Combining DNA and protein alignments to improve genome annotation with LiftOn

2024.04.27

Kuan-Hao Chao

khchao.com

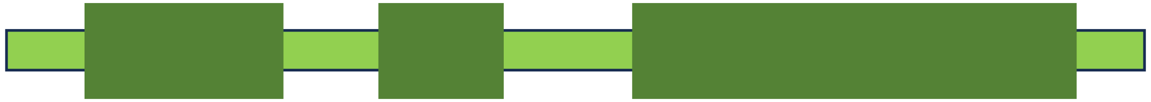
@KuanHaoChao

Kuanhao-Chao

Johns Hopkins University
Center for Computational Biology
CCB

Johns Hopkins
Whiting School of Engineering
Department of Computer Science

Genome annotation



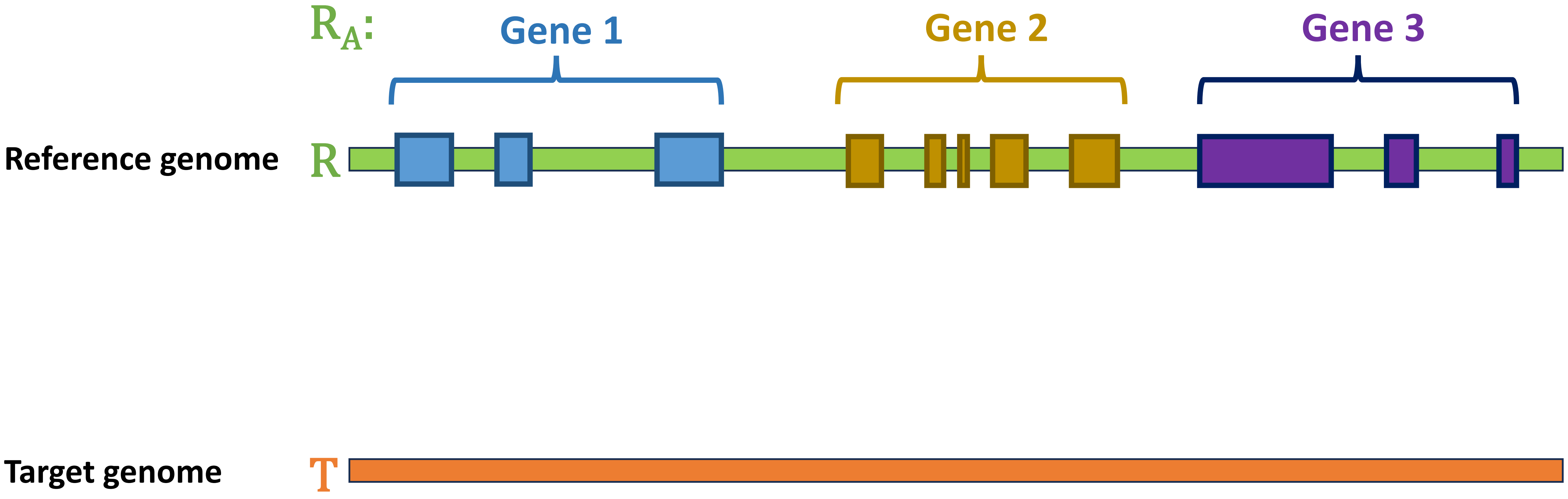
Genome
(FASTA)

```
CAGCCCCCGGAGACTtaaatacaggaagaaaaaggCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCT
GCCTCCTACCCTTGCGCCTCATGACCAGCTTGTTGAAGAGATCCGACATCAAGTGCCCACCTTGGCTCGTGGCTCTCACT
GCAACGGGAAAGCCACAGACTGGGGTGAAGAGTTCAGTCACATGCGACCGGTgactccctgtccccaccccatgACACT
CCCCAGCCCTCCAAGGCCACTGTGTTTCCCAGTTAGCTCAGAGCCTCAGTCGATCCCTGACCCAGCACCGGGCACTGATG
AGACAGCGGCTGTTTGAGGagccacctcccagccacctcggggccagggccagggtgtGCAGCACCACTGTACAATGGGG
AAACTGGCCCAGAGAGGTGAGGCAGCTTGCCTGGGGTCACAGAGCAAGGCAAAAGCAGCGCTGGGTACAAGCTCAAAACC
ATAGTGCCCAGGGCACTGCCGCTGCAGGCGCAGGCATCGCATCACACCAGTGTCTGCGTTCACAGCAGGCATCATCAGTA
```

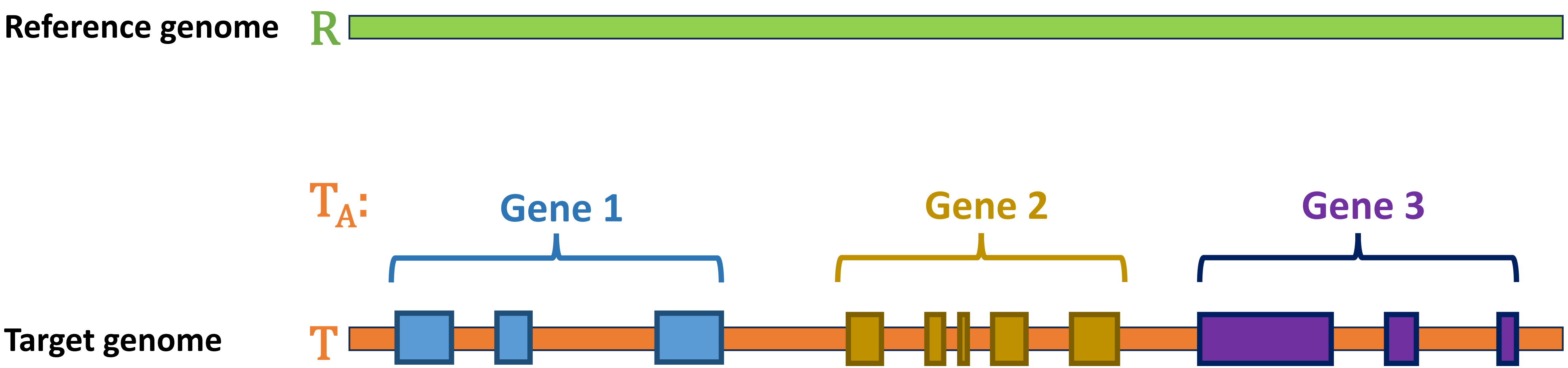
Annotation
(GFF / GTF)

```
chr1   BestRefSeq   gene   450740  451678  .   -   .   ID=gene-OR4F29;
chr1   BestRefSeq   mRNA   450740  451678  .   -   .   ID=rna-NM_001005221.2;Parent=gene-OR4F29;
chr1   BestRefSeq   exon   450740  451678  .   -   .   ID=exon-NM_001005221.2-1;Parent=rna-NM_001005221.2;
chr1   BestRefSeq   exon   452658  453675  .   -   .   ID=exon-NM_001005221.2-2;Parent=rna-NM_001005221.2;
chr1   BestRefSeq   exon   454672  459678  .   -   .   ID=exon-NM_001005221.2-3;Parent=rna-NM_001005221.2;
chr1   BestRefSeq   CDS    450740  451678  .   -   0   ID=cds-NP_001005221.2-1;Parent=rna-NM_001005221.2;
chr1   BestRefSeq   CDS    452658  453675  .   -   0   ID=cds-NP_001005221.2-2;Parent=rna-NM_001005221.2;
```

Lift-over Problem Definition:



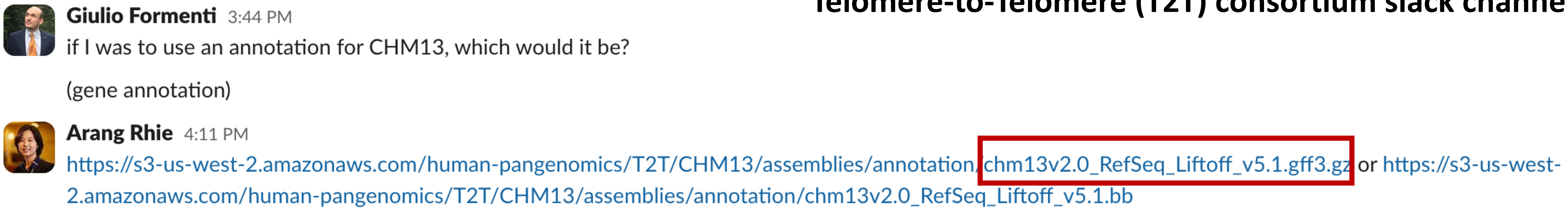
Lift-over Problem Definition:



If you were to use a CHM13 annotation …

Which tool to use?

