JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING
Department of Computer Science

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

Combining DNA and protein alignments to improve genome annotation with LiftOn

Kuan-Hao Chao

2024.07.15

khchao.com

@KuanHaoChao

Kuanhao-Chao

Genome annotation

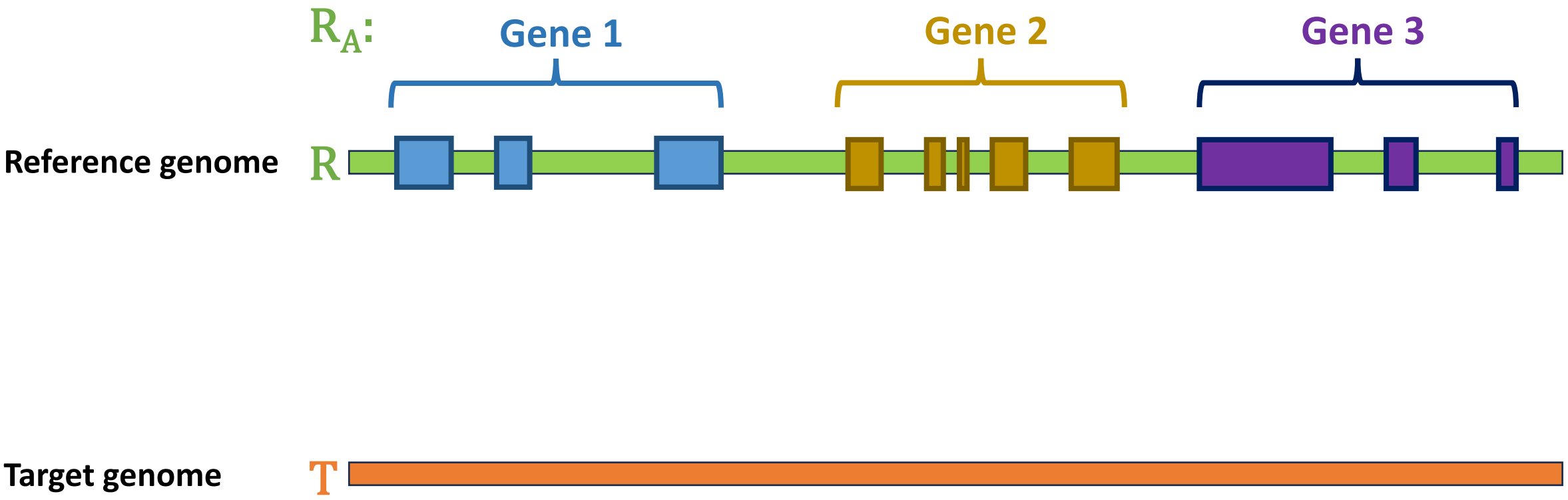
Genome (FASTA)

```
CAGCCCCCGGAGACTtaaatacaggaagaaaaaggCAGGACAGAATTACAAGGTGCTGGCCCAGGGCGGGCAGCGGCCCT
GCCTCCTACCCTTGCGCCTCATGACCAGCTTGTTGAAGAGATCCGACATCAAGTGCCCACCTTGGCTCGTGGCTCTCACT
GCAACGGGAAAGCCACAGACTGGGGTGAAGAGTTCAGTCACATGCGACCGGTgactccctgtccccacccccatgACACT
CCCCAGCCCTCCAAGGCCACTGTGTTTCCCAGTTAGCTCAGAGCCTCAGTCGATCCCTGACCCAGCACCGGGCACTGATG
AGACAGCGGCTGTTTGAGGagccacctcccagccacctcggggccagggccagggtgtGCAGCACCACTGTACAATGGGG
AAACTGGCCCAGAGAGGTGAGGCAGCTTGCCTGGGGTCACAGAGCAAGGCAAAAGCAGCGCTGGGTACAAGCTCAAAACC
ATAGTGCCCAGGGCACTGCCGCTGCAGGCGCAGGCATCGCATCACACCAGTGTCTGCGTTCACAGCAGGCATCATCAGTA
```

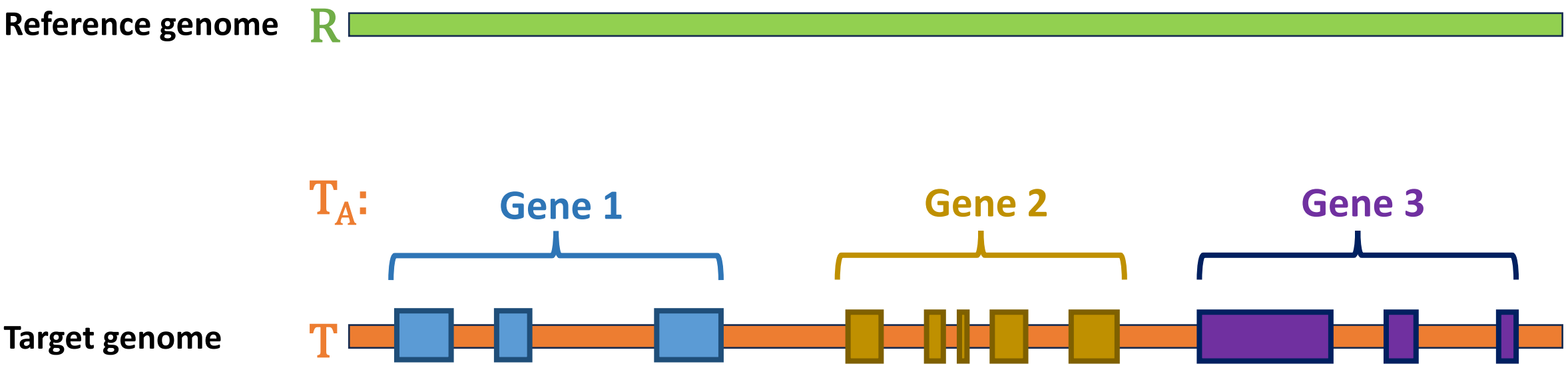
Annotation (GFF / GTF)

```
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:



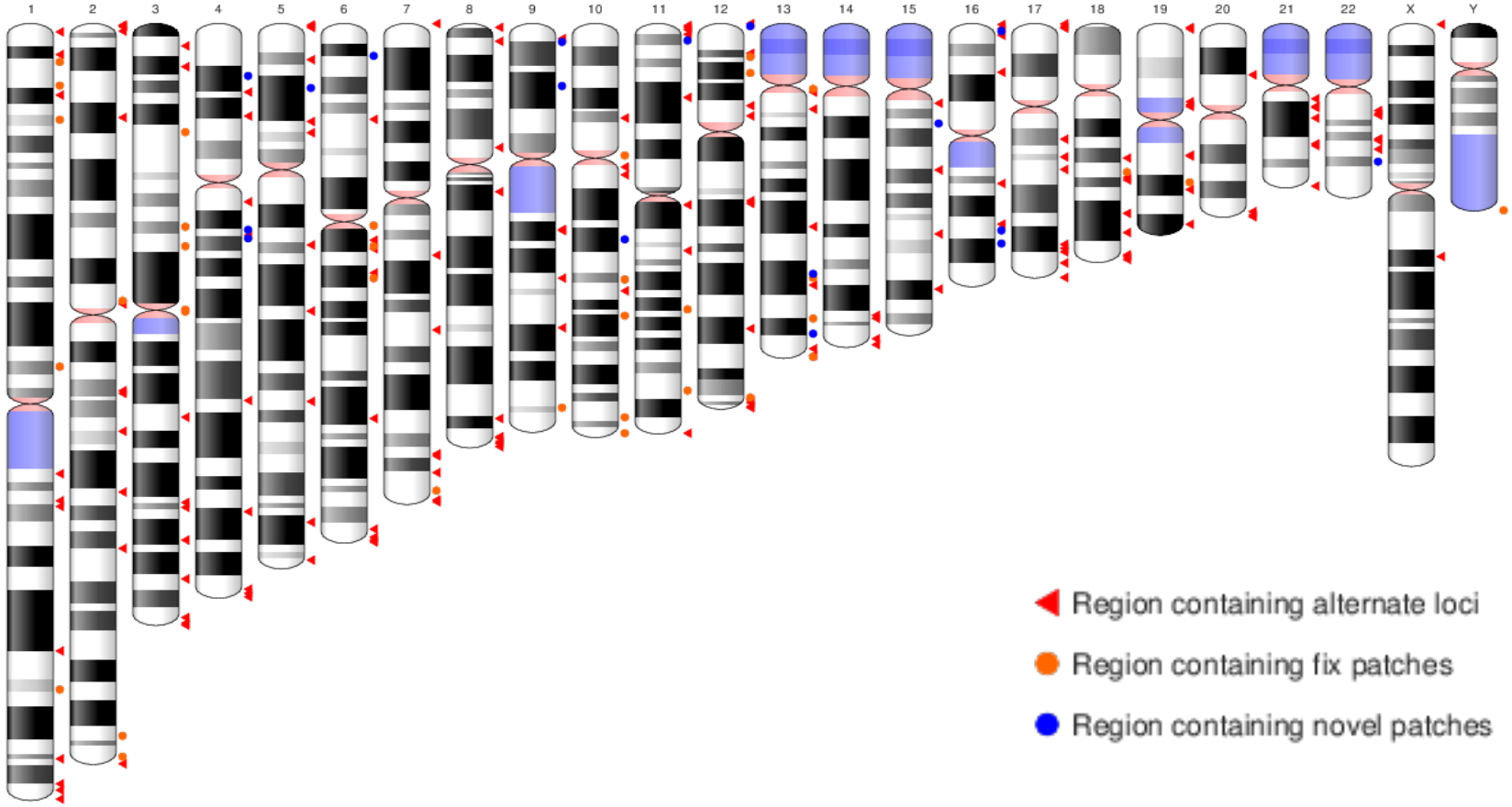
Lift-over Problem Definition:



https://www.sanger.ac.uk/data/genome-reference-consortium/

Application: GRCh38 to T2T-CHM13 lift-over

“**Finished**” the human genome project




nature

Explore content · About the journal · Publish with us

nature › article

Original Article | Published: 01 February 2001

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

Science

Current Issue · First release papers

