Combining DNA and protein alignments to improve genome annotation with LiftOn

Johns Hopkins Whiting School of Engineering
Department of Computer Science

Johns Hopkins University
Center for Computational Biology
CCB

Kuan-Hao Chao[1,2,*], Jakob M. Heinz[5], Celine Hoh[1,2], Alan Mao[1,2,3], Alaina Shumate[2,3], Steven Salzberg[1,2,3,4,*], Mihaela Pertea[1,2,3,*]

[1]Department of Computer Science, [2]Center for Computational Biology, [3]Department of Biomedical Engineering, [4]Department of Biostatistics, Johns Hopkins University , [4]Department of Biomedical Informatics, Harvard University

kh.chao@cs.jhu.edu

https://ccb.jhu.edu/lifton/

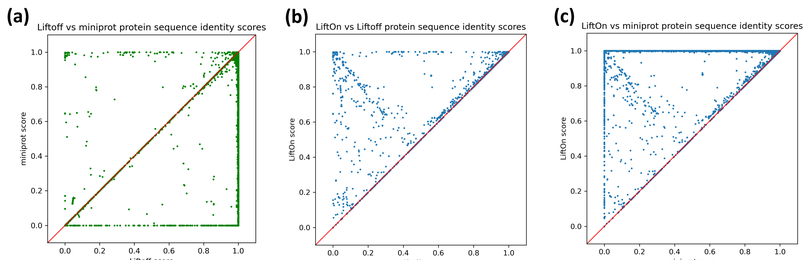
https://github.com/Kuanhao-Chao/LiftOn

https://khchao.com

Introduction

- 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 species, at least as divergent as mouse and rat.

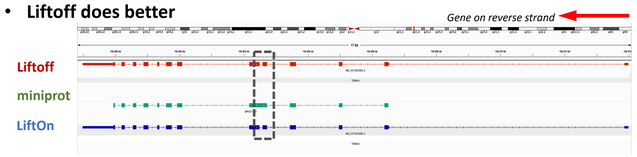
LifOn vs Liftoff vs miniprot



The scatter plot of protein sequence identity comparing between **(a)** miniprot vs Liftoff, **(b)** LiftOn vs Liftoff, and **(c)** LiftOn vs miniprot. Each dot represent a protein-coding transcript.

