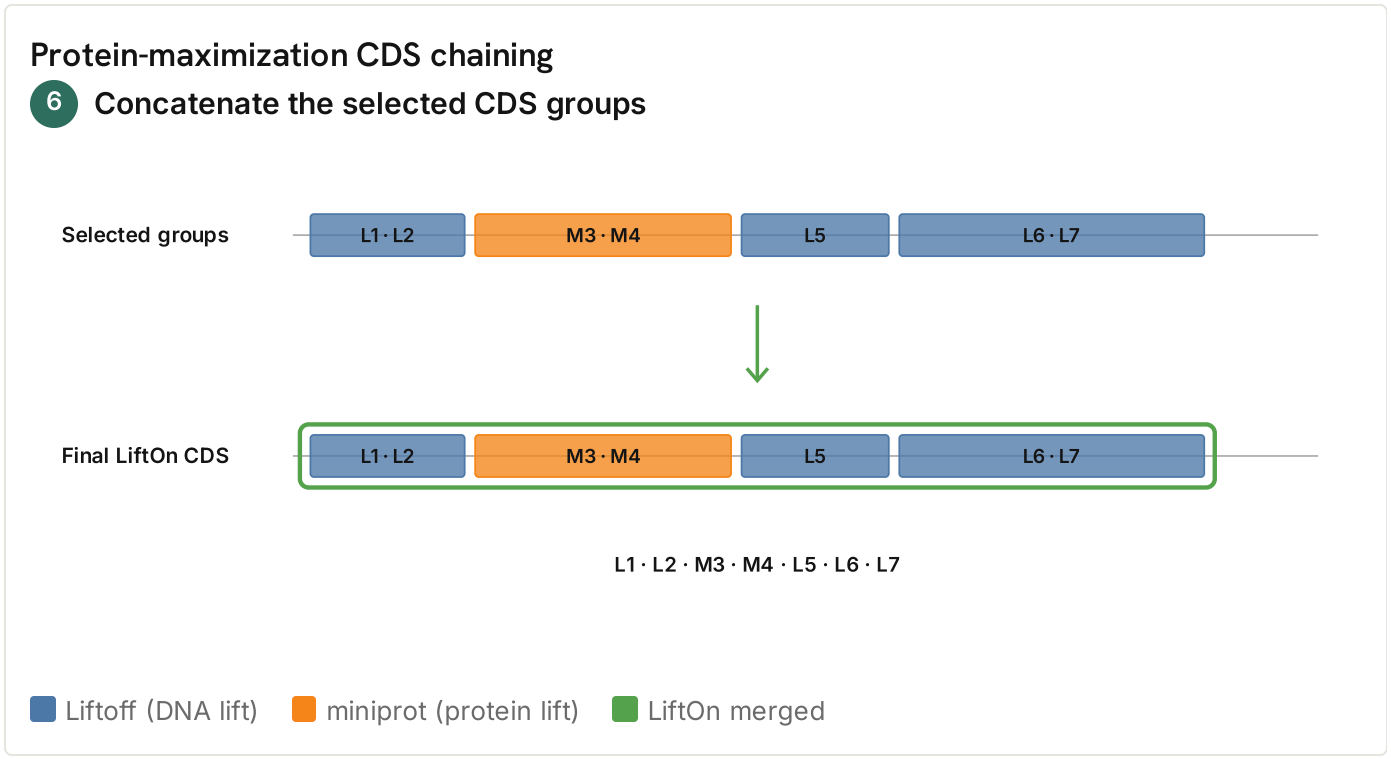


2. Synchronize and chain the CDS groups

For a matched pair, LiftOn translates both candidate CDS models and globally aligns each protein to the full-length reference protein. It then projects every CDS boundary onto those protein alignments. Starting at the 5′ end, the algorithm walks both CDS lists until Liftoff and miniprot have consumed the same cumulative number of reference amino acids. Those shared endpoints delimit directly comparable chunks, even when one source uses one CDS and the other uses several to cover the same reference interval.

Within each chunk, LiftOn calculates partial protein identity and keeps the higher-scoring source. An exact tie goes to Liftoff, preserving the DNA-derived structure when protein evidence offers no improvement. Concatenating the selected chunks creates a new CDS model assembled from the strongest local evidence rather than choosing either input wholesale.

The walkthrough follows the worked example in Figure 1 of the paper: the two protein alignments are annotated with their CDS boundaries, divided into five comparable groups, and resolved into the final L1–L2–M3–M4–L5–L6–L7 chain.