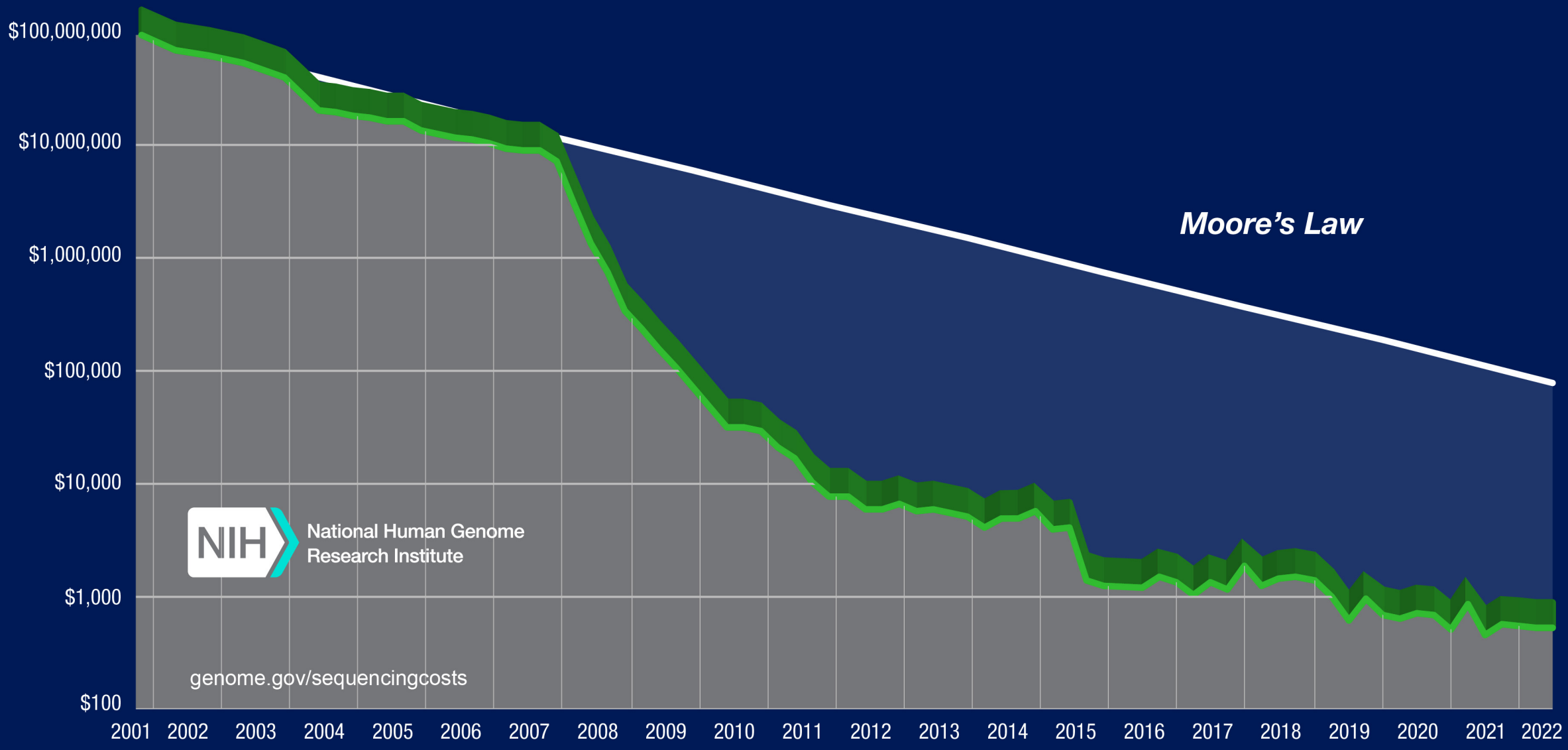
HOME › SCIENCE › VOL. 291, NO. 5507 › THE SEQUENCE OF THE HUMAN GENOME

SPECIAL REVIEWS

The Sequence of the Human Genome

J. CRAIG VENTER, MARK D. ADAMS, EUGENE W. MYERS, PETER W. LI, [...], AND XIAOHONG ZHU +269 authors Authors Info & Affiliations

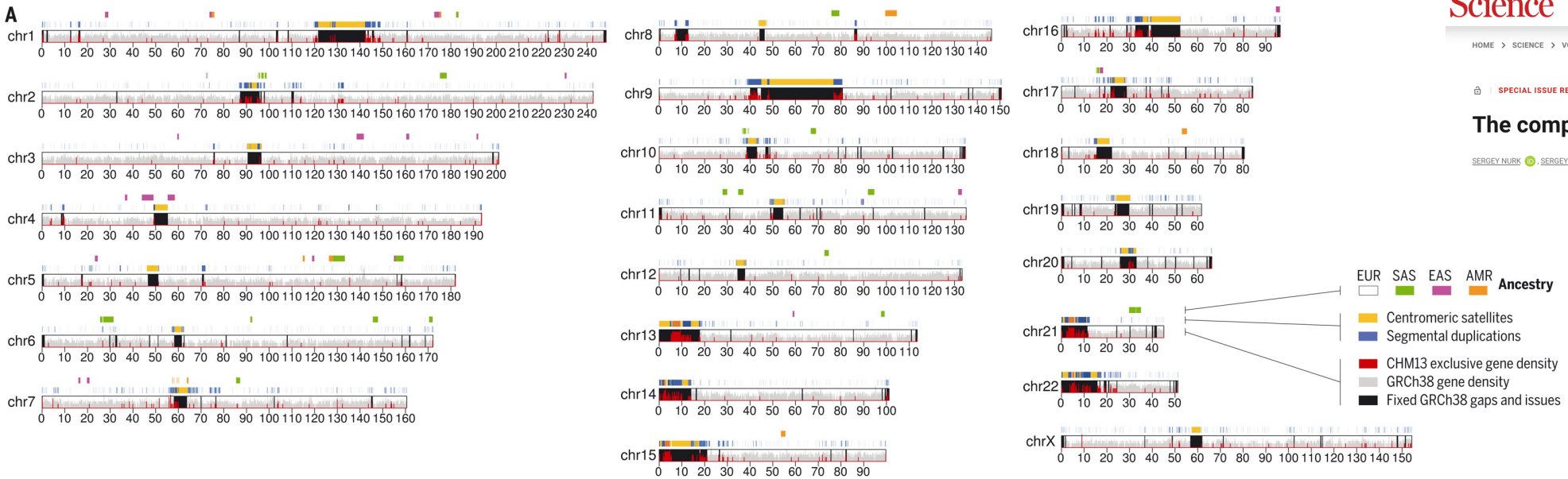
Cost per Human Genome



https://youtu.be/MbYvTyIMc84?si=YjijRyLwq31MY5HM

Application: GRCh38 to T2T-CHM13 lift-over

- 238 Mbp added and corrected
- 180 Mbp of centromeric satellites
- 68 Mbp of segmental duplications
- 10 Mbp of rDNAs
- 182 Mbp of entirely novel sequence
- 1956 novel genes including 99 protein-coding




Science

Current Issue First release p

HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, [...], AND ADAM M. PHILLIPPY +95 authors

If you were to use a CHM13 annotation …

Which lift-over tool to use?

Telomere-to-Telomere (T2T) consortium slack channel

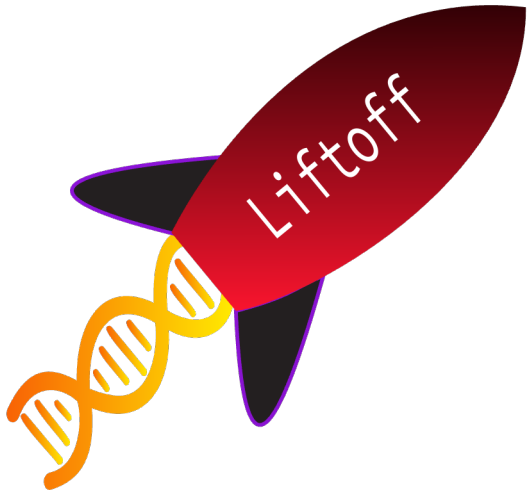
Giulio Formenti 3:44 PM

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

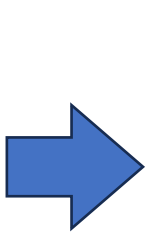
Liftoff: accurate mapping of gene annotations FREE

Alaina Shumate ✉, Steven L Salzberg

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

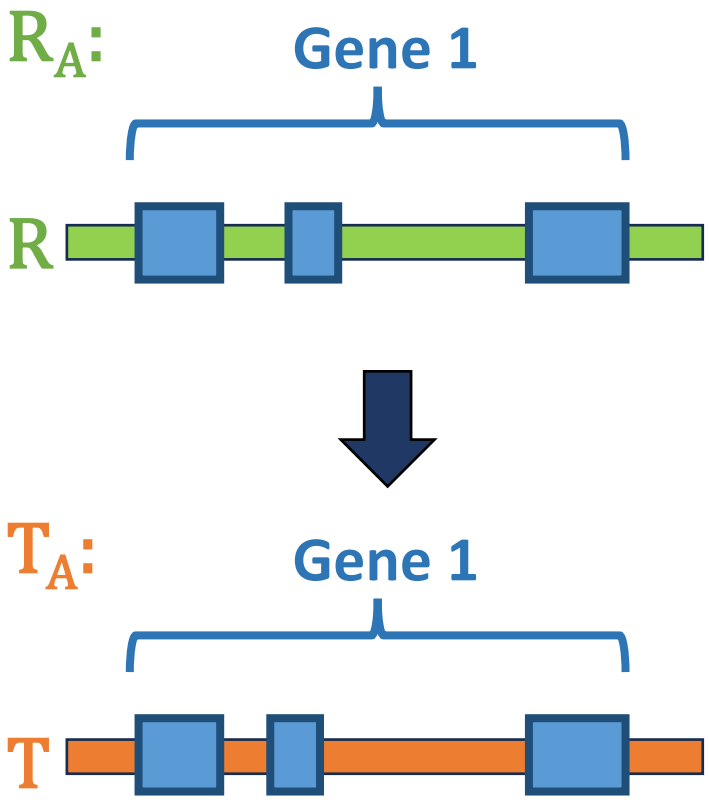
Published: 09 May 2021 **Article history** ▾

~ 390 citation



LiftOn

Lift-over problem, what methods are available?



Chain file

UCSU liftOver

Kuhn et al.

CrossMap

Zhao et al.

DNA-based

Liftoff

github.com/agshumate/Liftoff

Shumate et al.