IGV screenshots of gene loci examples

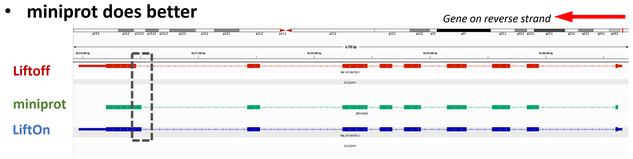
- **Liftoff does better**



TDRKH (NM_001083965.2)
chr1:150896981 - 150913985

	Protein identity
Liftoff	100%
miniprot	31.48%
LiftOn	100%

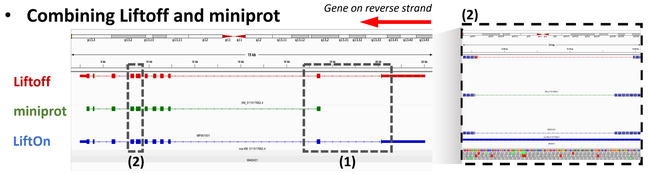
- **miniprot does better**



SLC22A31 (NM_001384763.1)
chr16:95276205 - 95280662

	Protein identity
Liftoff	76.70%
miniprot	99.55%
LiftOn	99.55%

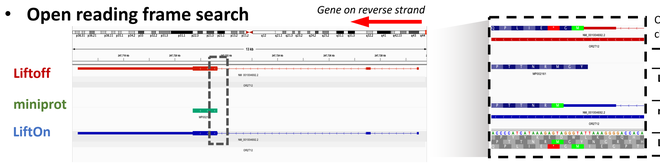
- **Combining Liftoff and miniprot**



WASHC1 (XM_011517662.4)
chr19:6990 - 22049

	Protein identity
Liftoff	38.92%
miniprot	99.14%
LiftOn	99.35%

- **Open reading frame search**

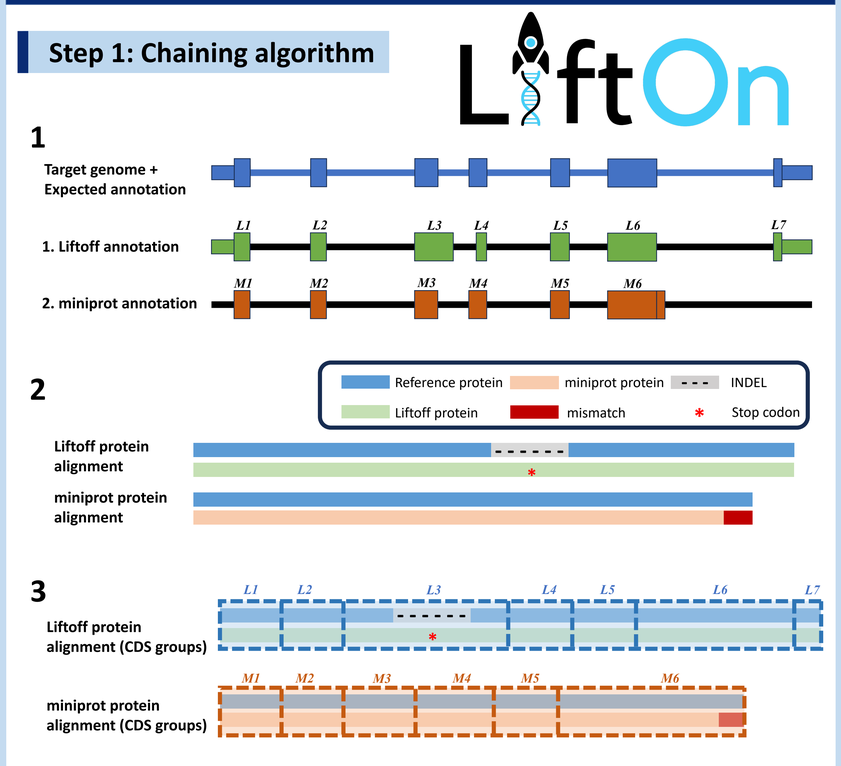


OR2T12 (NM_001004692.2)
chr1:247716221 - 247729507

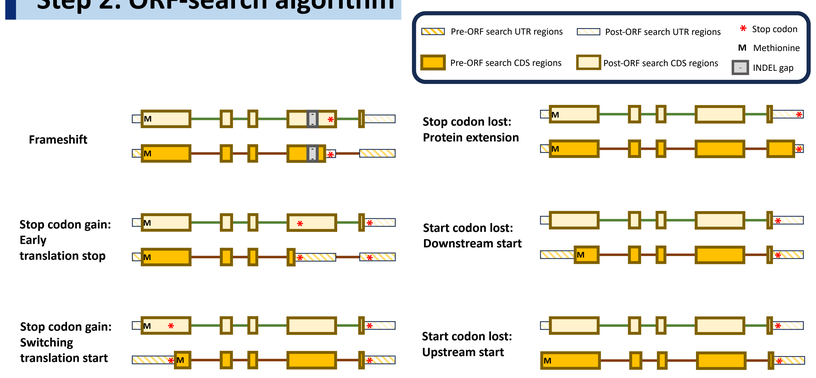
	Protein identity
Liftoff	00.27%
miniprot	97.51%
LiftOn	97.51%

Protein-maximization algorithm

Step 1: Chaining algorithm



Step 2: ORF-search algorithm



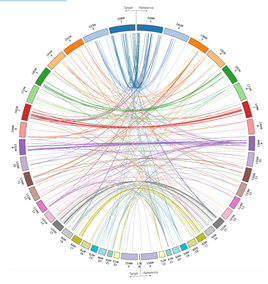
Results

Mapping RefSeq v220 from GRCh38 to CHM13

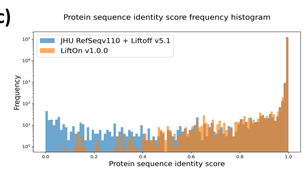
(a)

		Total feature count	Protein-coding feature count			Non-coding feature count		
Gene	Reference (GRCh38)	37,986			19,927			18,059
	Target (CHM13)	38,886	*Single copy*	*>1 copy*	*Extra copy*	*Single copy*	*>1 copy*	*Extra copy*
			19,738	85	291	17,715	289	768
					20,114			18,772
Transcript	Reference (GRCh38)	160,561			130,528			30,033
	Target (CHM13)	16,1671	*Single copy*	*>1 copy*	*Extra copy*	*Single copy*	*>1 copy*	*Extra copy*
			129,967	238	544	29,488	410	1,024
					130,749			30,922

(b)



(c)

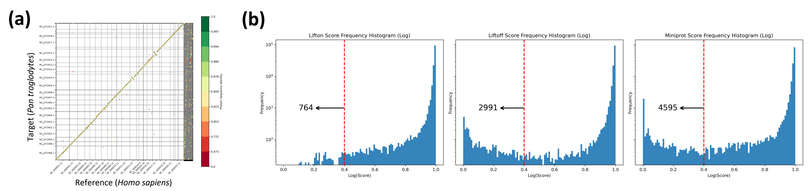


(d)

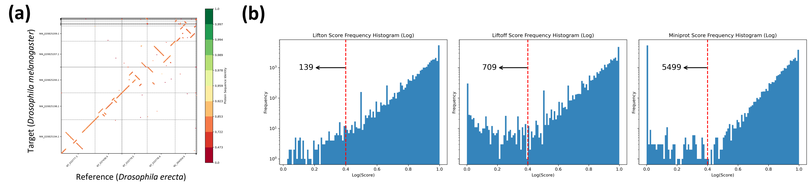
Frameshift	Start loss	Stop missing	Stop gain	Inframe insertion	Inframe deletion	Nonsynony -mous
126	5	106	781	2	21	5

(a) LiftOn GRCh38-to-CHM13 mapping summary of protein-coding and non-coding genes and transcripts.
(b) Circos plot shows relative positions of extra gene copies between target (left) and reference (right)
(c) Protein-coding sequence identity frequency plots comparing LiftOn and JHU RefSeqv110+Liftoff v5.1
(d) Mutation reports of LiftOn-generated human annotation.

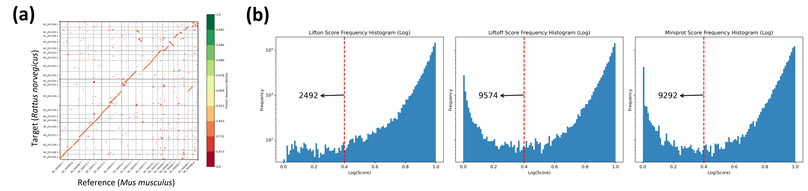
Mapping from *Homo sapiens* to *Pan troglodytes*



Mapping from *Drosophila melanogaster* to *Drosophila erecta*



Mapping from *Mus musculus* to *Rattus norvegicus*



(a) Protein-coding gene order plot with x-axis for reference genome and y-axis for target genome, using a logarithmic color scale from green (identical) to red for protein sequence identities.
(b) Logarithmic frequency plots of protein sequence identity for LiftOn, Liftoff, and miniprot.

Acknowledgements: supported by NIH under grants R01-HG006677, R35-GM130151, and T32HG002295 and by NSF under grants DBI-1759518.