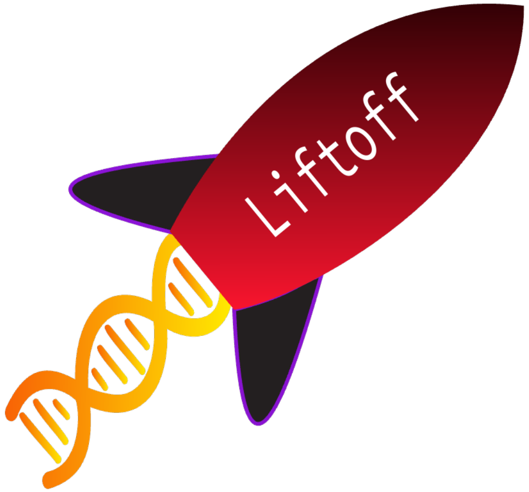
Telomere-to-Telomere (T2T) consortium slack channel

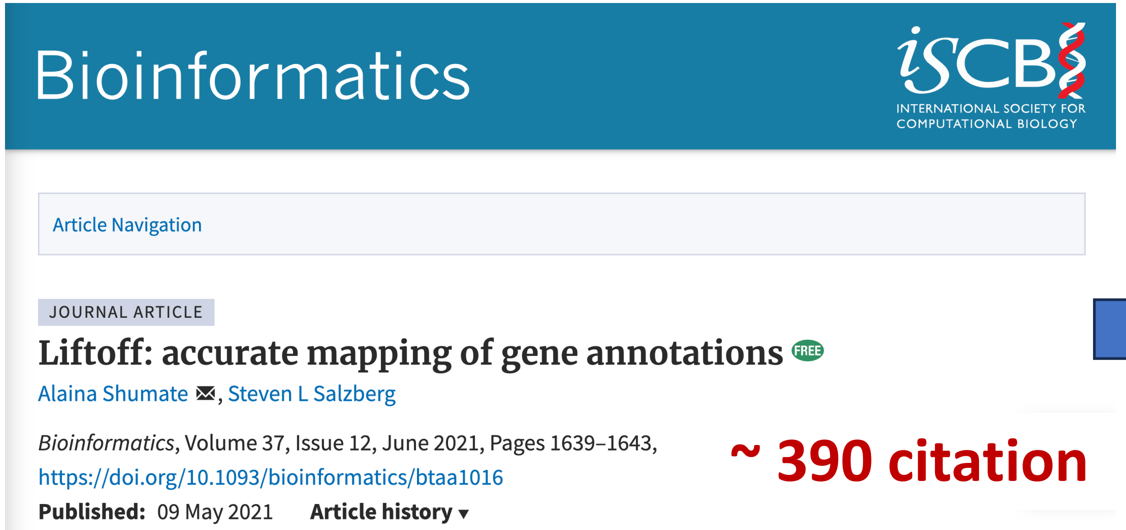


Giulio Formenti 3:44 PM
if I was to use an annotation for CHM13, which would it be?

(gene annotation)

Arang Rhie 4:11 PM
https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RefSeq_Liftoff_v5.1.gff3.gz or https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RefSeq_Liftoff_v5.1.bb





Bioinformatics

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Article Navigation

JOURNAL ARTICLE

Liftoff: accurate mapping of gene annotations FREE

Alaina Shumate ✉, Steven L Salzberg

Bioinformatics, Volume 37, Issue 12, June 2021, Pages 1639–1643, https://doi.org/10.1093/bioinformatics/btaa1016

Published: 09 May 2021 **Article history ▾**

~ 390 citation

LiftOn: Successor to Liftoff

Result 1
- outperforms state-of-the-art DNA- and protein-based liftover approaches

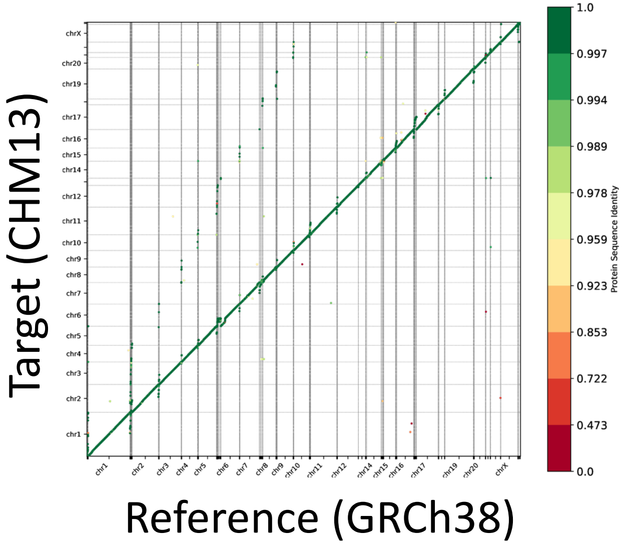
Result 2
- improves the annotation of protein-coding genes in T2T-CHM13 genome

Result 3
- Improves the annotation lift-over between relatively distant species, at least as divergent as mouse and rat.

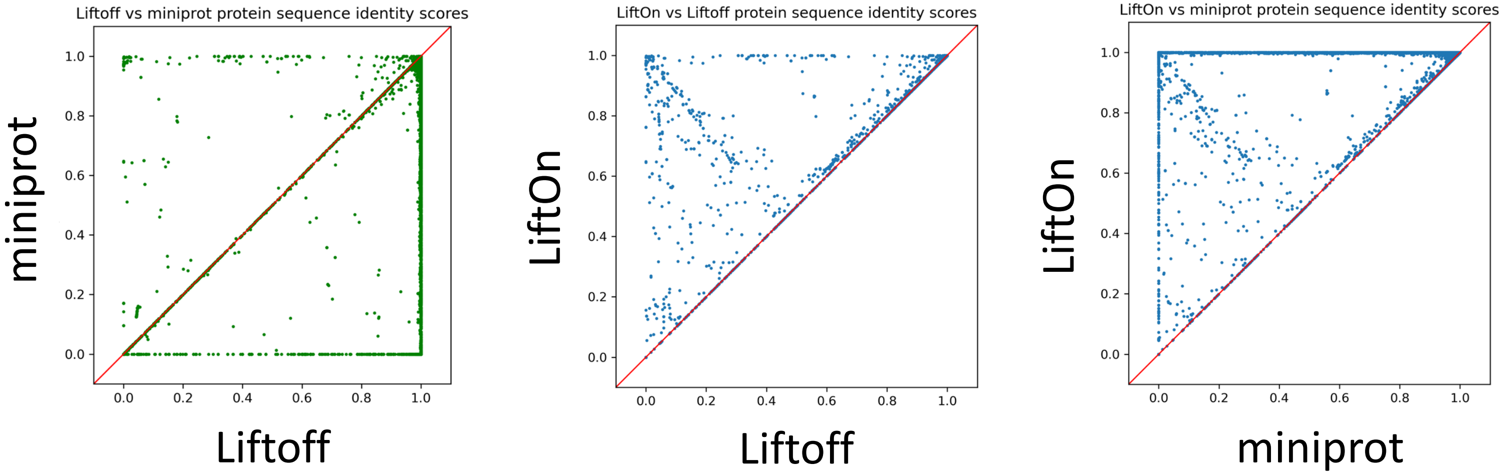
Methods
- Takes **DNA-**genome and **protein-**genome alignments and accurately maps annotations between genome assemblies of the same or different species.