minimap2

github.com/lh3/minimap2

Li

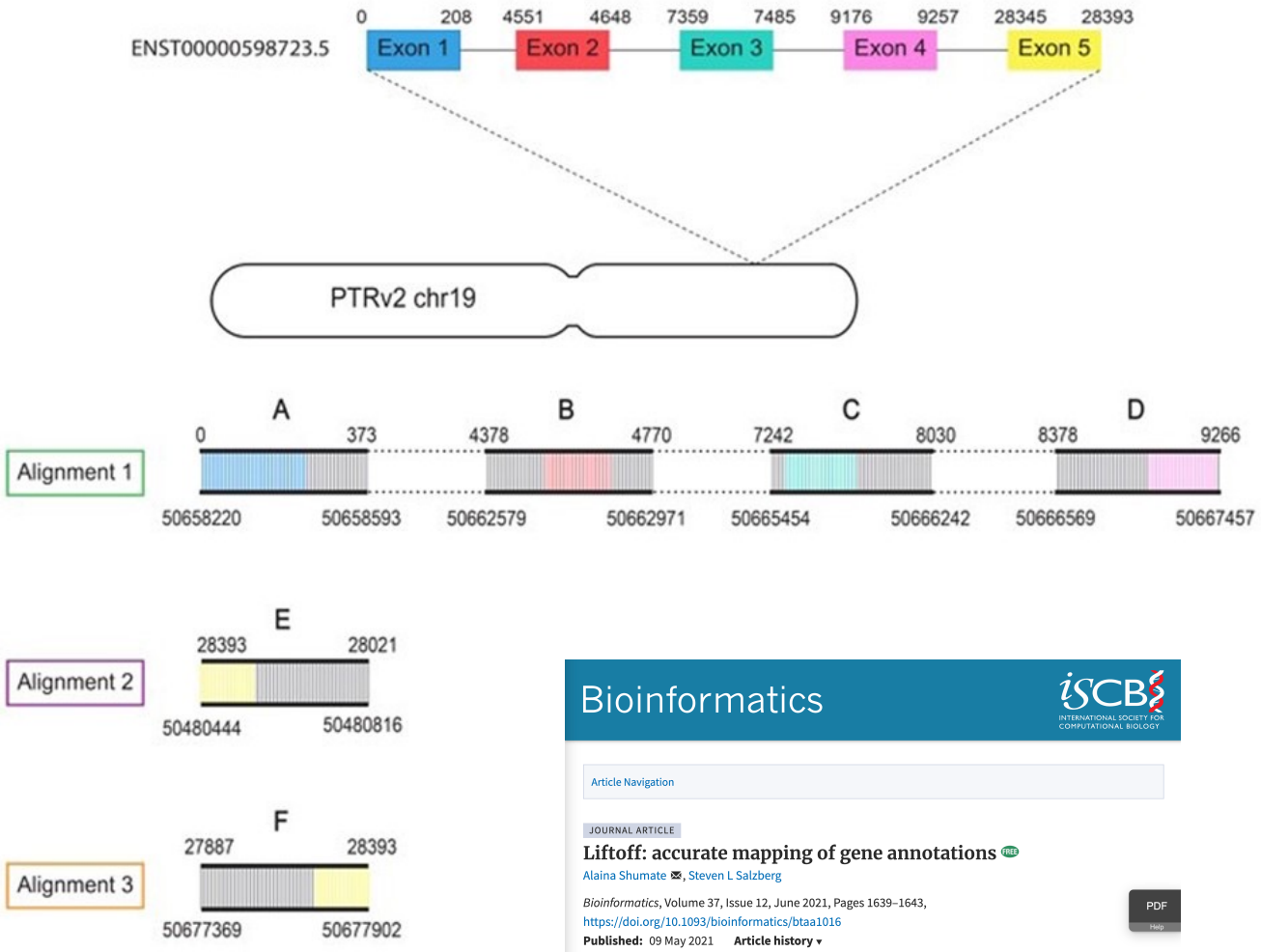
Protein-based

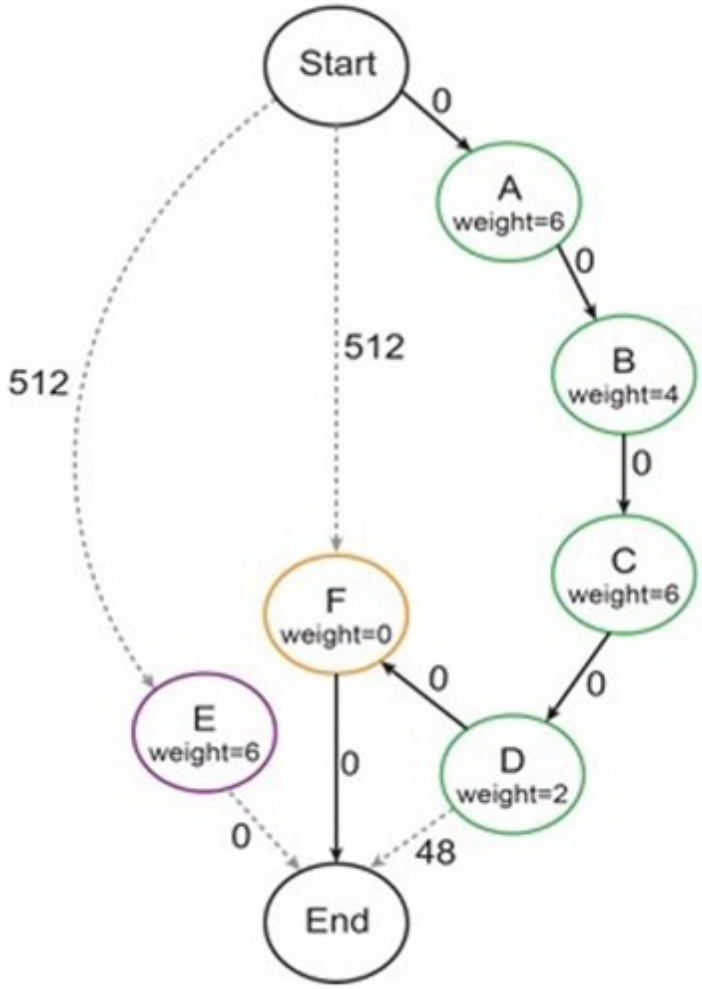
miniprot

github.com/lh3/miniprot

Li

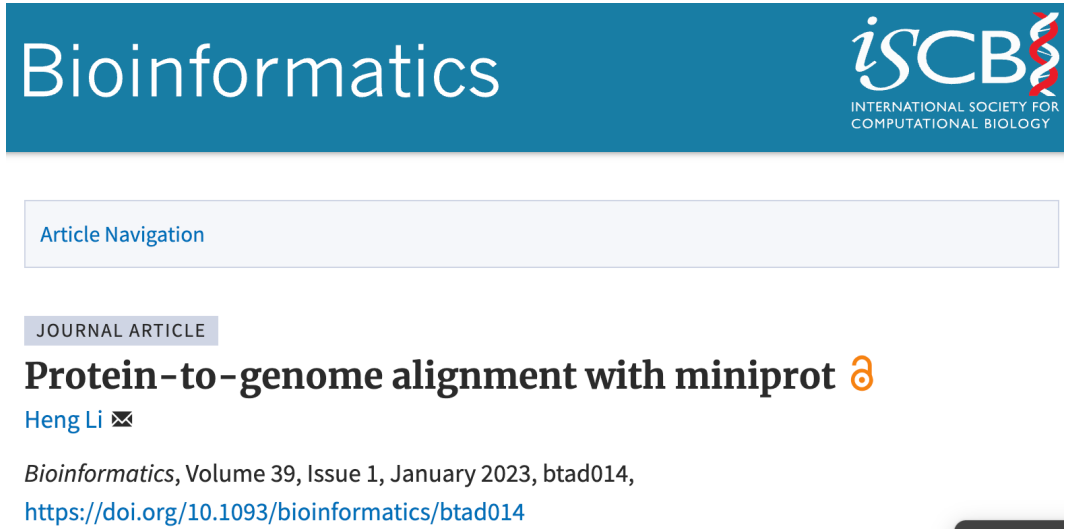
Methods – Liftoff / minimap2





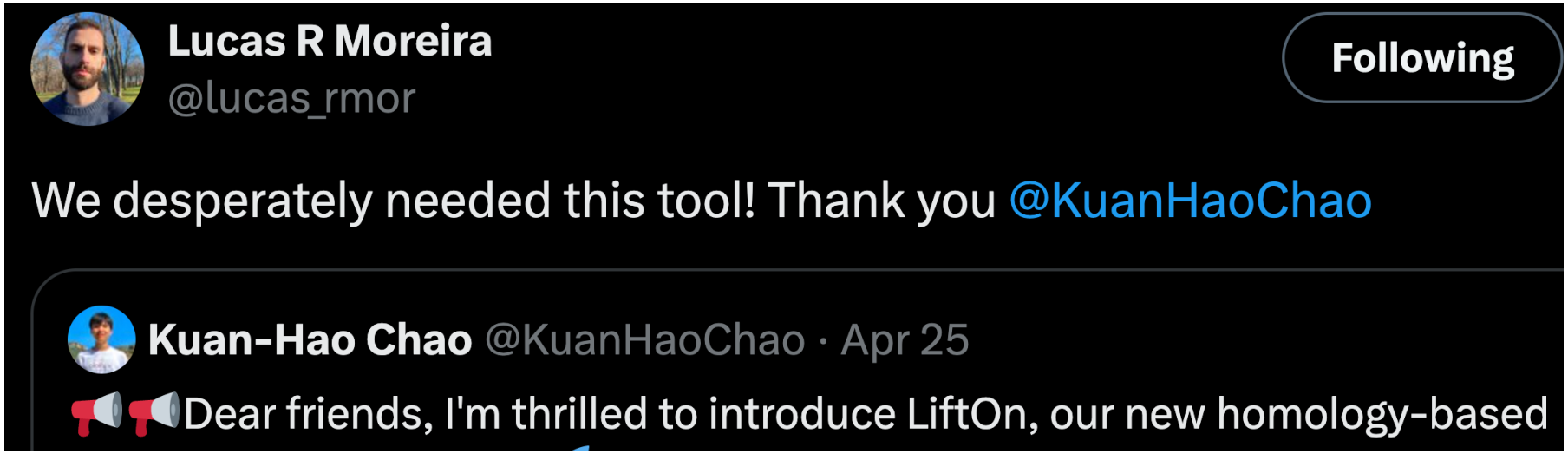
Methods – miniprot

- **Index step:**
  - Translate reference genome to amino acids in 6 phases, filter out ORFs
- **Initial chaining:**
  - Extract 6-mer syncmers from protein query
  - Look up index for seed matches
- **Refined chaining**
  - Redo seeding and chaining
  - Sliding 5-mers from reference and protein
- **Final alignment**



Methods – LiftOn

- How can we do better?
  - Combining DNA- and protein-based alignments!