3. Search for a better open reading frame

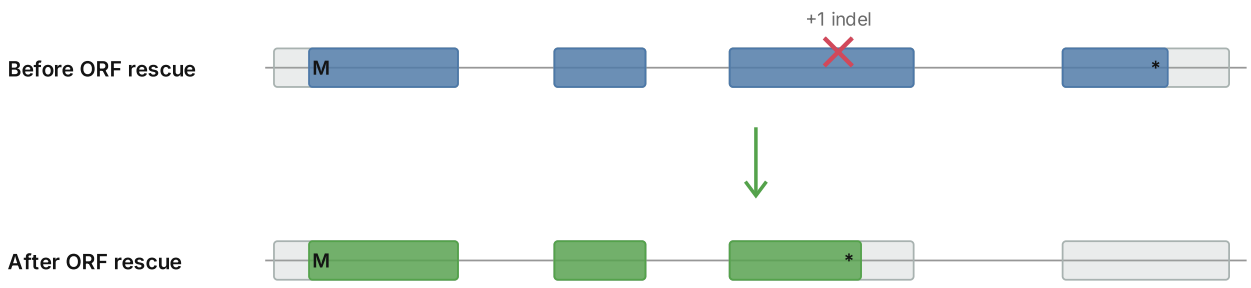
Chaining can still leave a frameshift (an insertion or deletion that shifts the reading frame out of register), a premature stop (a stop codon that appears earlier than the true one, truncating the protein), a lost stop codon, or a lost start codon. For transcripts with these damaging patterns, LiftOn scans the spliced transcript in all three reading frames. It retains the longest valid ORF from each frame, translates those candidates, aligns each one to the reference protein, and selects the highest-identity candidate. The CDS boundaries are updated only when that ORF improves on the current annotation.

This search changes the annotation, not the underlying genome. Figure 1 shows six possible outcomes: repairing a frameshift, stopping at an earlier valid stop, switching to a downstream start after stop gain, extending into the 3′ UTR after stop loss, and moving a lost start either downstream or upstream into the 5′ UTR. The selector below applies the same search rule to each case.

Repair the open reading frame

Frameshift

5 Adopt the candidate only when identity improves

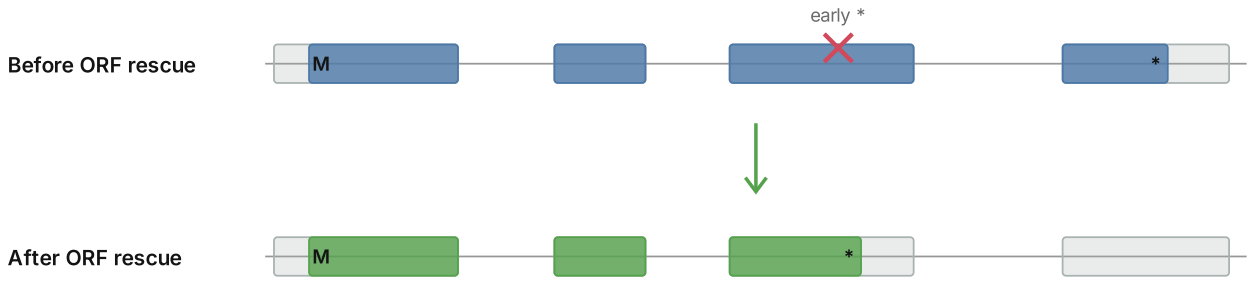


move the CDS stop before the disrupted translation

adopt only if id(o*) > id(current); then map boundaries back through exon coordinates

Stop gain · earlier stop

5 Adopt the candidate only when identity improves

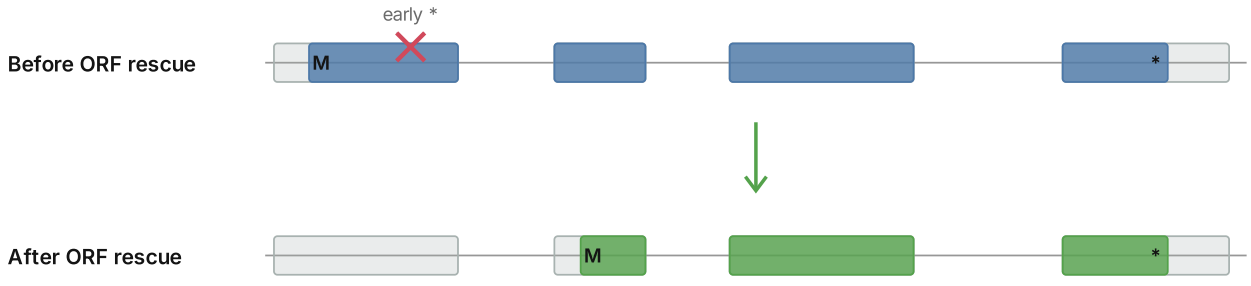


use the first valid stop codon

adopt only if id(o*) > id(current); then map boundaries back through exon coordinates