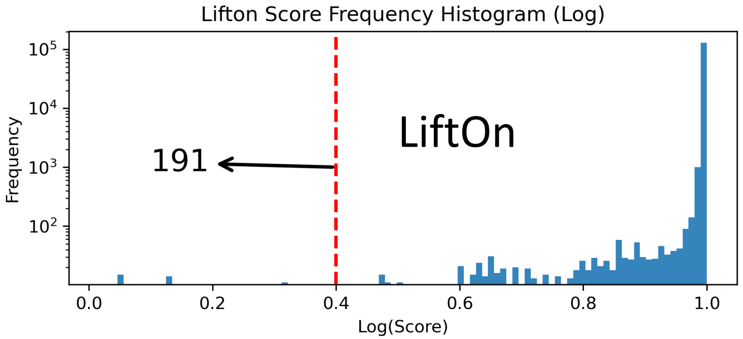
Result 1: improves DNA & protein-based lift-over

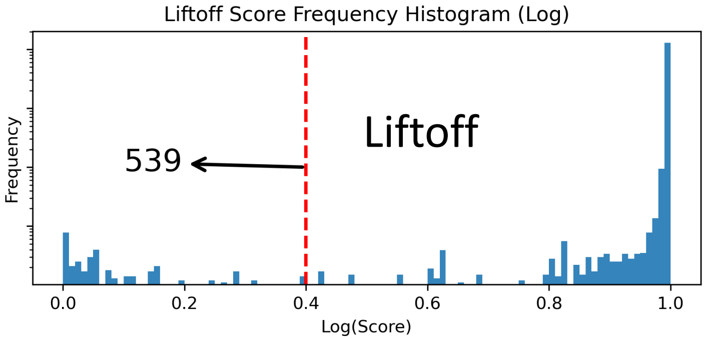
Map RefSeq v220 from GRCh38 -> CHM13V2.0