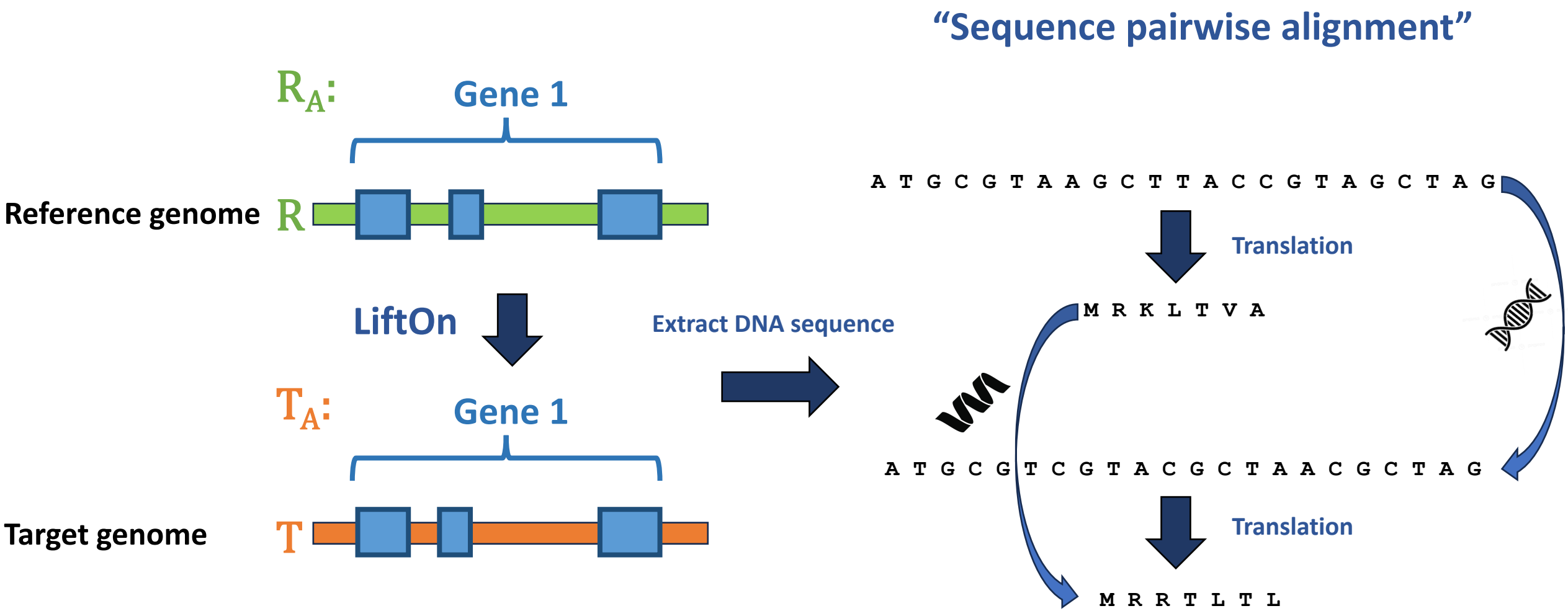


LiftOn

Results

- outperforms state-of-the-art DNA- and protein-based liftover methods
- improves the annotation of protein-coding genes in T2T-CHM13 genome
- Improves annotation lift-over between distant species, such as mouse and rat

Evaluation Metrics



https://lh3.github.io/2018/11/25/on-the-definition-of-sequence-identity

Evaluation Metrics

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

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

https://lh3.github.io/2018/11/25/on-the-definition-of-sequence-identity

Evaluation Metrics

Protein sequence identity

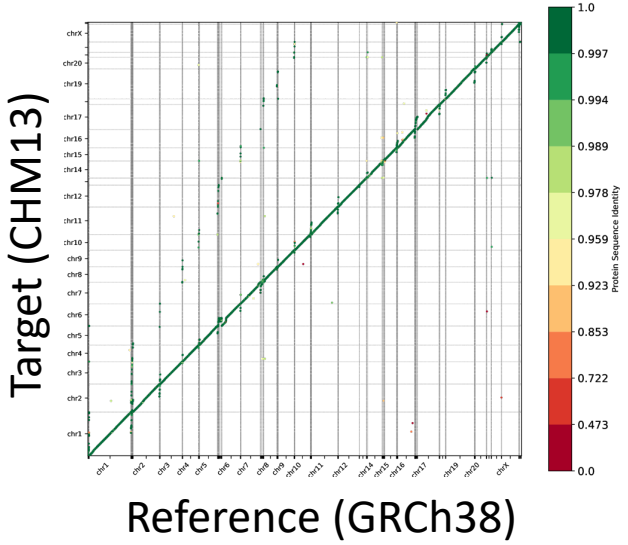
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Reference	M	G	L	V	-	-	-	-	R	W	S	Y	K	K	N	P	T	A	F	E	H	I	I	C	D	*
Target	M	G	L	V	R	W	S	S	R	W	S	Y	Q	K	N	P	T	A	-	-	H	I	-	C	D	*

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

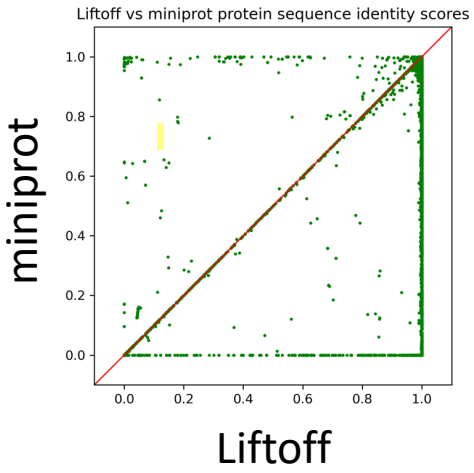
“Do not penalize longer proteins”

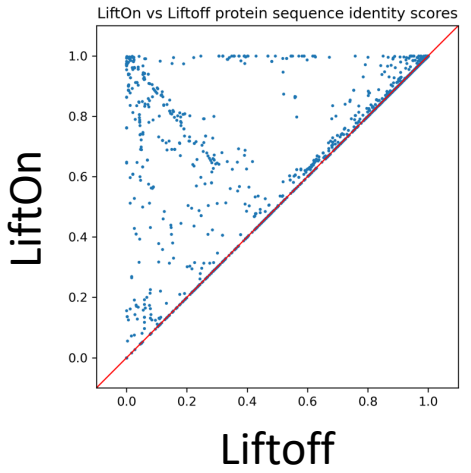
Result 1: improves DNA & protein-based lift-over

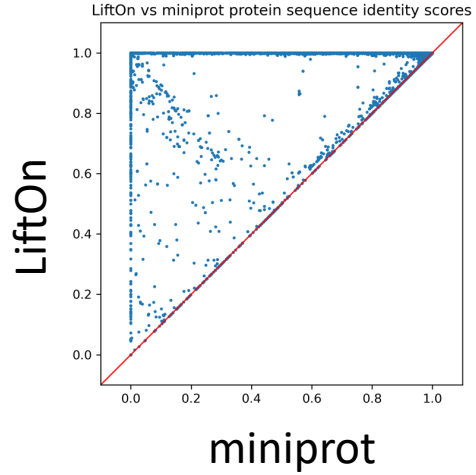
Map RefSeq v220 from GRCh38 -> CHM13V2.0

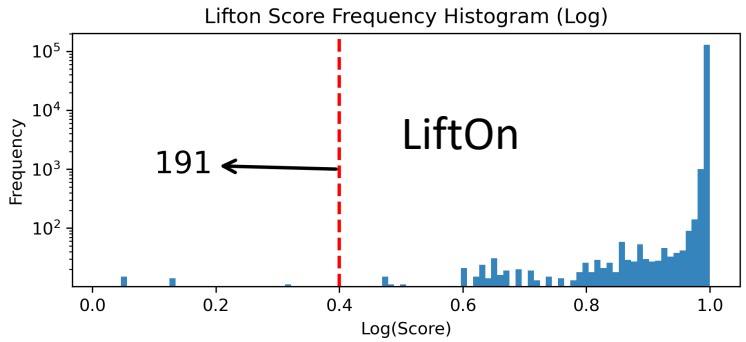


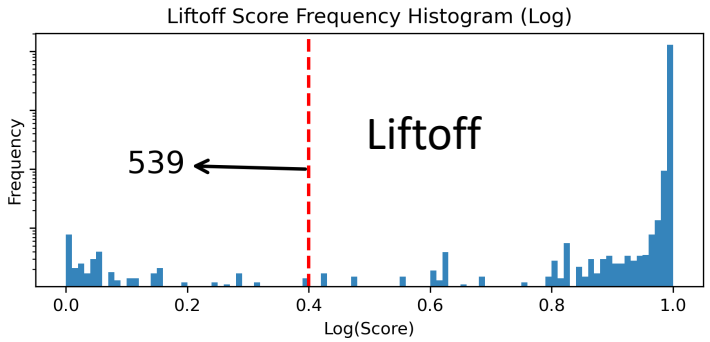
Compressed-gap protein sequence identity

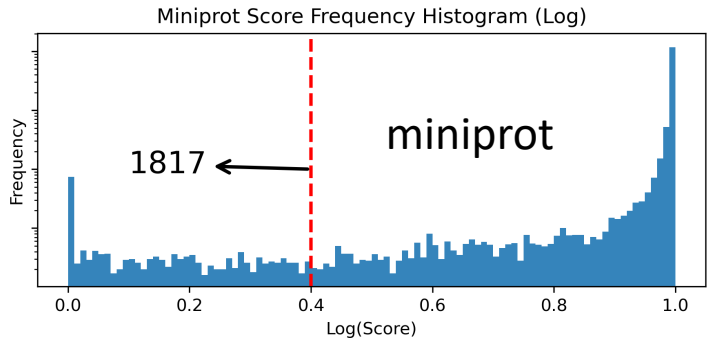








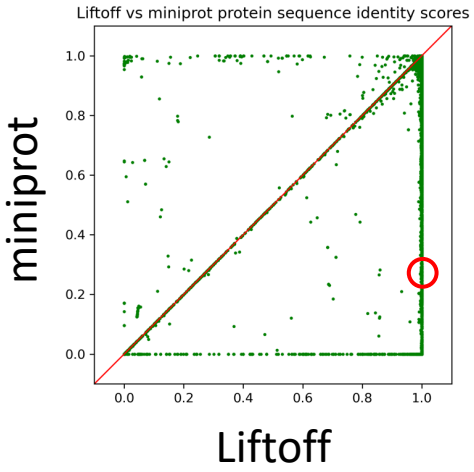