Stop gain · downstream start

5 Adopt the candidate only when identity improves

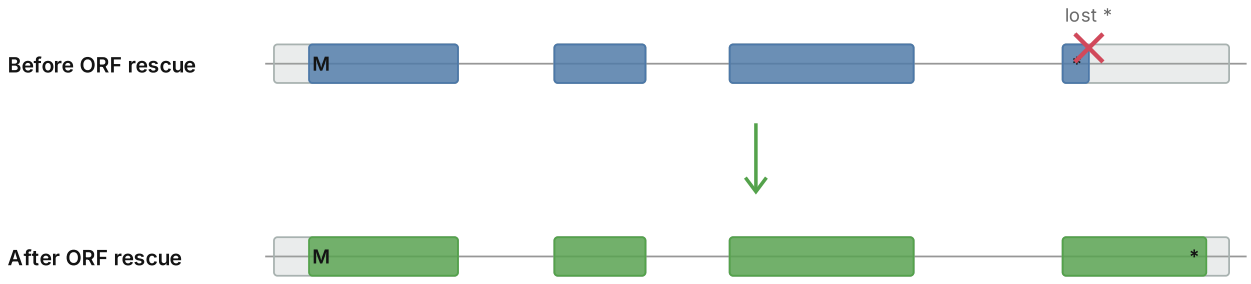


switch to a downstream start codon

adopt only if id(o*) > id(current); then map boundaries back through exon coordinates

Stop loss · 3′ extension

5 Adopt the candidate only when identity improves

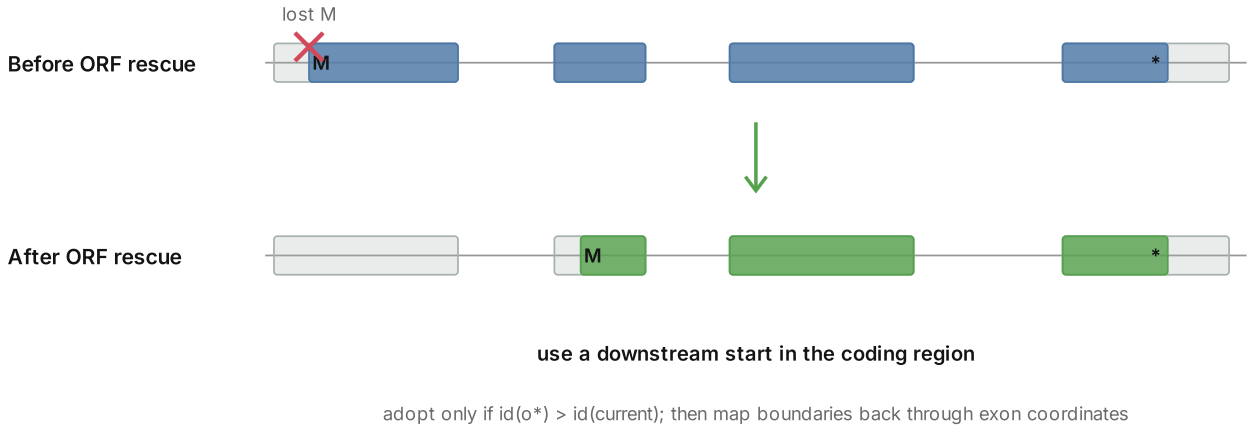


extend into the 3′ UTR to a new stop

adopt only if id(o*) > id(current); then map boundaries back through exon coordinates