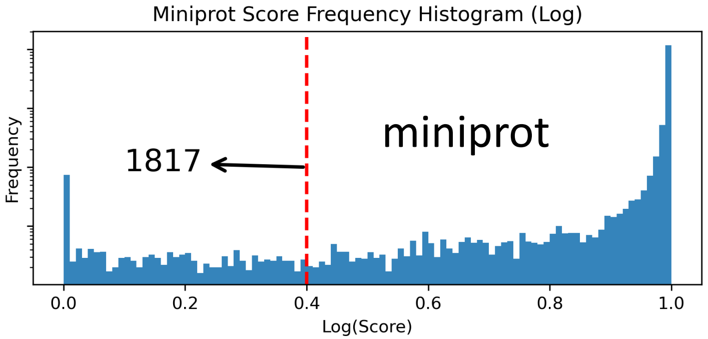


Compressed-gap protein sequence identity



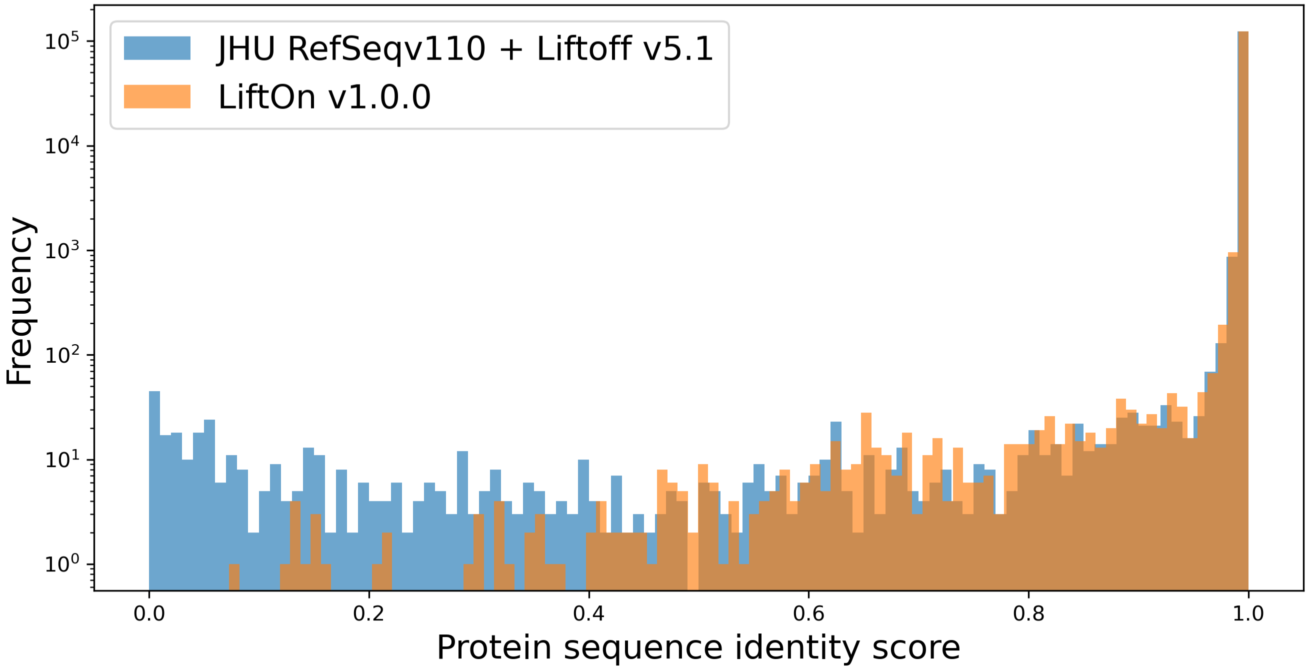


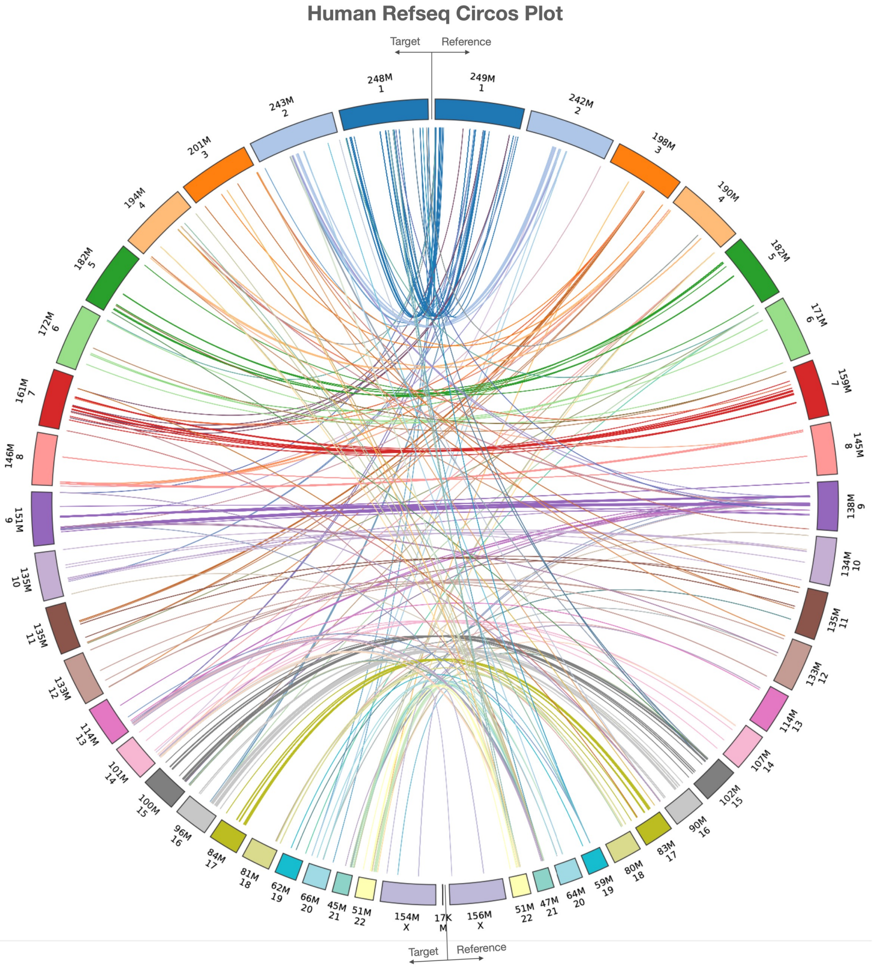




Result 2: improve CHM13 protein annotations

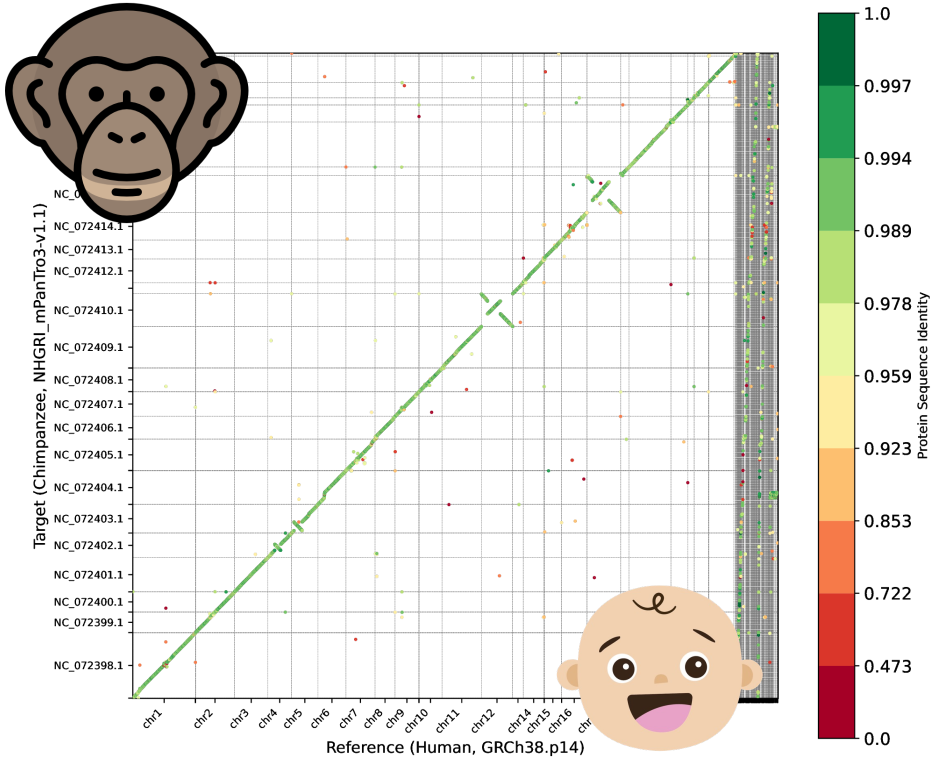
Protein sequence identity score frequency histogram