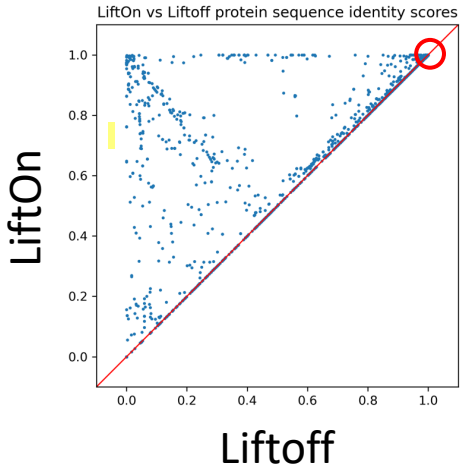


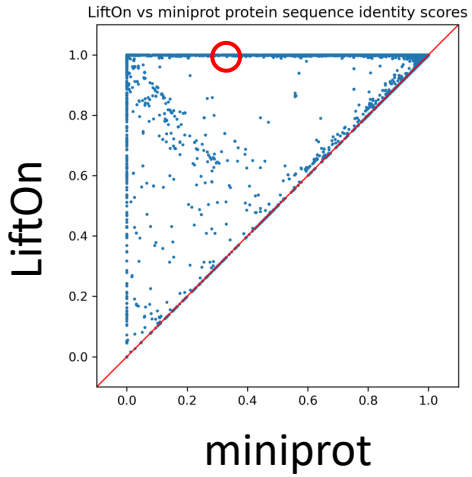
Result 1: improves DNA & protein-based lift-over

Map RefSeq v220 from GRCh38 -> CHM13V2.0

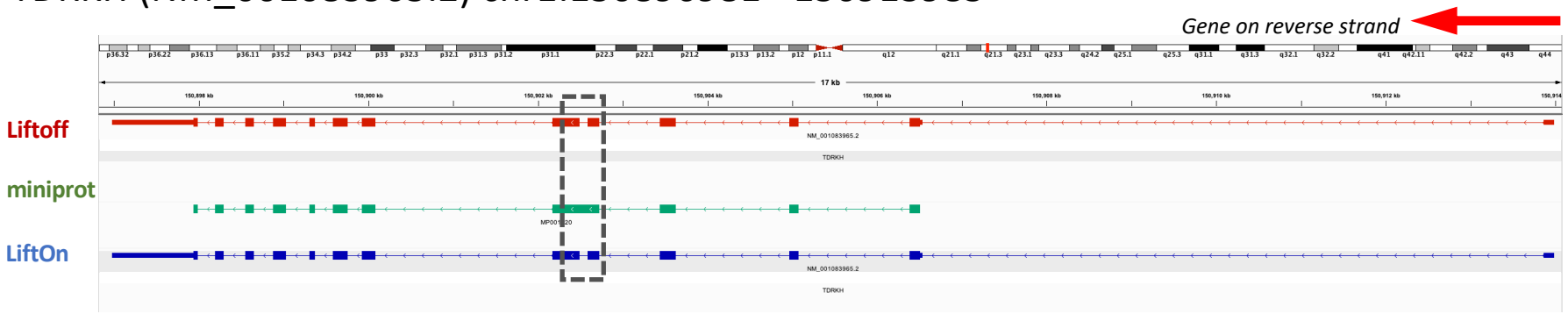
Compressed-gap protein sequence identity







TDRKH (NM_001083965.2) chr1:150896981 - 150913985

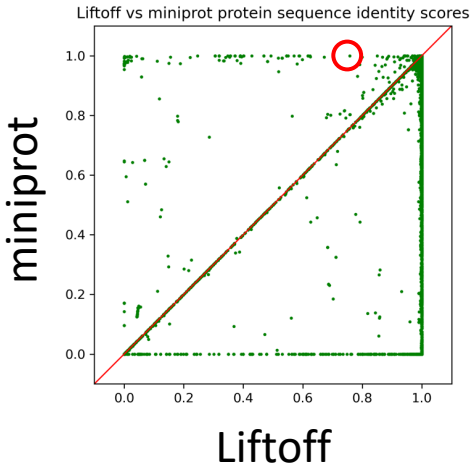


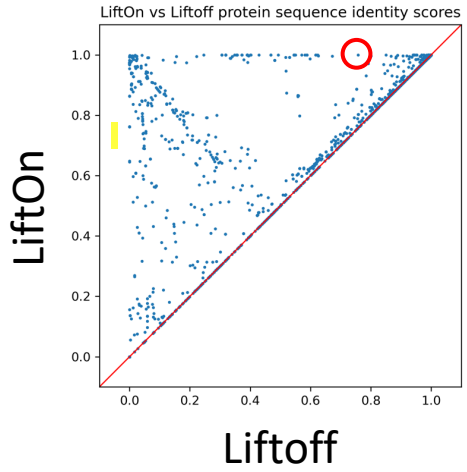
	Protein identity
Liftoff	100%
miniprot	31.48%
LiftOn	100%

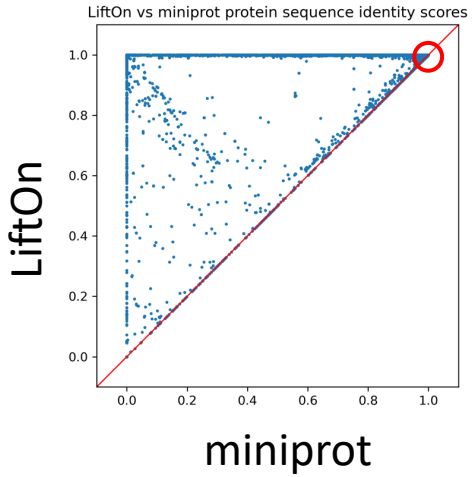
Result 1: improves DNA & protein-based lift-over

Map RefSeq v220 from GRCh38 -> CHM13V2.0

Compressed-gap protein sequence identity







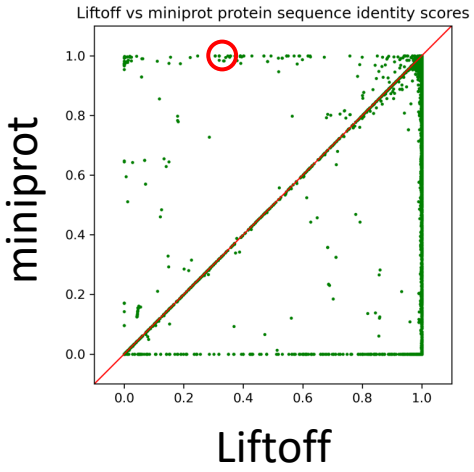
SLC22A31 (NM_001384763.1) chr16:95276205 - 95280662

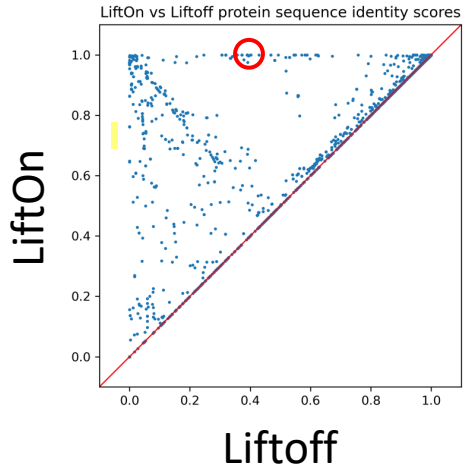
	Protein identity
Liftoff	76.70%
miniprot	99.55%
LiftOn	99.55%

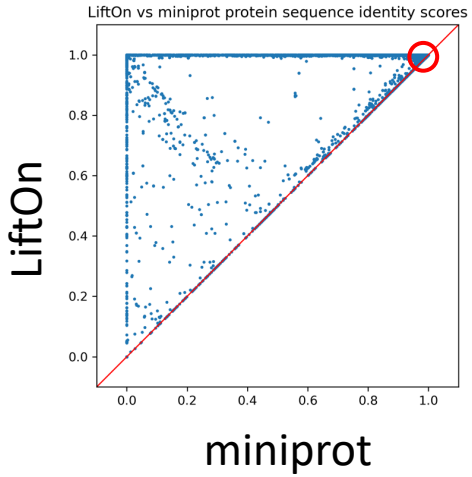
Result 1: improves DNA & protein-based lift-over

Map RefSeq v220 from GRCh38 -> CHM13V2.0

Compressed-gap protein sequence identity

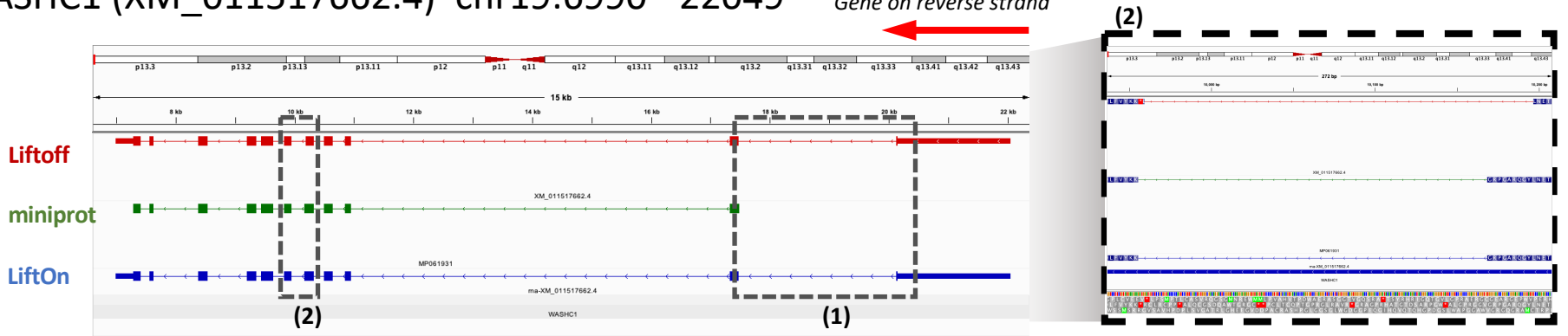






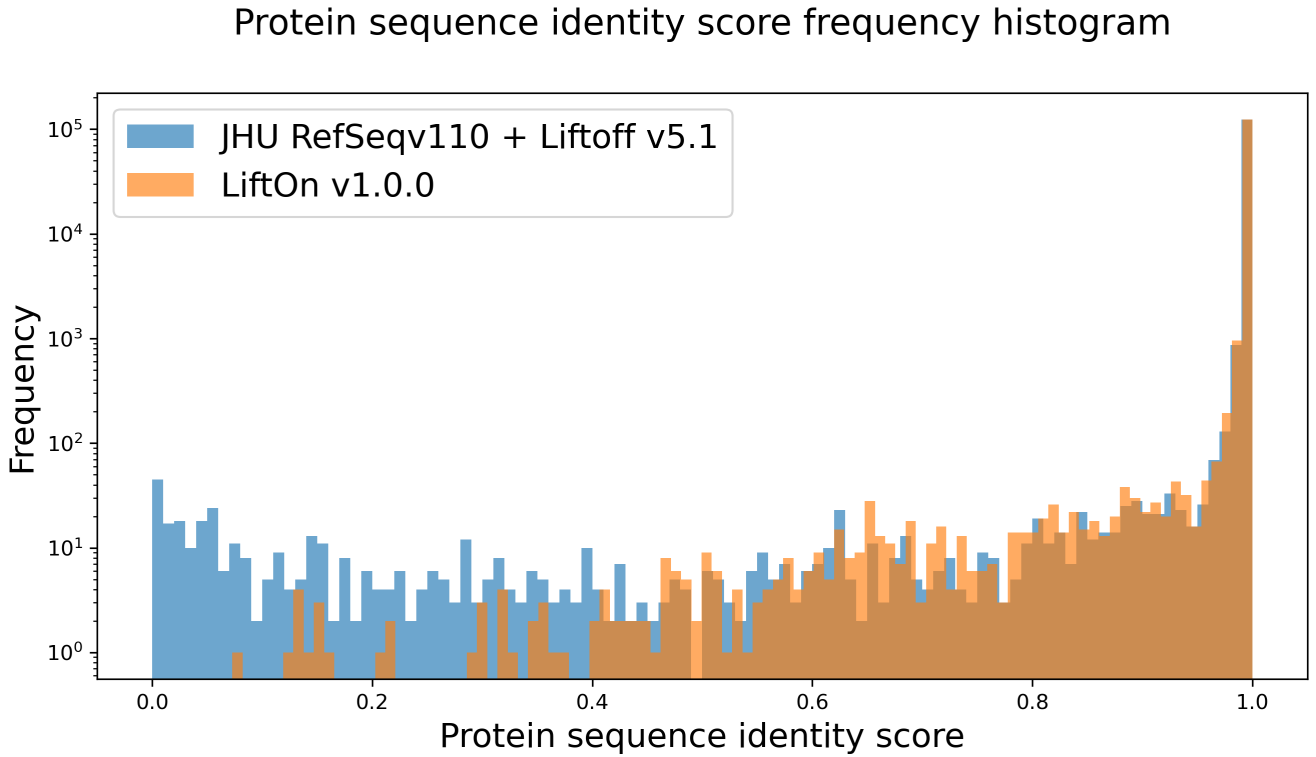
WASHC1 (XM_011517662.4) chr19:6990 - 22049

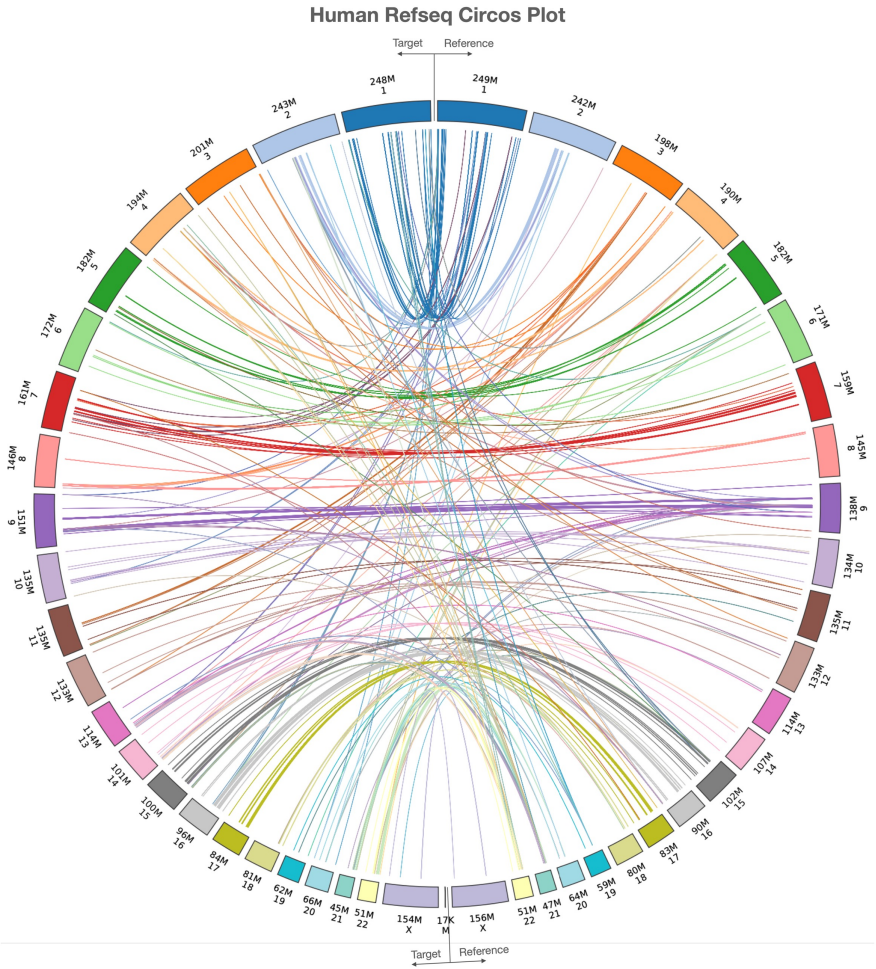
Gene on reverse strand



	Protein identity
Liftoff	38.92%
miniprot	99.14%
LiftOn	99.35%

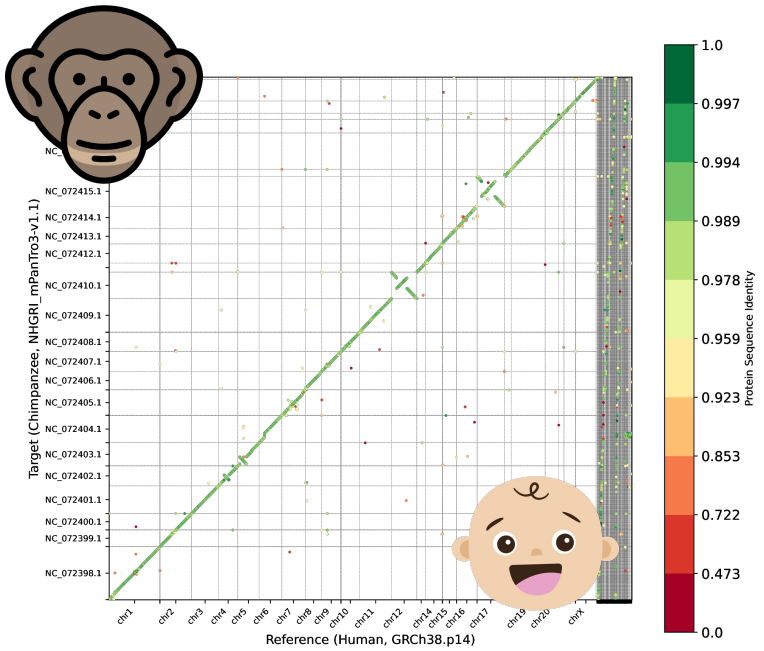
Result 2: improve CHM13 protein annotations