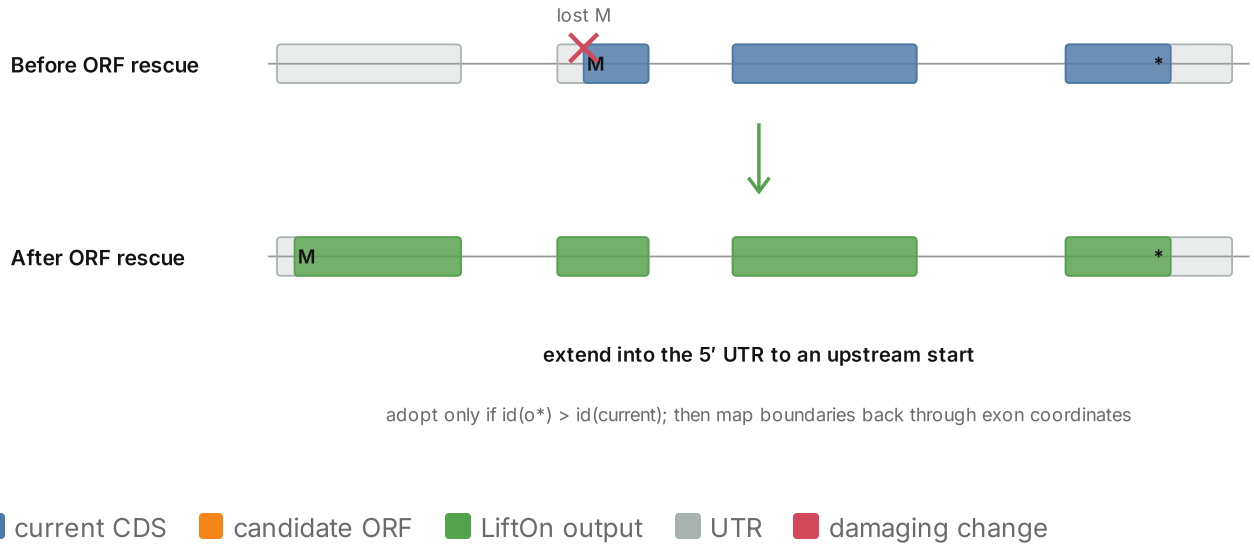
Start loss · downstream start

5 Adopt the candidate only when identity improves



Start loss · upstream start

5 Adopt the candidate only when identity improves



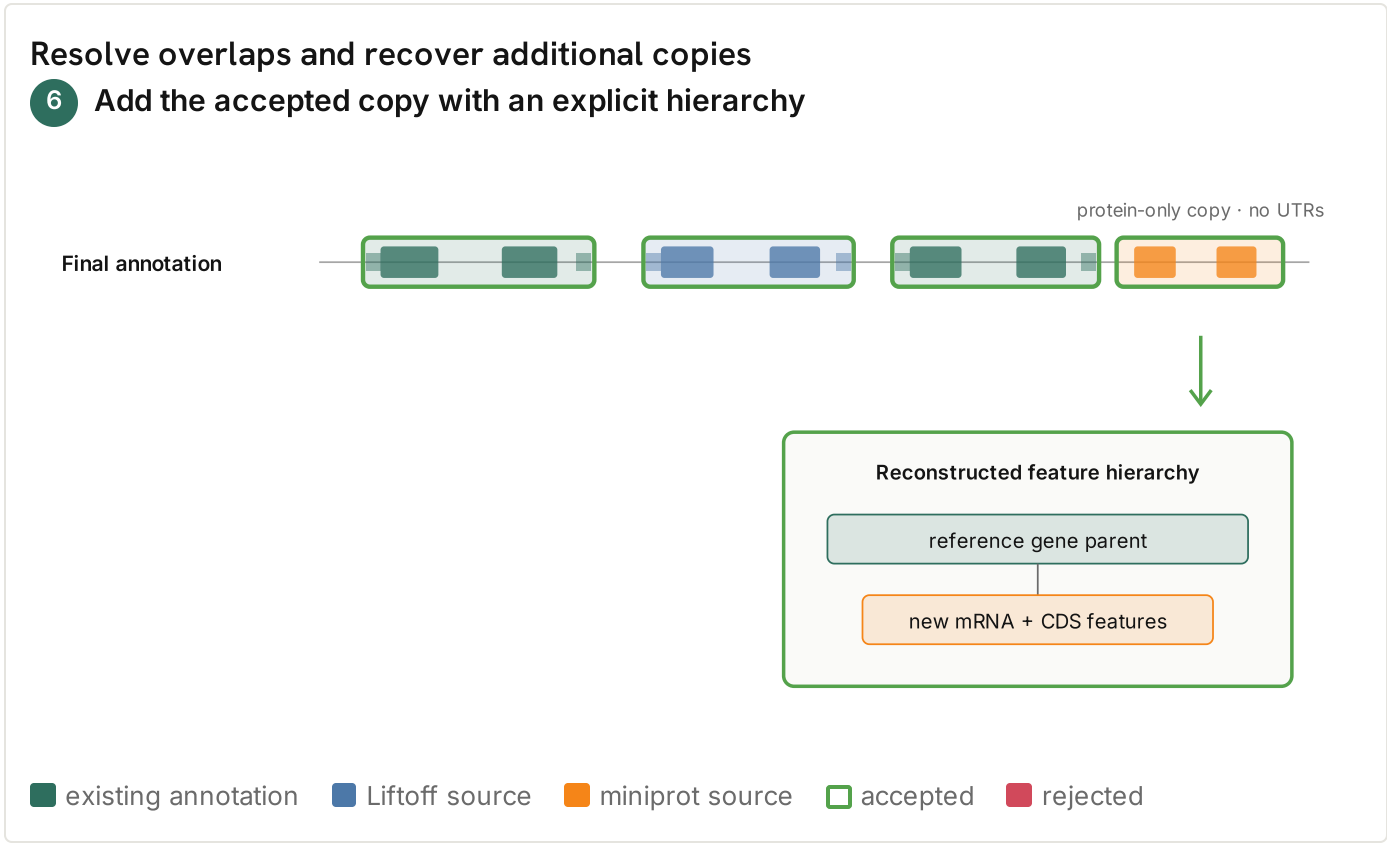
current CDS · candidate ORF · LiftOn output · UTR · damaging change

4. Resolve overlaps and recover additional copies

A one-to-one transfer is not enough when the target contains duplicated genes. LiftOn lets Liftoff search for DNA-supported copies first because DNA mapping can preserve UTRs and the full feature hierarchy. New loci must satisfy the requested sequence-identity threshold and may overlap existing genes only where that overlap exists in the reference or remains within the allowed 10%.

LiftOn then considers protein mappings that Liftoff missed. These miniprot-only candidates face stricter safeguards against potential processed or partial mappings: at most 10% overlap with accepted loci; a

single CDS only when the reference transcript is also single-CDS; and a coding-length ratio between 0.9 and 1.5 relative to the reference gene's longest isoform. An accepted protein-only copy receives a reconstructed gene–mRNA–CDS hierarchy, but no UTRs, because protein alignment does not define them. Noncoding copies therefore depend on Liftoff.