Result 3: improve distant species lift-over

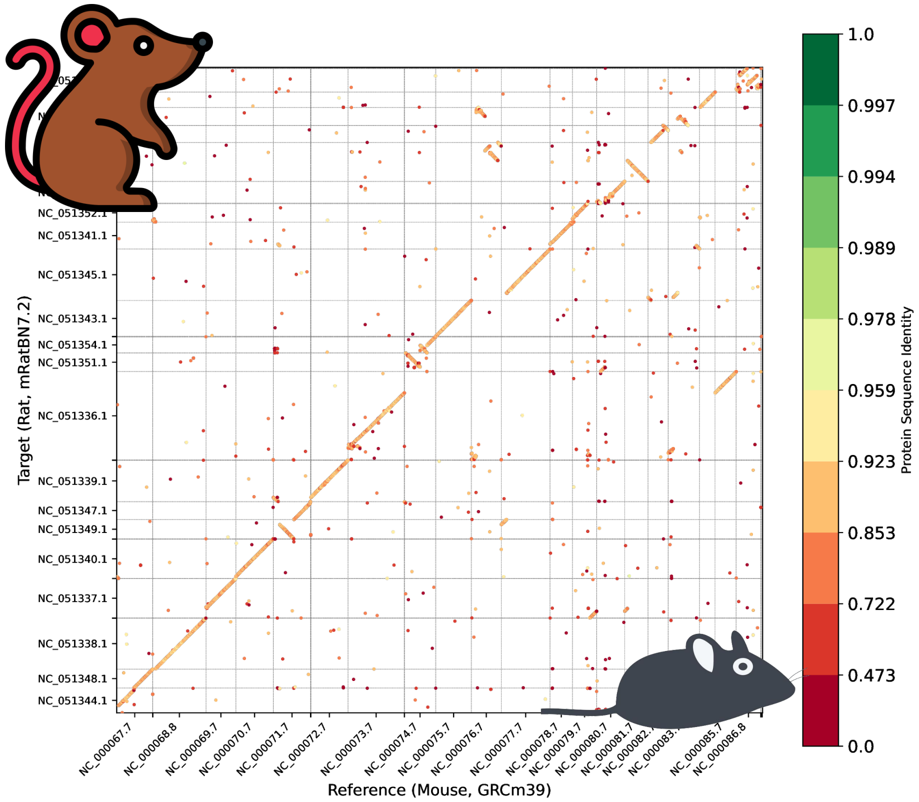
human to chimp



Mash : 0.013

Dashing2 : 0.47

mouse to rat

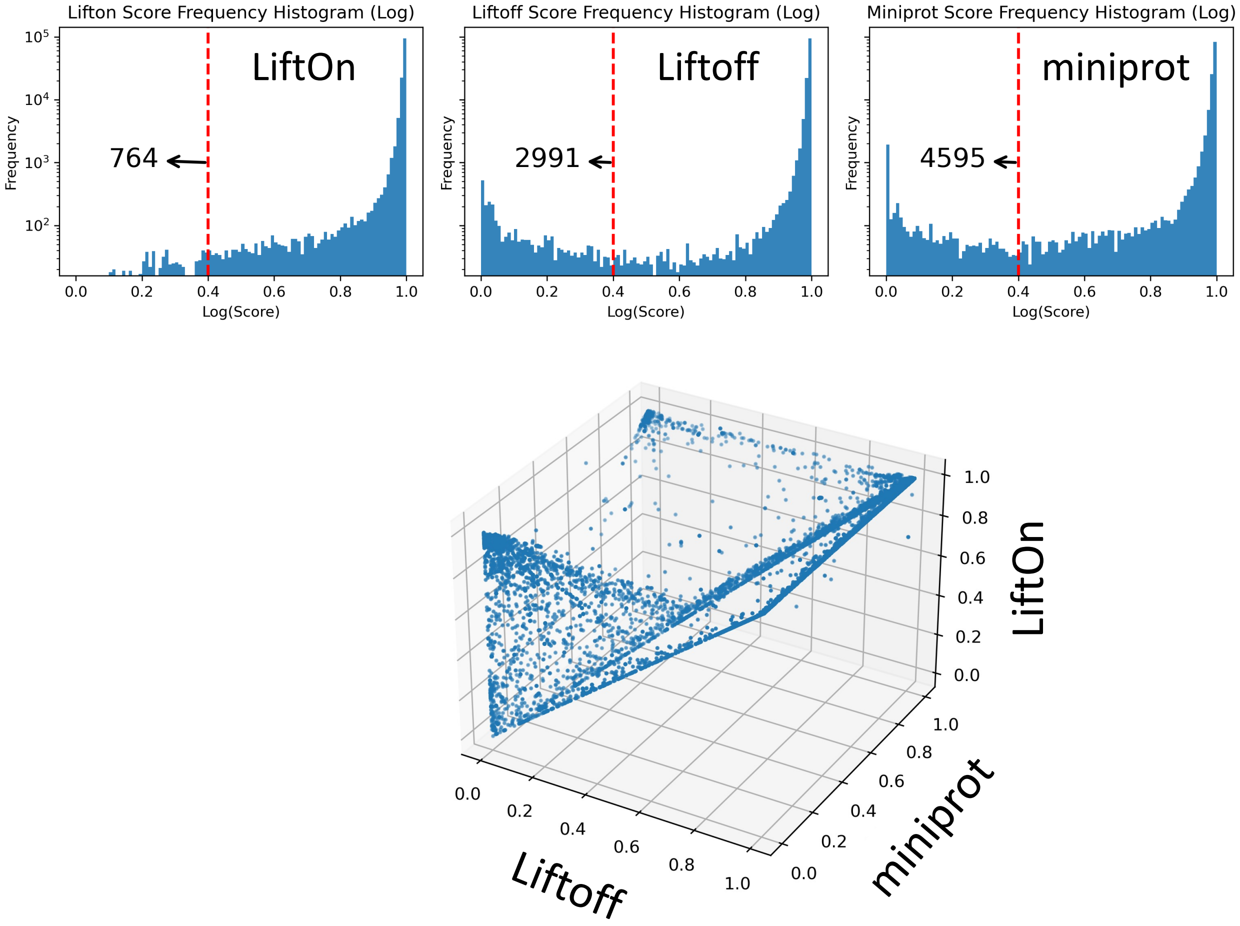


Mash : 0.120

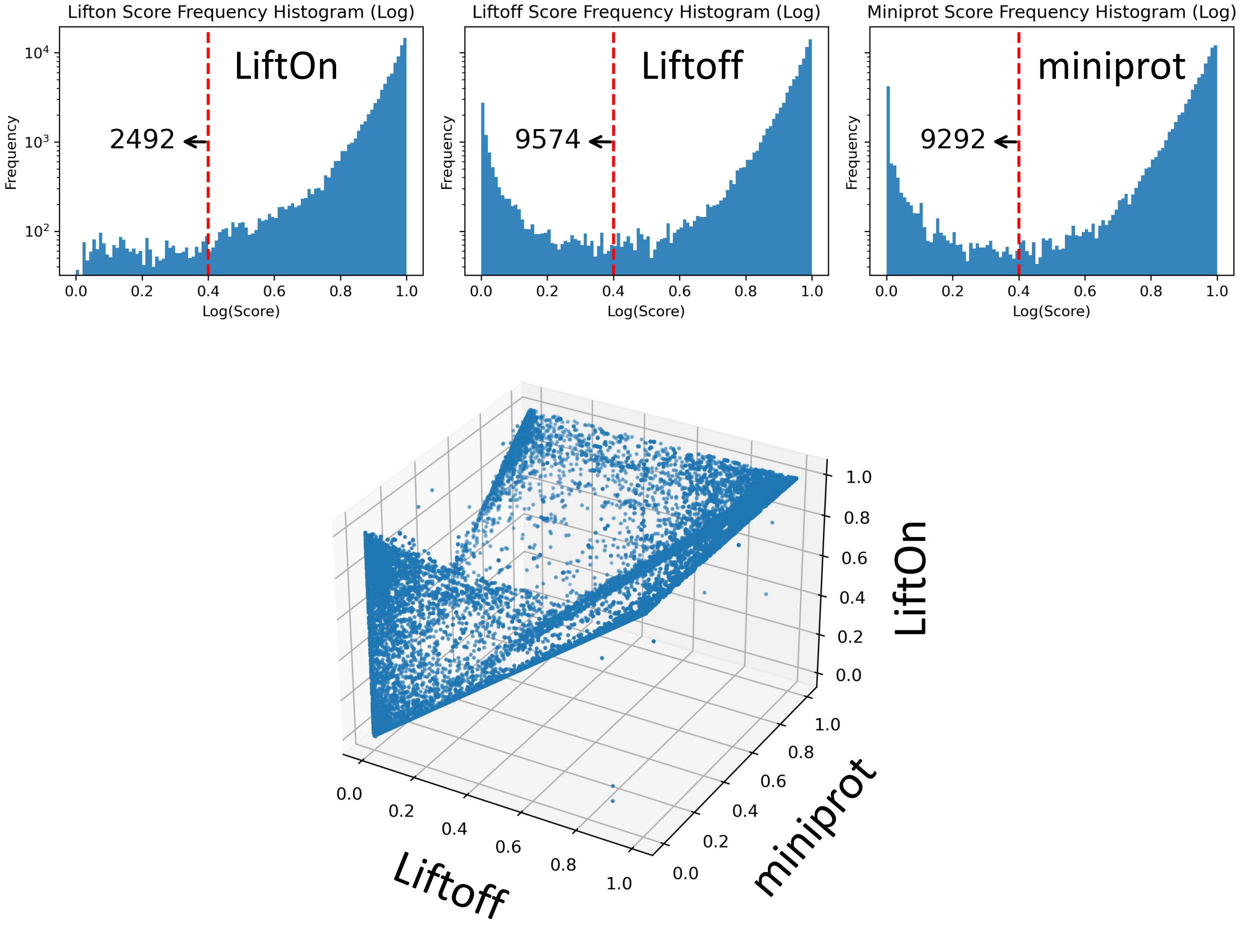
Dashing2 : 0.01

Result 3: improve distant species lift-over

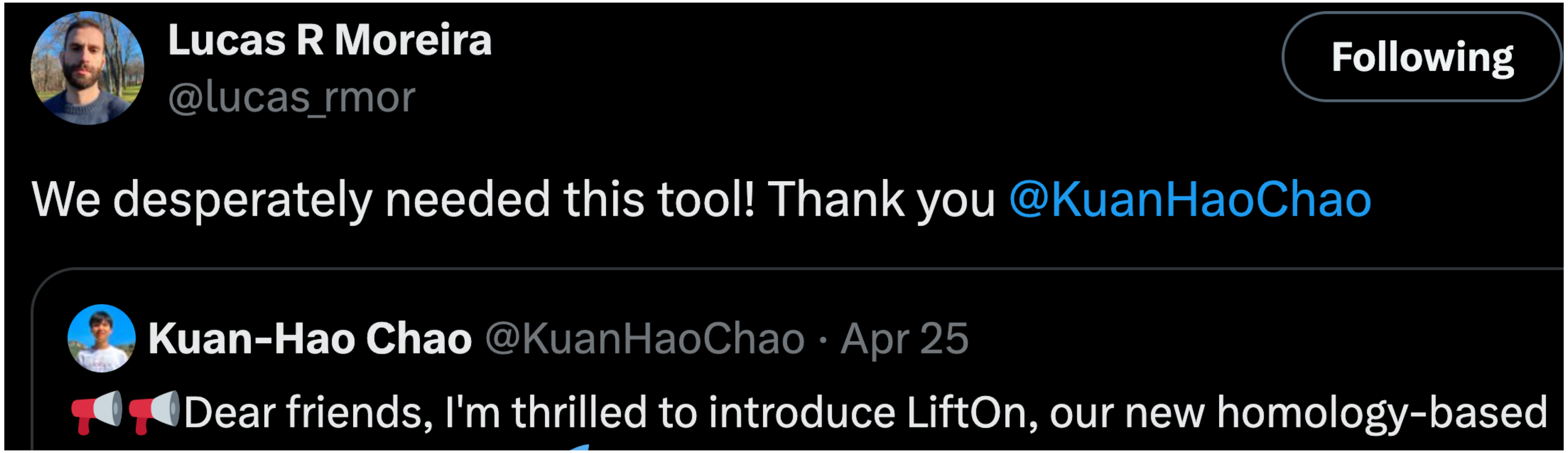
human to chimp



mouse to rat



Methods



Lucas R Moreira
@lucas_rmor

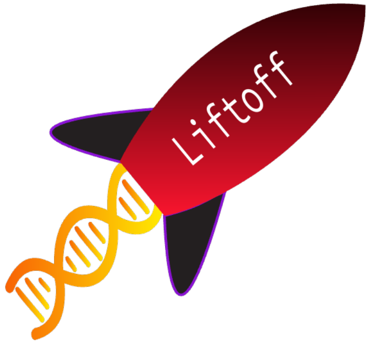
Following

We desperately needed this tool! Thank you @KuanHaoChao

Kuan-Hao Chao @KuanHaoChao · Apr 25