Result 3: improve distant species lift-over

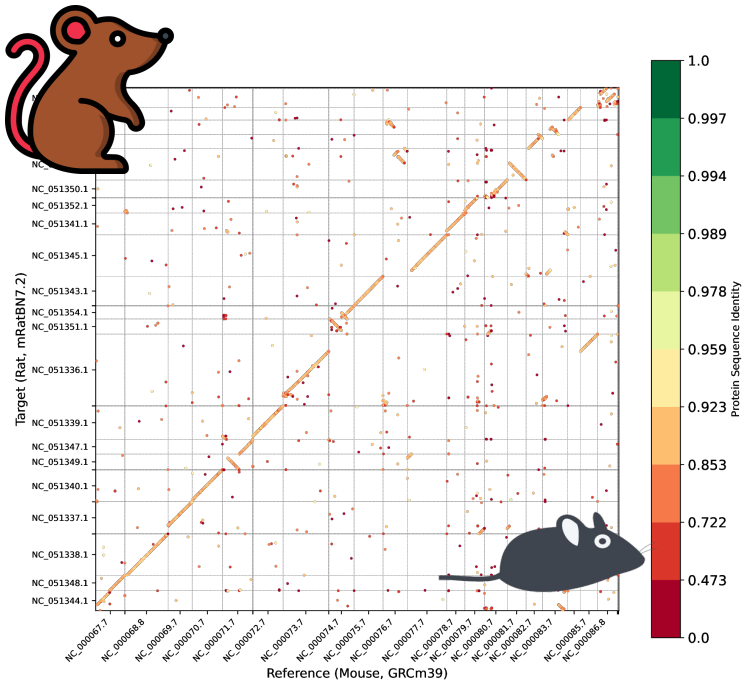
human to chimp



Mash : 0.013

Dashing2 : 0.47

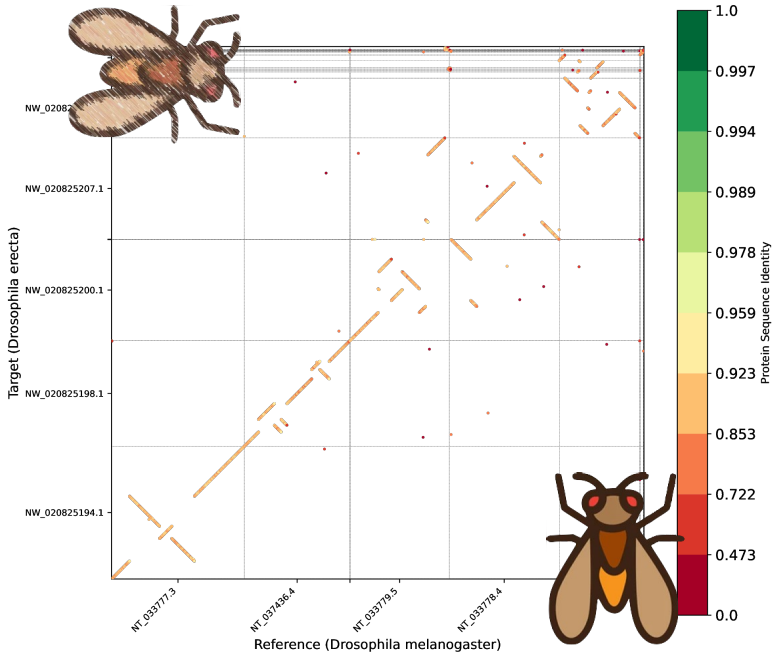
mouse to rat



Mash : 0.120

Dashing2 : 0.01

Drosophila m. to *Drosophila e.*



Mash : 0.077

Dashing2 : 0.07

Method

Genomic sketching with multiplicities and locality-sensitive hashing using Dashing 2

Daniel N. Baker and Ben Langmead

Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218-2683, USA

Ondov *et al. Genome Biology* (2016) 17:132
DOI 10.1186/s13059-016-0997-x

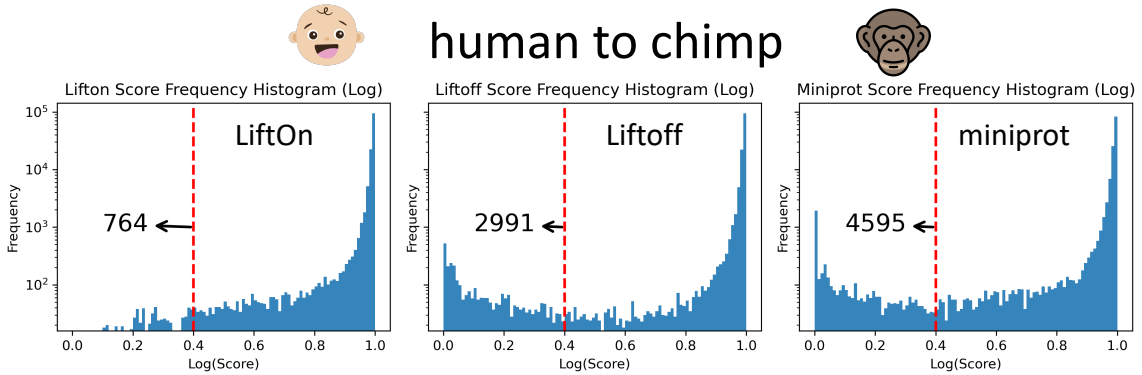
Genome Biology

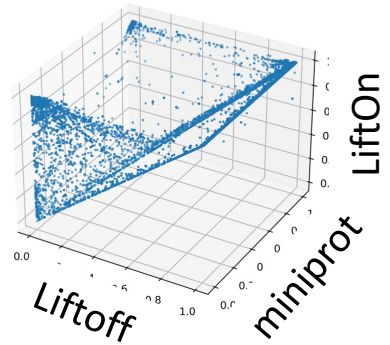
SOFTWARE Open Access

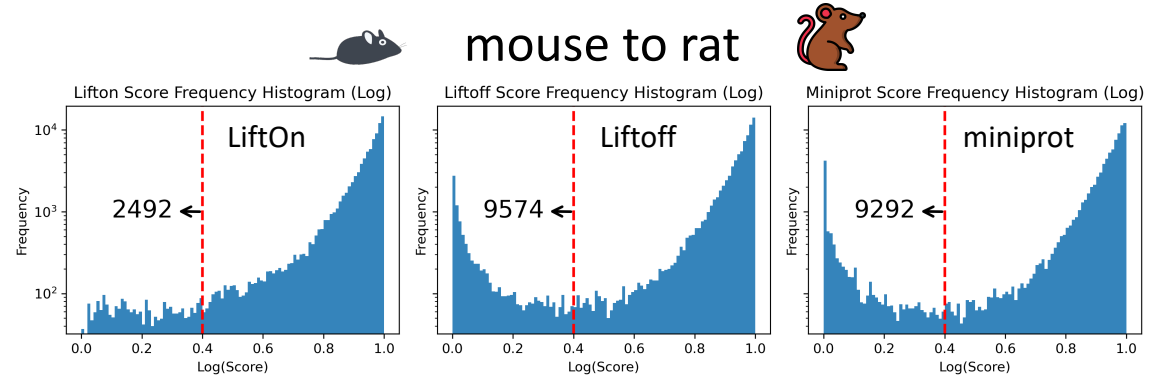
Mash: fast genome and metagenome distance estimation using MinHash

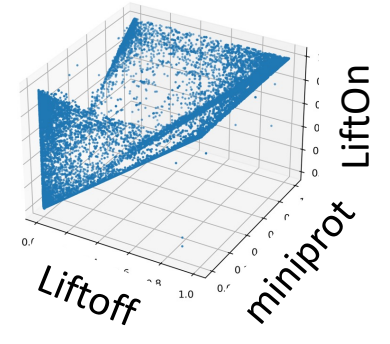
Brian D. Ondov[1], Todd J. Treangen[1], Páll Melsted[2], Adam B. Mallonee[1], Nicholas H. Bergman[1], Sergey Koren[3] and Adam M. Phillippy[3*]

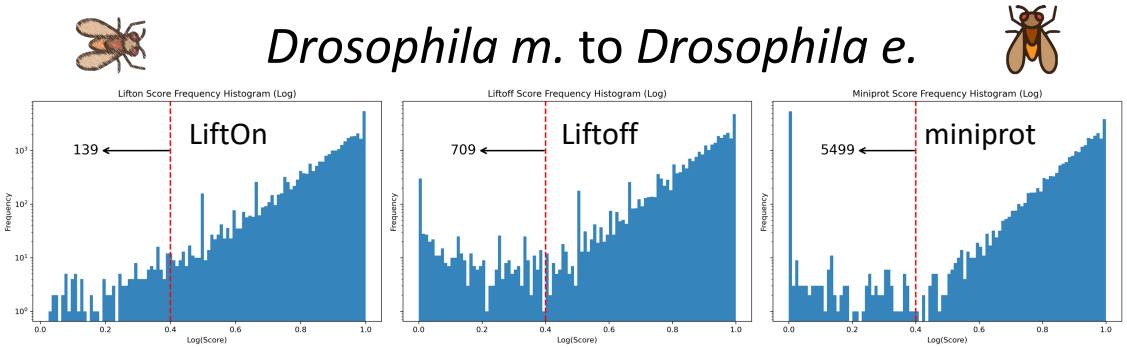
Result 3: improve distant species lift-over

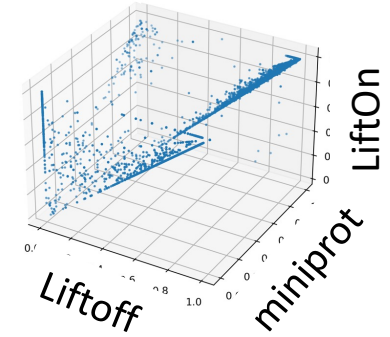












Result 3: improve distant species lift-over

House mouse
(*Mus musculus*)

Yeast
(*Saccharomyces cerevisiae*)

Honey bee
(*Apis mellifera*)

Thale cress
(Arabidopsis thaliana)

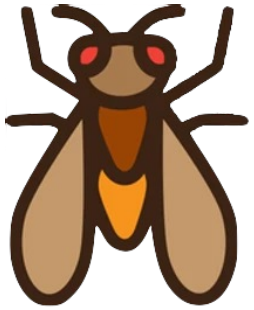
New Results

Follow this preprint