Together, these stages explain what "combining DNA and protein evidence" means operationally. DNA supplies location and transcript structure; protein supplies a conserved scoring coordinate; LiftOn makes each replacement or addition under an explicit rule rather than blending the signals into an opaque consensus.

What it showed

It improves where the two inputs disagree. The first test was the human genome itself: lifting RefSeq release 220 (O'Leary et al., 2016) from GRCh38 onto the complete telomere-to-telomere assembly, T2T-CHM13 (Nurk et al., 2022). LiftOn mapped 37,828 of 37,986 genes, a 99.6% gene mapping rate, while the protein-coding transcript analysis showed the central reason for combining evidence: Liftoff and miniprot often made different errors, and LiftOn usually matched or improved on each input by reference-protein identity.

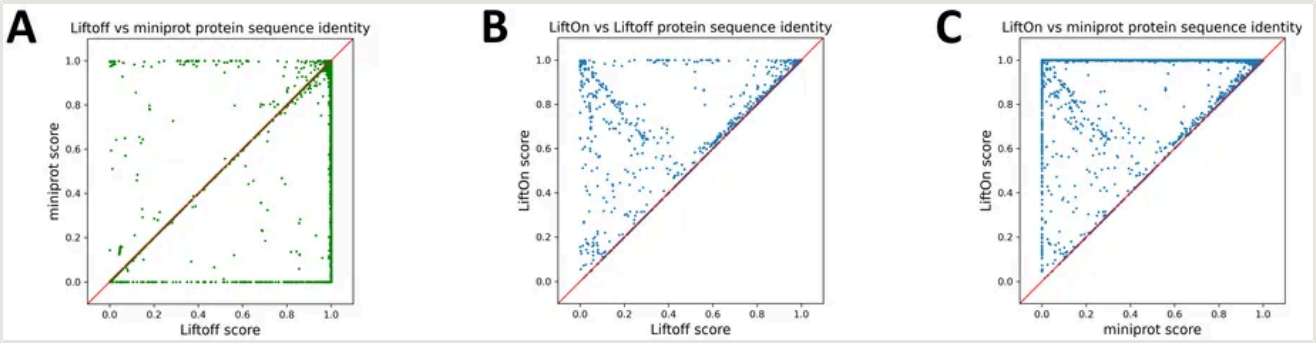


Figure 1. LiftOn compared with its two evidence sources. Each dot is a protein-coding transcript lifted from GRCh38 to T2T-CHM13, plotted by gap-compressed protein sequence identity. **(A)** miniprot vs Liftoff — neither method dominates the other. **(B, C)** LiftOn vs Liftoff, and LiftOn vs miniprot — most points sit on or above the diagonal, meaning the protein-maximization algorithm usually matches or improves on each input.

Panel A shows that Liftoff and miniprot disagree for many transcripts: one preserves the reference protein better where the other does not. Panels B and C quantify what LiftOn gains from that disagreement. It improved 866 protein-coding transcripts over Liftoff, with 113 reaching 100% protein identity, and improved 30,266 over miniprot, with 22,746 reaching 100% identity. LiftOn is not averaging the annotations; it selects better-supported CDS segments under a reference-protein objective.

It finds gene copies that a one-to-one map would miss. Because T2T-CHM13 is complete — it fills in the repetitive, duplicated regions that earlier assemblies left as gaps — it actually contains extra copies of some genes. A strict one-to-one lift-over can’t represent those; LiftOn looks for them explicitly, and found extra copies of 86 protein-coding genes, adding 320 new gene loci.