📢📢Dear friends, I'm thrilled to introduce LiftOn, our new homology-based

minimap2



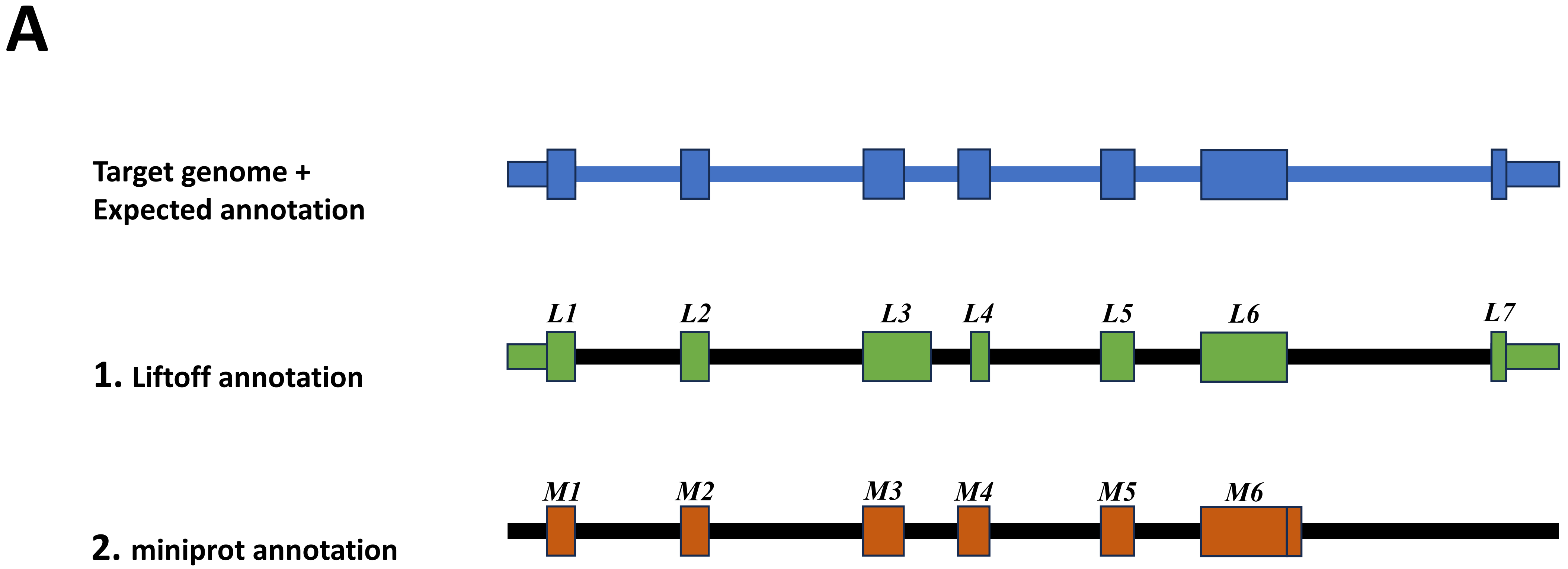
miniprot

github.com/lh3/minimap2

github.com/agshumate/Liftoff

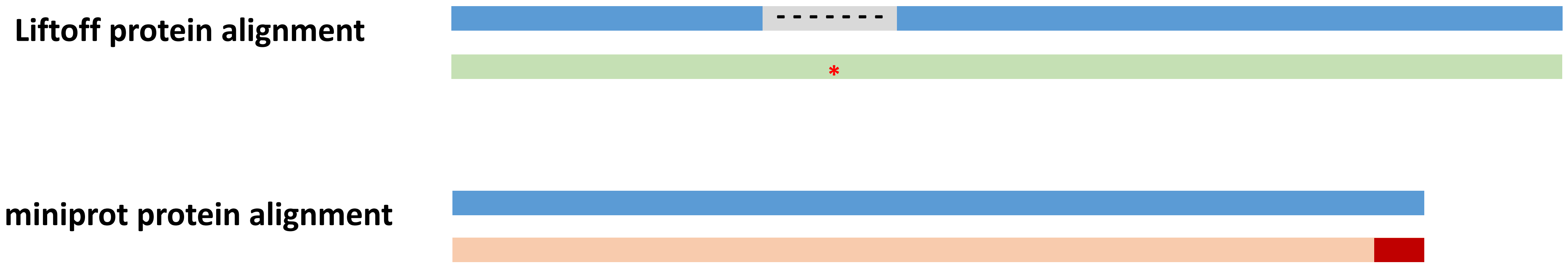
github.com/lh3/miniprot

LiftOn: Protein-maximization algorithm



LiftOn: Protein-maximization algorithm

B Step 1: Align Liftoff & miniprot proteins to reference protein



Liftoff protein alignment

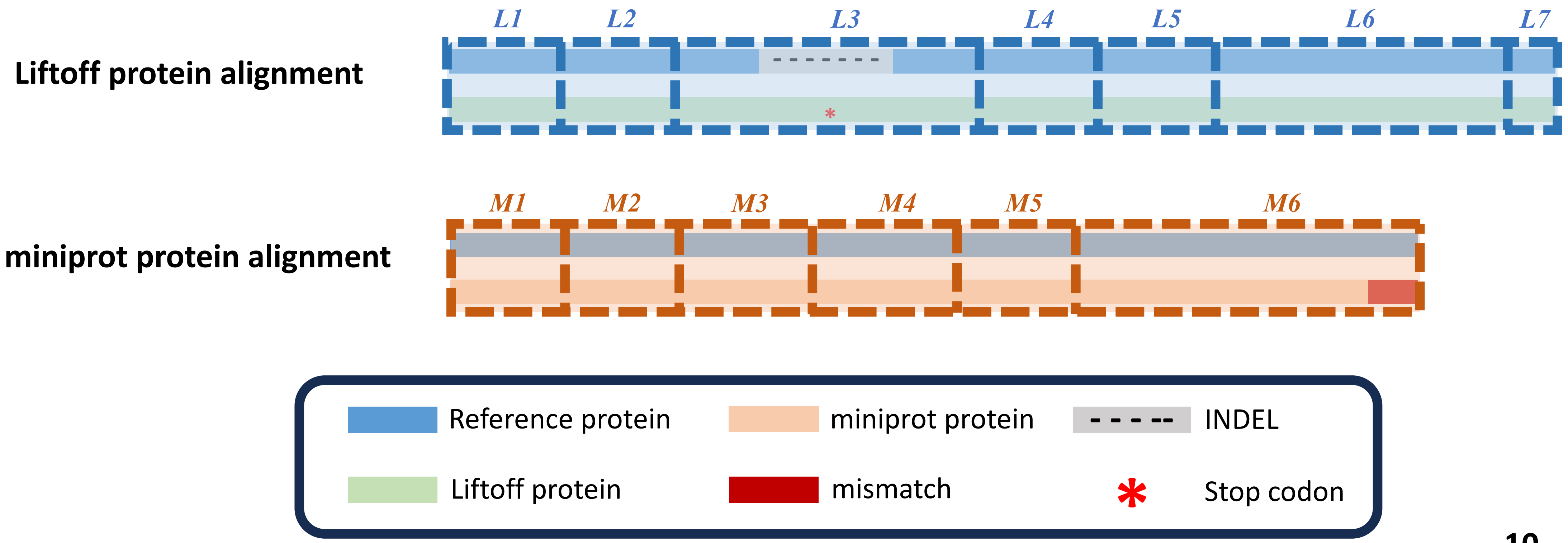
miniprot protein alignment

🟦 Reference protein 🟧 miniprot protein - - - -- INDEL