Combining DNA and protein alignments to improve genome annotation with LiftOn

Kuan-Hao Chao, Jakob M. Heinz, Celine Hoh, Alan Mao, Alaina Shumate, Mihaela Pertea, Steven L Salzberg

doi: https://doi.org/10.1101/2024.05.16.593026



fruit fly
(*Drosophila melanogaster*)

Rice
(Oryza sativa)

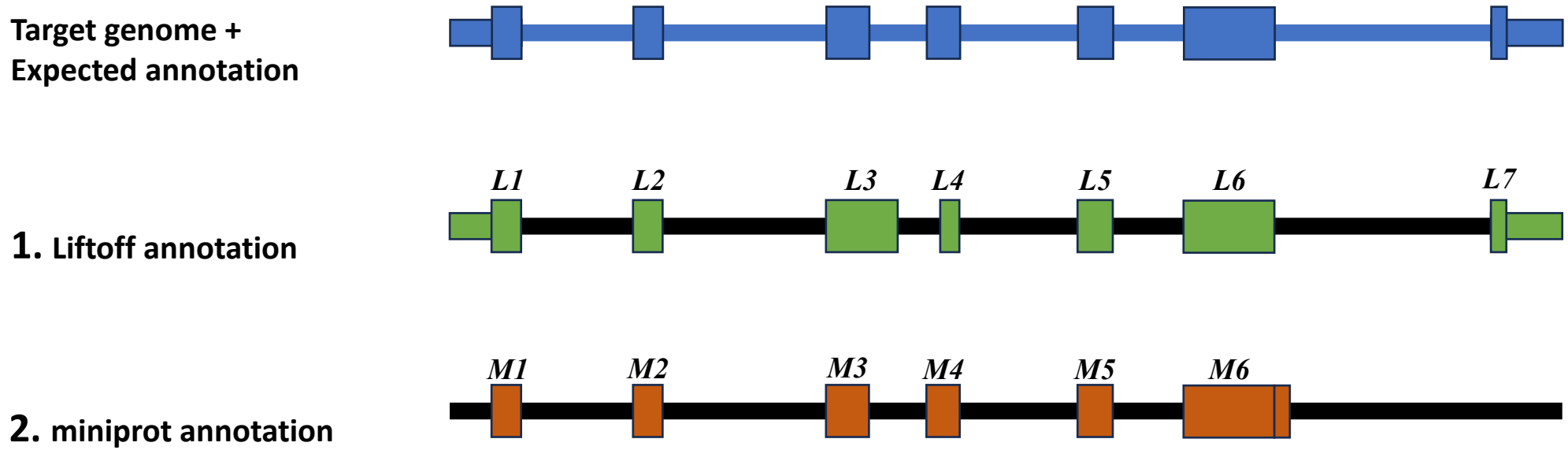
LiftOn

Methods in Details

- Protein-maximization algorithm
  - Step 1: chaining CDSs
  - Step 2: ORF search

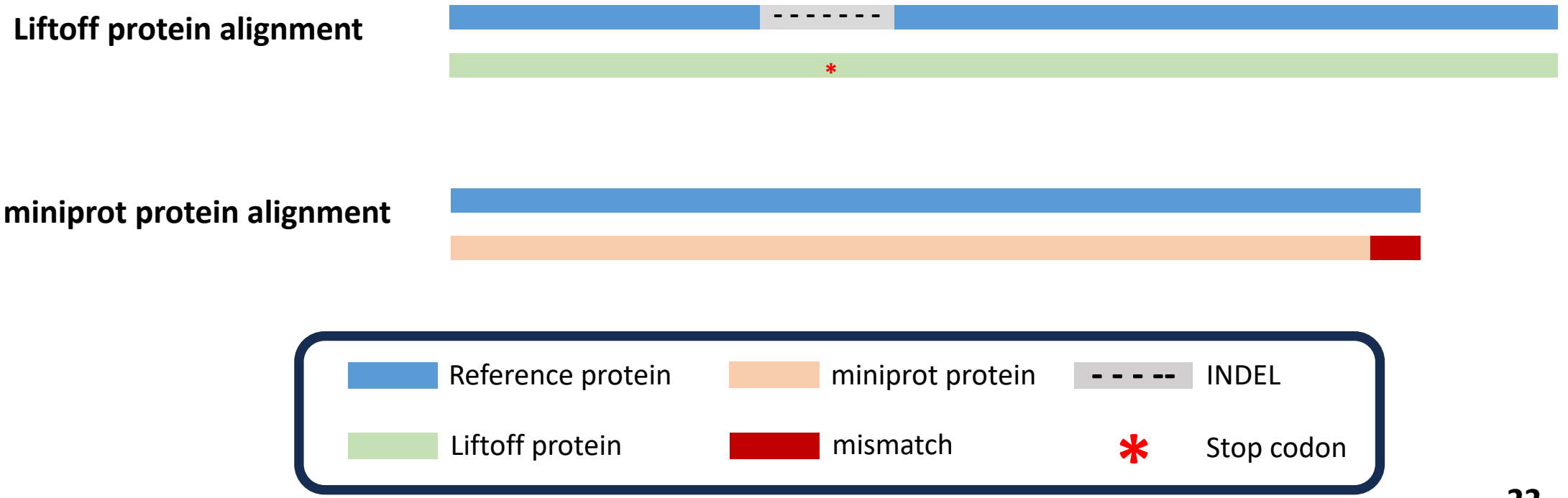
LiftOn: Protein-maximization algorithm

A



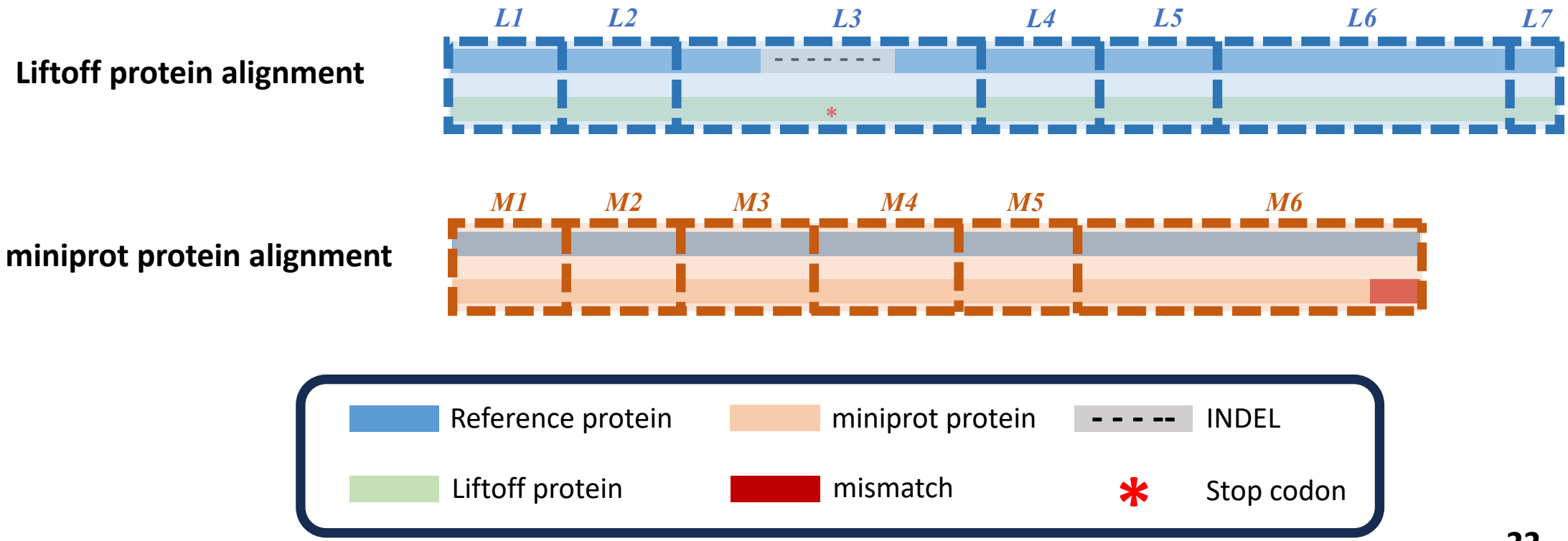
LiftOn: Protein-maximization algorithm

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



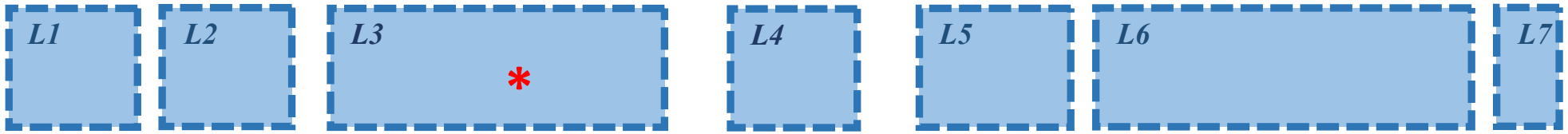
LiftOn: Protein-maximization algorithm

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



LiftOn: Protein-maximization algorithm

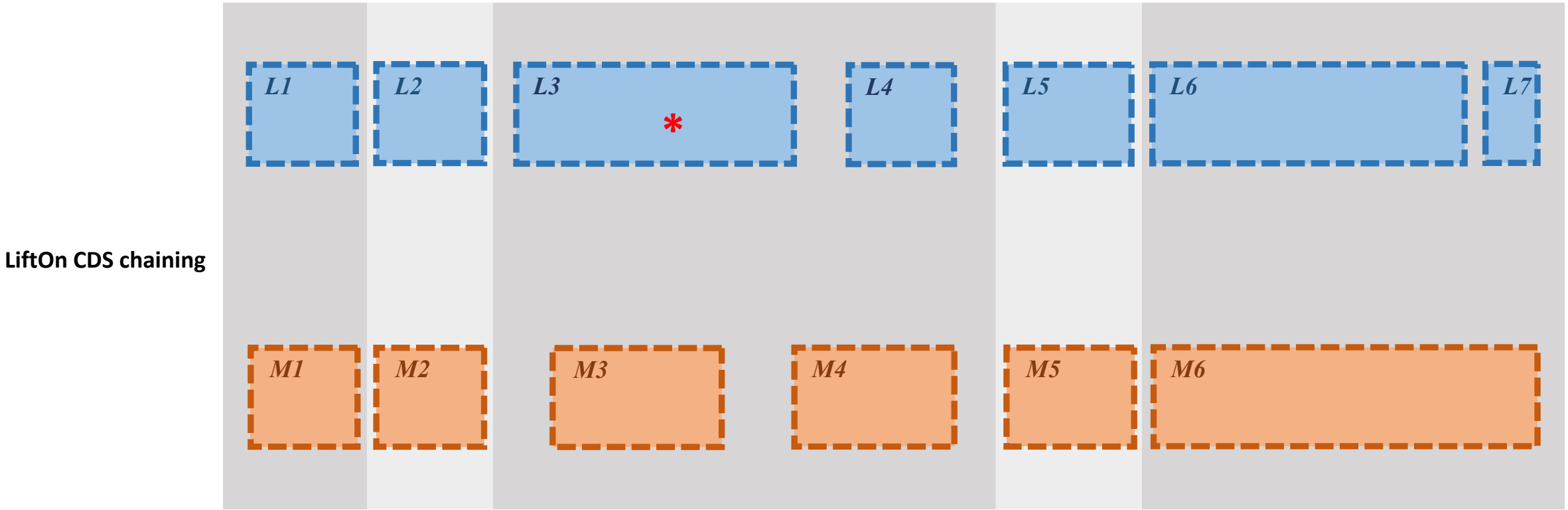
D **Step 3: group CDSs by “accumulated AA in the reference protein”**



LiftOn CDS chaining

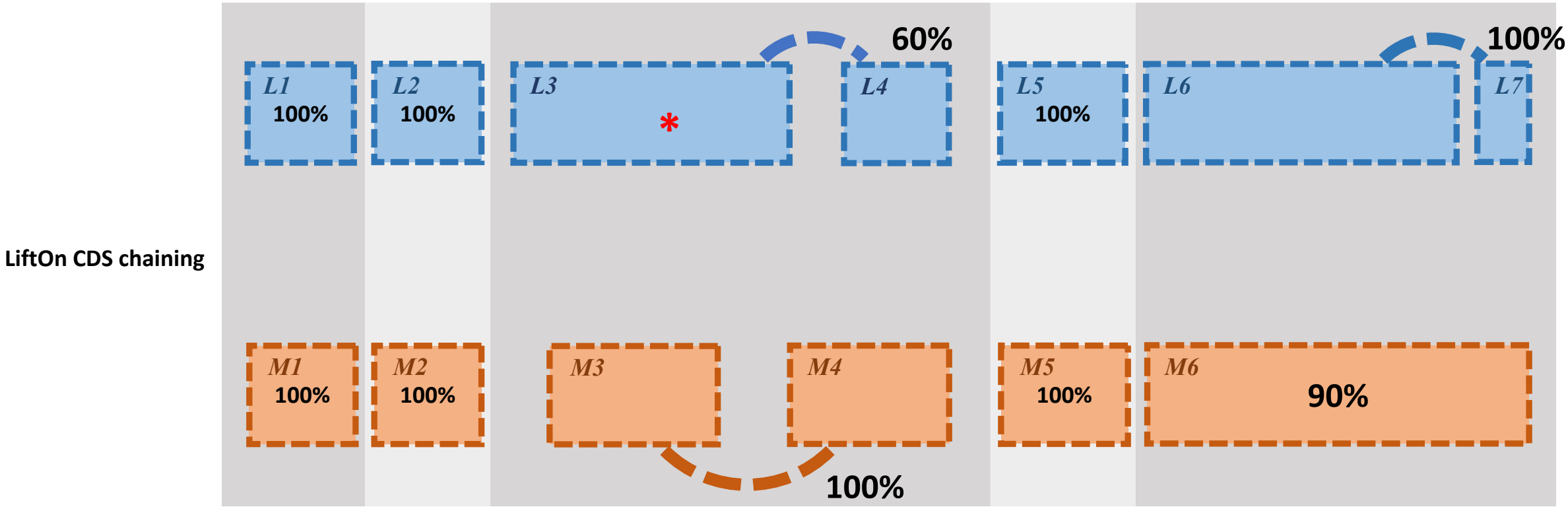
LiftOn: Protein-maximization algorithm

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



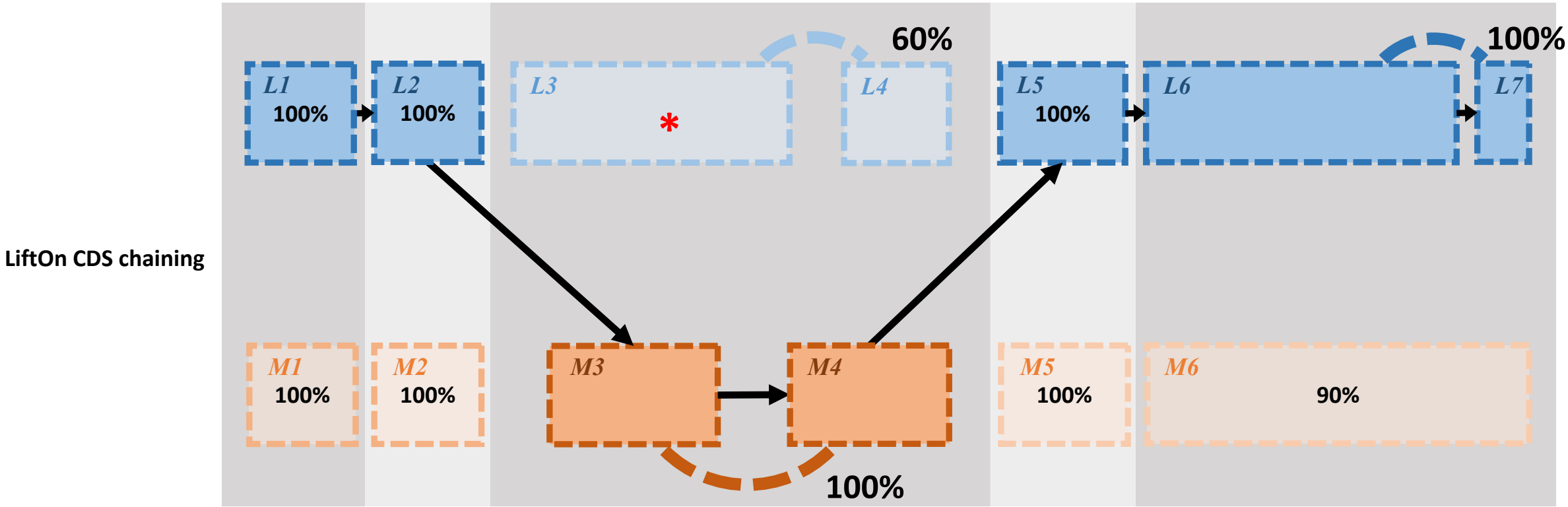
LiftOn: Protein-maximization algorithm

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

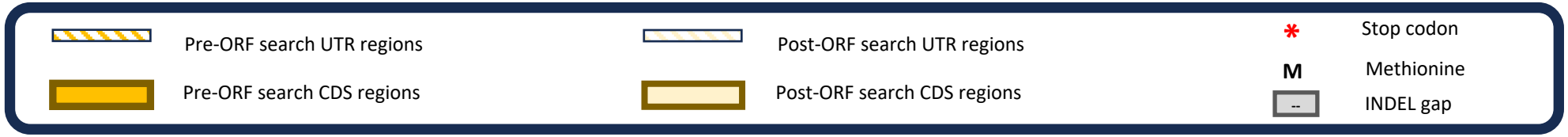


LiftOn: Protein-maximization algorithm

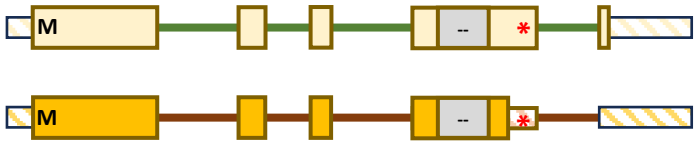