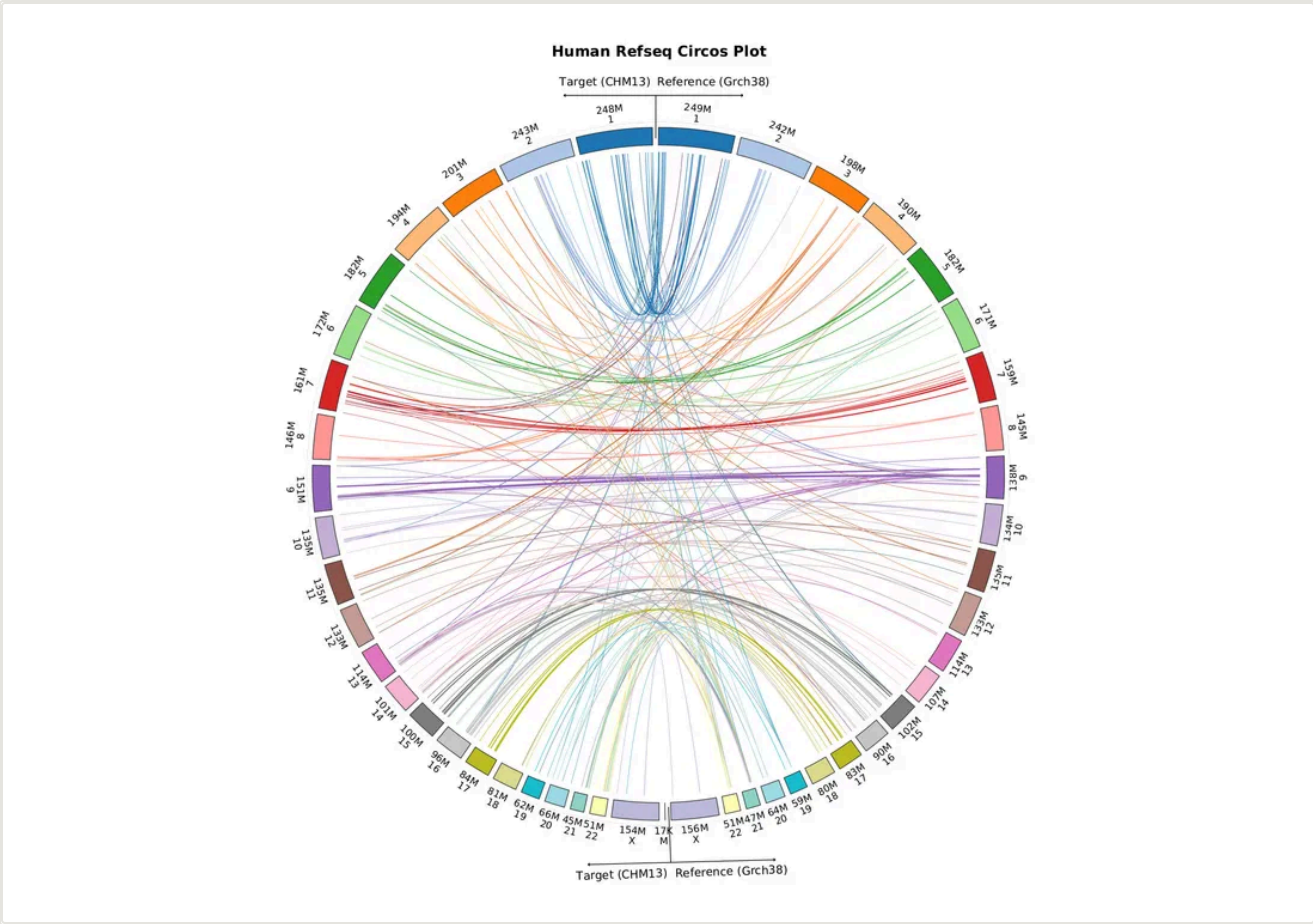


Figure 2. Extra gene copies LiftOn recovered in T2T-CHM13. Each ribbon connects an original gene copy to an additional copy LiftOn placed on the complete T2T-CHM13 assembly, colored by the chromosome of the original. In the paper, LiftOn identified 86 protein-coding genes with at least one extra copy, for 320 additional protein-coding gene loci.

It improves the evaluated T2T-CHM13 annotation in specific cases. The paper also compared LiftOn with the RefSeq-derived T2T-CHM13 annotation used in the study. On chromosomes 1-22 and X, LiftOn produced 665 protein-coding transcripts with higher identity to the corresponding GRCh38 reference proteins.