🟩 Liftoff protein 🟥 mismatch * Stop codon

LiftOn: Protein-maximization algorithm

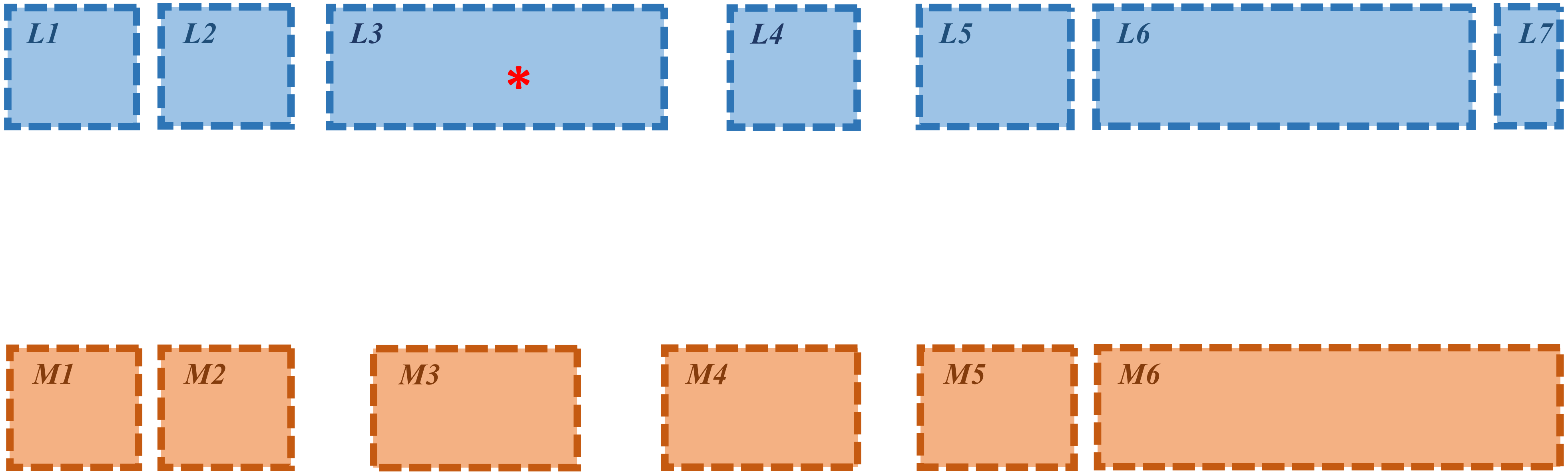
C Step 2: Mapped CDS boundaries onto Liftoff & miniprot protein alignments



LiftOn: Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”

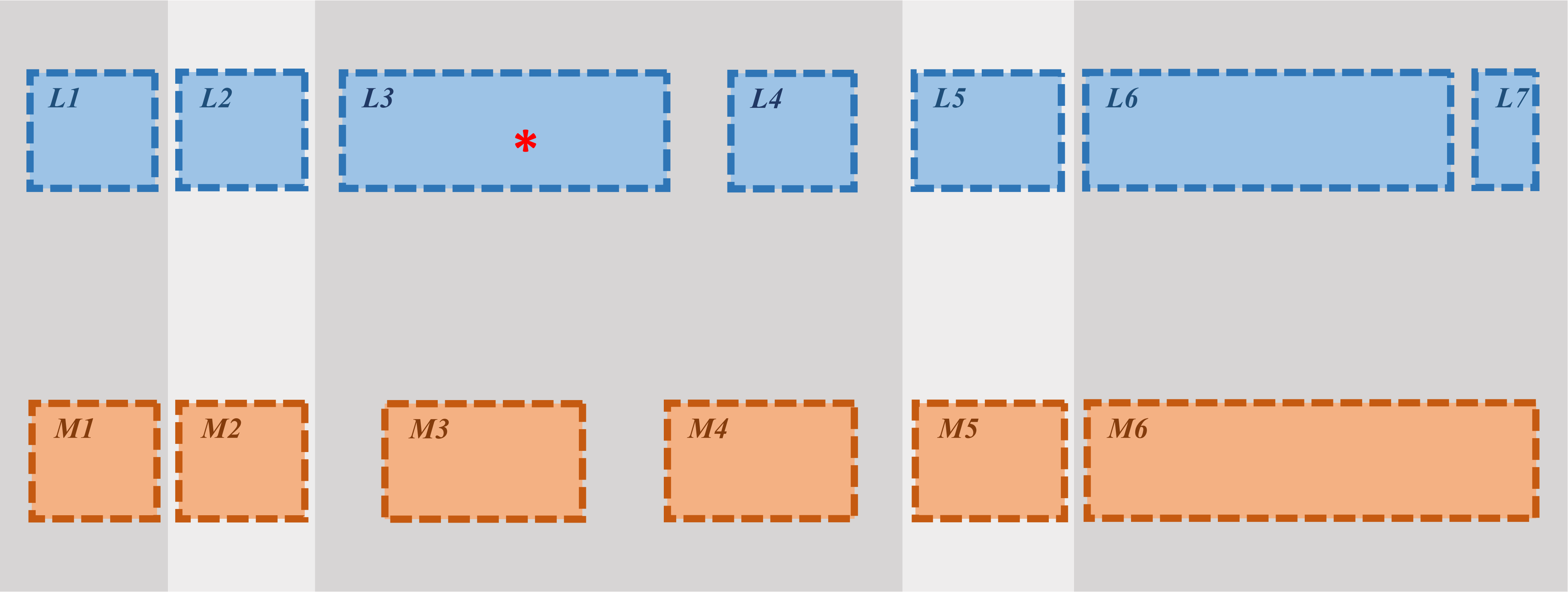
LiftOn CDS chaining



LiftOn: Protein-maximization algorithm

D **Step 3: group CDSs by "accumulated AA in the reference protein"**

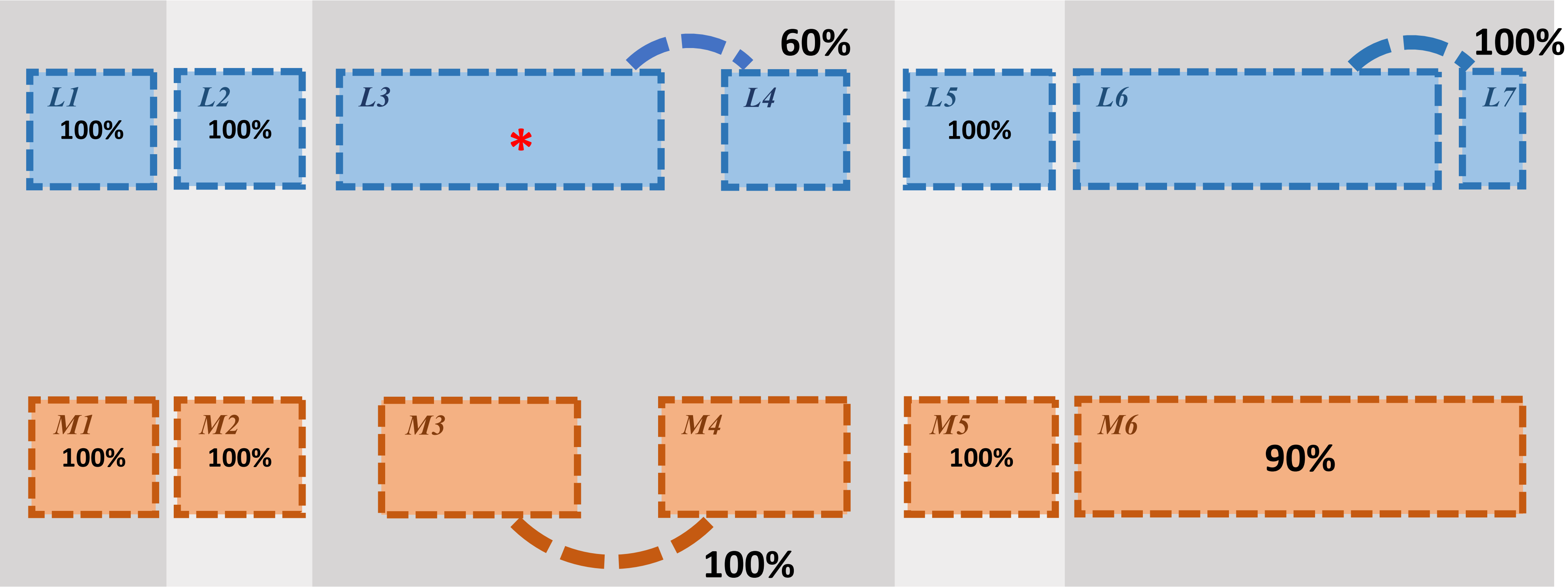
LiftOn CDS chaining



LiftOn: Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”