D **Step 3: group CDSs by "accumulated AA in the reference protein"**



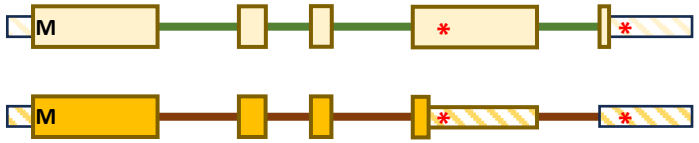
LiftOn: Protein-maximization algorithm



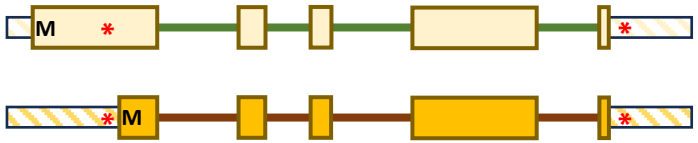
A Frameshift



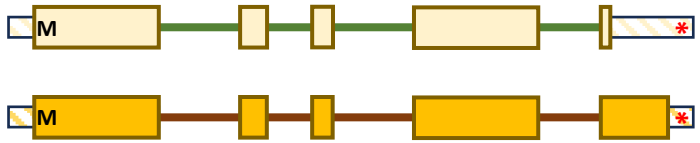
B Stop codon gain: Early translation stop



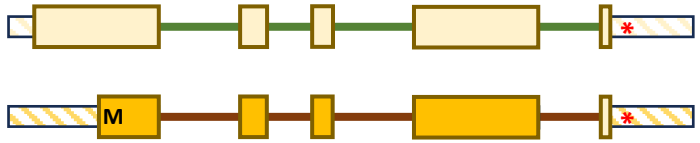
C Stop codon gain: Switching translation start



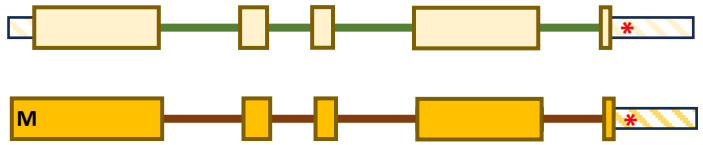
D Stop codon lost: Protein extension



E Start codon lost: Downstream start



F Start codon lost: Upstream start



LiftOn : Summary

- LiftOn is a promising new tool to study comparative genomics
- 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.

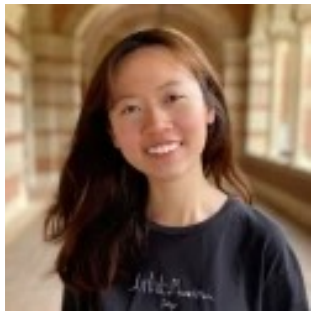
Acknowledgement

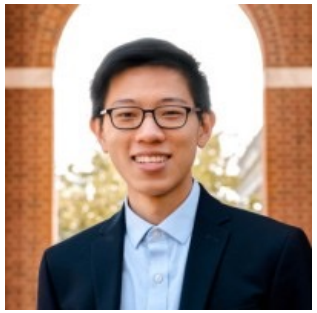

Steven Salzberg


Mihaela Pertea

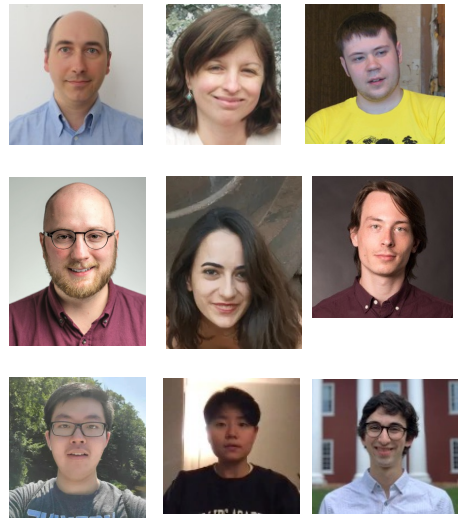

Alaina Shumate


Jakob Heinz


Celine Hoh


Alan Mao

Salzberg & Pertea Lab



LiftOn

Google search:
“LiftOn genome”

Chao, K. H., Heinz, J. M., Hoh, C., Mao, A., Shumate, A., Pertea, M., & Salzberg, S. L. (2024). Combining DNA and protein alignments to improve genome annotation with LiftOn. **bioRxiv**.

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

ccb.jhu.edu/lifton

GPL-3.0 license

46 stars 1 fork 1 watching

1 Branch 5 Tags Activity

ccb.jhu.edu/lifton

github.com/Kuanhao-Chao/LiftOn