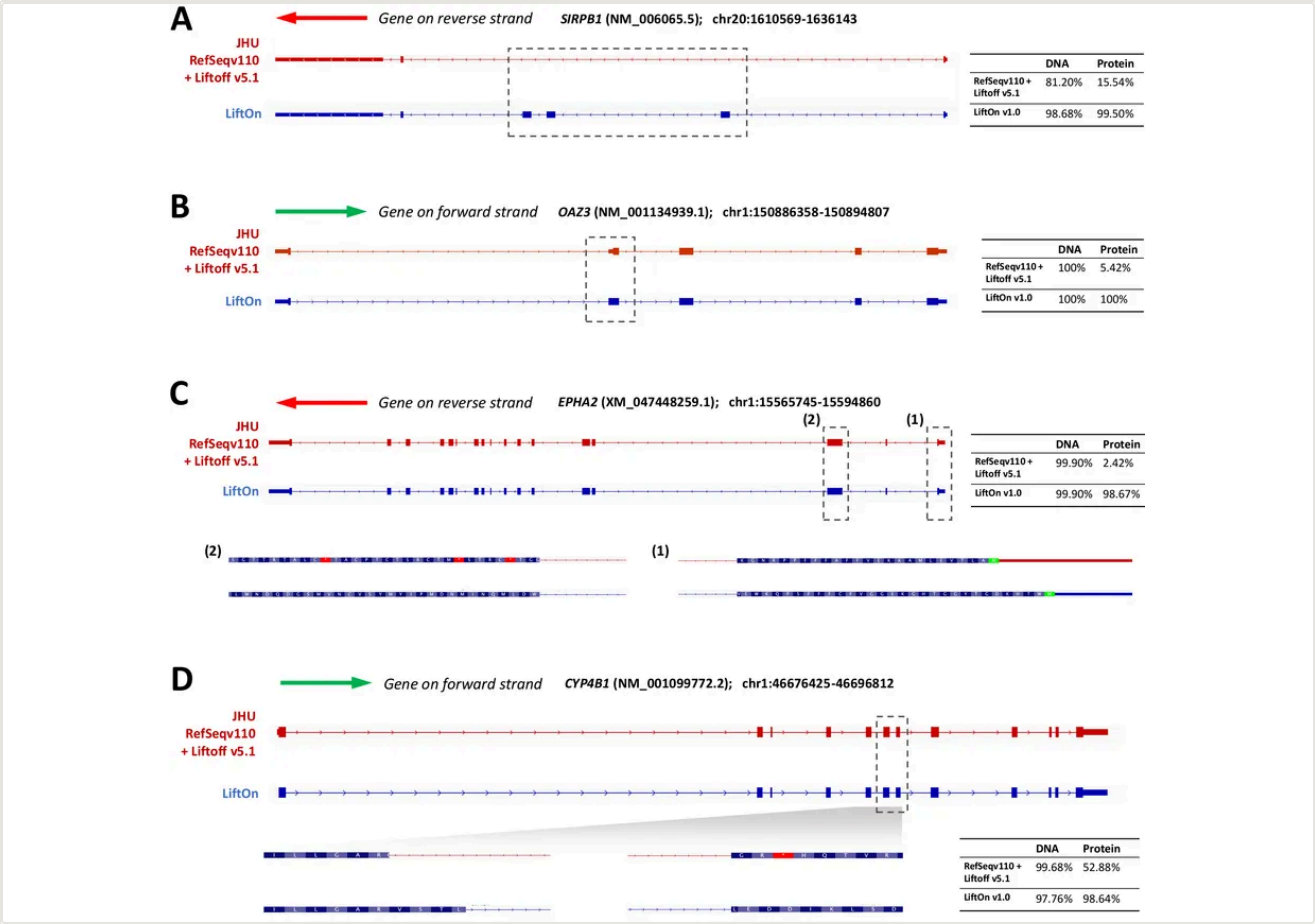


Figure 3. Four genes where LiftOn produces a model with higher reference-sequence identity than the T2T-CHM13 annotation evaluated in the paper (red = evaluated annotation, blue = LiftOn). **(A)** SIRPB1: LiftOn recovers three coding exons, raising DNA identity from 81% to 98%. **(B)** OAZ3: the protein model rises from 5.42% to 100% identity. **(C)** EPHA2: selecting a different start raises protein identity from 2.4% to 98.7%. **(D)** CYP4B1: an 11-nucleotide donor-site shift resolves the predicted frameshift and raises protein identity from 53% to 99%.

In each example, the existing annotation had a concrete protein-coding problem — missing exons, a truncated protein, the wrong start codon, or a frameshift — and LiftOn's protein-maximizing search produced a model with much higher fidelity to the GRCh38 reference protein. The broader comparison matters here: these four examples illustrate a larger set of 665 CHM13 protein-coding transcripts where LiftOn had higher protein identity than the published annotation under the paper's evaluation.

And it holds up across species. The real motivation, though, was distance — annotating genomes that aren't just another assembly of the same species. So we pushed LiftOn outward: human to chimpanzee, fruit fly to a different fruit fly (Drosophila melanogaster to D. erecta), and mouse to rat. The mouse-to-rat jump is the hardest of these, with tens of millions of years between the two.