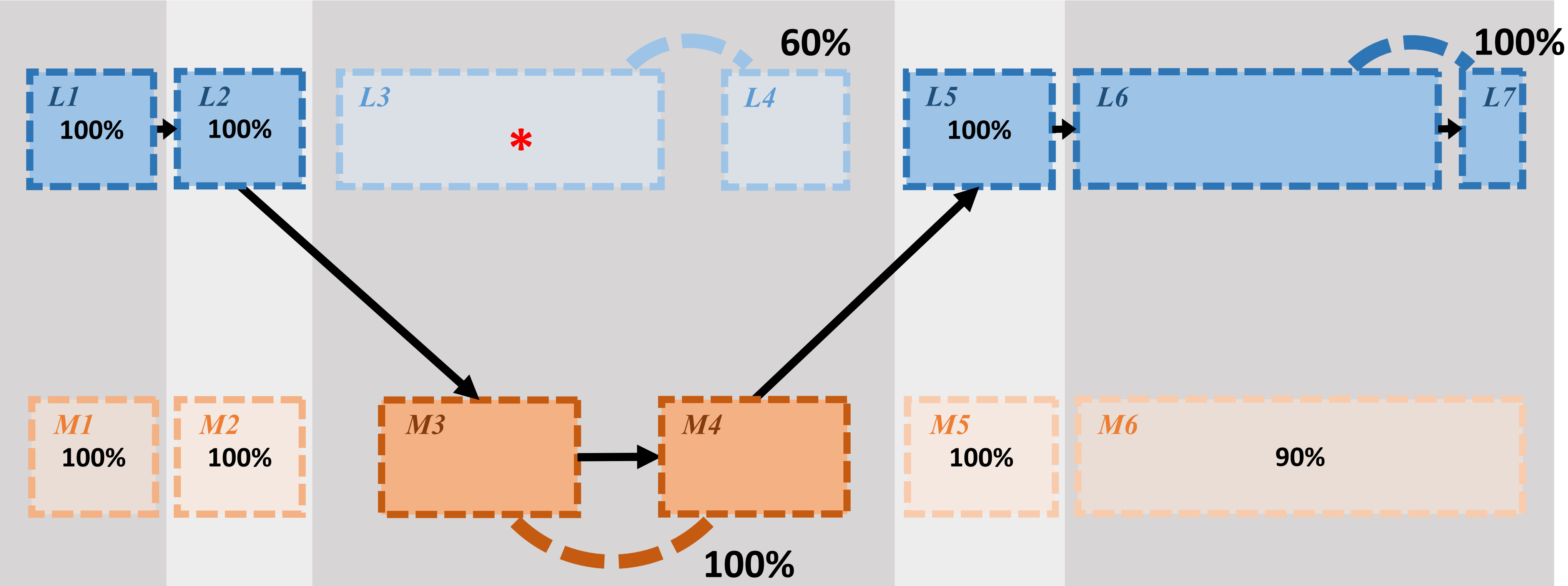
LiftOn CDS chaining



LiftOn: Protein-maximization algorithm

D Step 3: group CDSs by “accumulated AA in the reference protein”

LiftOn CDS chaining



Summary

- LiftOn uses both DNA-DNA alignments (from Liftoff) & protein-DNA alignments (from miniprot) to map annotations between genome assemblies of the same or different species.
- LiftOn's protein-maximization algorithm improves the annotation of protein-coding genes in the T2T- CHM13 genome.
- LiftOn can map annotation between relatively distant species, at least as divergent as mouse and rat.

Acknowledgement

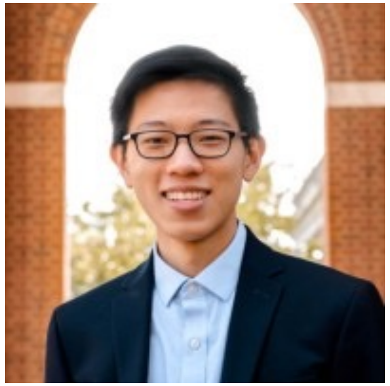
Steven Salzberg

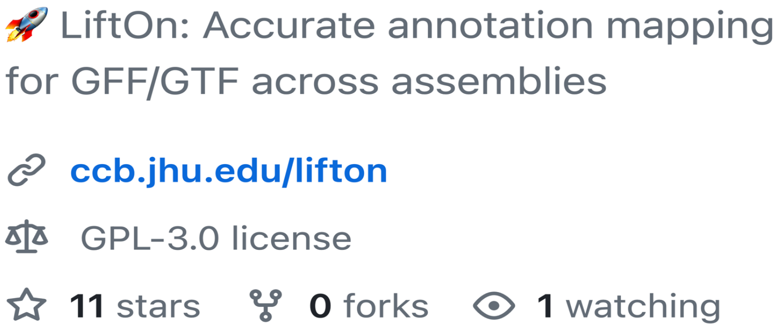
Mihaela Pertea

Alaina Shumate

Jakob Heinz

Celine Hoh


Alan Mao


🚀 LiftOn: Accurate annotation mapping for GFF/GTF across assemblies

ccb.jhu.edu/lifton

GPL-3.0 license

☆ **11** stars **0** forks **1** watching

LiftOn

Preprint coming soon!

ccb.jhu.edu/lifton

github.com/Kuanhao-Chao/LiftOn

Protein sequence identity

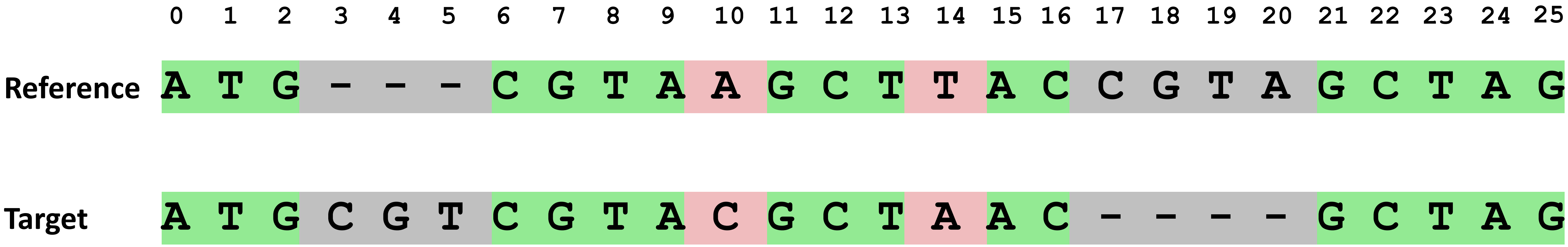
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Reference	M	G	L	V	-	-	-	-	R	W	S	Y	K	K	N	P	T	A	F	E	H	I	I	C	D	*
Target	M	G	L	V	R	W	S	S	R	W	S	Y	Q	K	N	P	T	A	-	-	H	I	-	C	D	*

$$\frac{\#Matched\_AA}{\#alignment\ column\ -\ \#gaps\ in\ reference\ protein} = \frac{18}{26-4} = 81.8\%$$

Do not penalize longer proteins

S1

DNA sequence identity



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Reference	A	T	G	-	-	-	C	G	T	A	A	G	C	T	T	A	C	C	G	T	A	G	C	T	A	G
Target	A	T	G	C	G	T	C	G	T	A	C	G	C	T	A	A	C	-	-	-	-	G	C	T	A	G

$$\frac{\#Matched\_nucleotide}{\#alignment\ column} = \frac{17}{26} = 65.4\%$$

S2