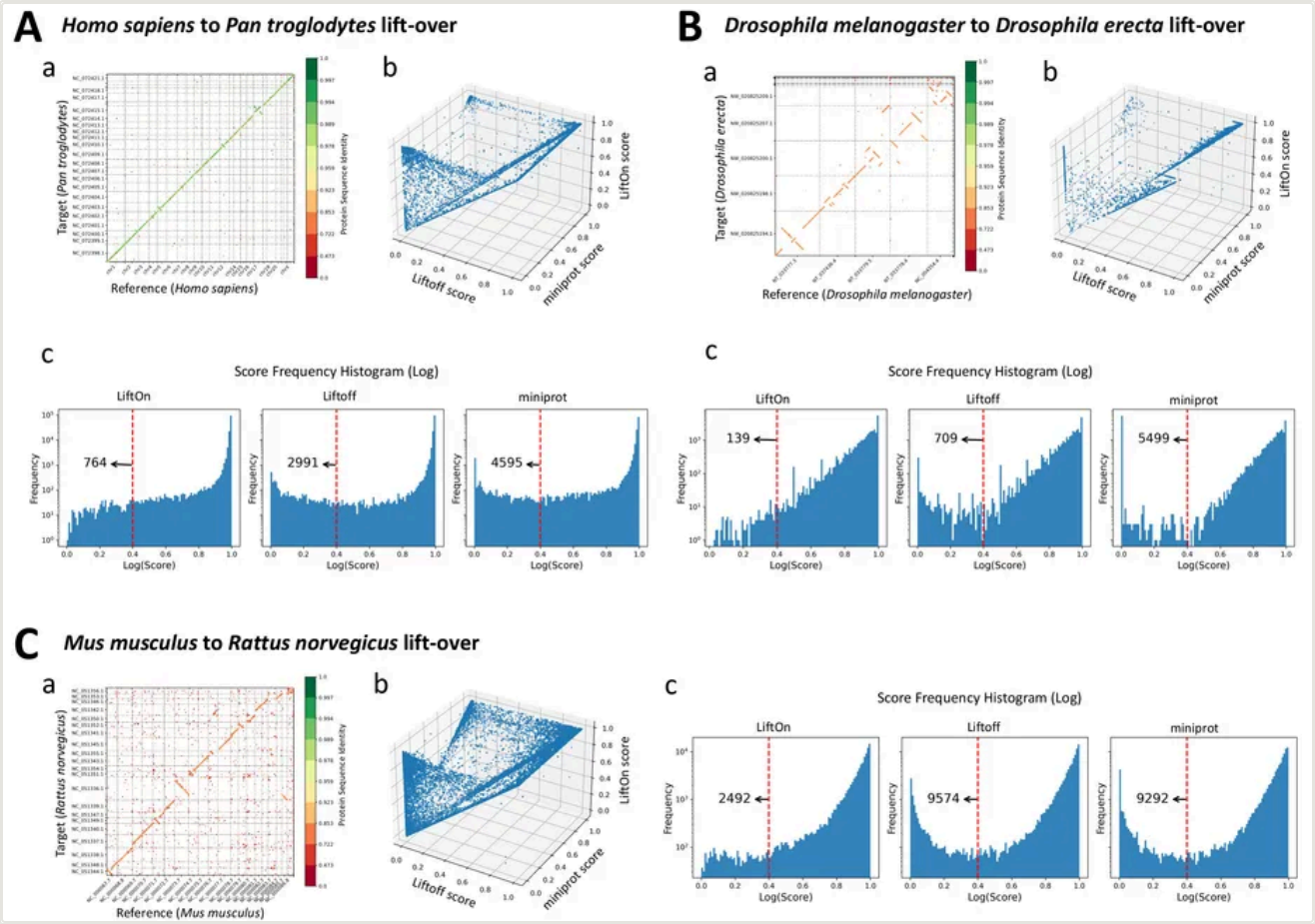


Figure 4. LiftOn across widening evolutionary distance: human→chimpanzee **(A)**, fly→fly **(B)**, and mouse→rat **(C)**. For each pair, the dot plot **(a)** shows the lifted genes preserve gene order and identity; the 3-D plot **(b)** compares per-transcript protein identity for Liftoff, miniprot, and LiftOn (points above the plane = LiftOn wins); and the histograms **(c)** show LiftOn's identity distribution shifted toward "identical" relative to either tool alone.

The pattern held, with the largest gains appearing in the more divergent examples tested in the paper. For mouse to rat, LiftOn improved 15,420 protein-coding transcripts over Liftoff and 30,574 over miniprot, mapped 94.3% of genes, and left far fewer protein-coding transcripts below the 40% identity threshold than either tool alone. In the closer human-to-chimpanzee experiment, LiftOn mapped 98.7% of genes; in the same-species lift-overs, honeybee reached 99.6%, rice reached 99.9%, and Arabidopsis reached 99.2% at the gene level.

The bigger picture

The bigger idea behind LiftOn is algorithmic: do not force one alignment signal to solve a problem it is not built to solve. DNA alignment is strongest for locus structure and UTRs; protein alignment is strongest for preserving the coding sequence across evolutionary distance. LiftOn turns their disagreement into a constrained optimization problem over coding segments, then checks the result with an ORF search and overlap logic. Annotation is the layer everything else stands on. If the gene models are wrong, every

downstream analysis inherits the error, quietly. As complete genomes and pangenomes multiply, the ability to carry accurate annotations from the genomes we understand onto the ones we have just sequenced becomes as important as the assembly itself.

The lesson I took from building it is about combining evidence without blurring the evidence together. DNA tells you a lot about where a gene is and how its transcript is structured; protein tells you what the coding sequence should preserve. LiftOn keeps those roles separate and then reconciles them with an explicit objective. The same instinct runs through my other work, like OpenSpliceAI, where the goal was a splice-site model you could retrain across the tree of life instead of one locked to a single species. Different problem, same conviction: build methods that travel.

LiftOn is free and open source, and it builds directly on Liftoff — one lab tool standing on the shoulders of another. Point it at a new genome and an annotation you trust, and see how much of the map carries over.

Read the paper in Genome Research, browse the code, or work through the documentation. LiftOn was built with Jakob M. Heinz, Celine Hoh, Alan Mao, Alaina Shumate, Mihaela Pertea, and Steven Salzberg at Johns Hopkins.

References

1. Chao, K.-H. et al. Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Research (2025). doi:10.1101/gr.279620.124
2. Shumate, A. and Salzberg, S. L. Liftoff: accurate mapping of gene annotations. Bioinformatics (2021). doi:10.1093/bioinformatics/btaa1016
3. Li, H. Protein-to-genome alignment with miniprot. Bioinformatics (2023). doi:10.1093/bioinformatics/btad014
4. 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
5. Nurk, S. et al. The complete sequence of a human genome. Science (2022). doi:10.1126/science.abj6987
