

Computational Life Science Seminar

Stefan Schmutz

2020-11-16

ARTICLES

<https://doi.org/10.1038/s41588-019-0483-y>

nature
genetics

Corrected: Publisher Correction

Inferring whole-genome histories in large population datasets

Jerome Kelleher^{ID}*, Yan Wong, Anthony W. Wohns^{ID}, Chaimaa Fadil^{ID}, Patrick K. Albers^{ID}
and Gil McVean^{ID}

What does this **Title** reveal?

Inferring whole-genome histories in large population datasets

SYNONYMS FOR *inferring*

ascertain

assume

construe

deduce

derive

interpret

presume

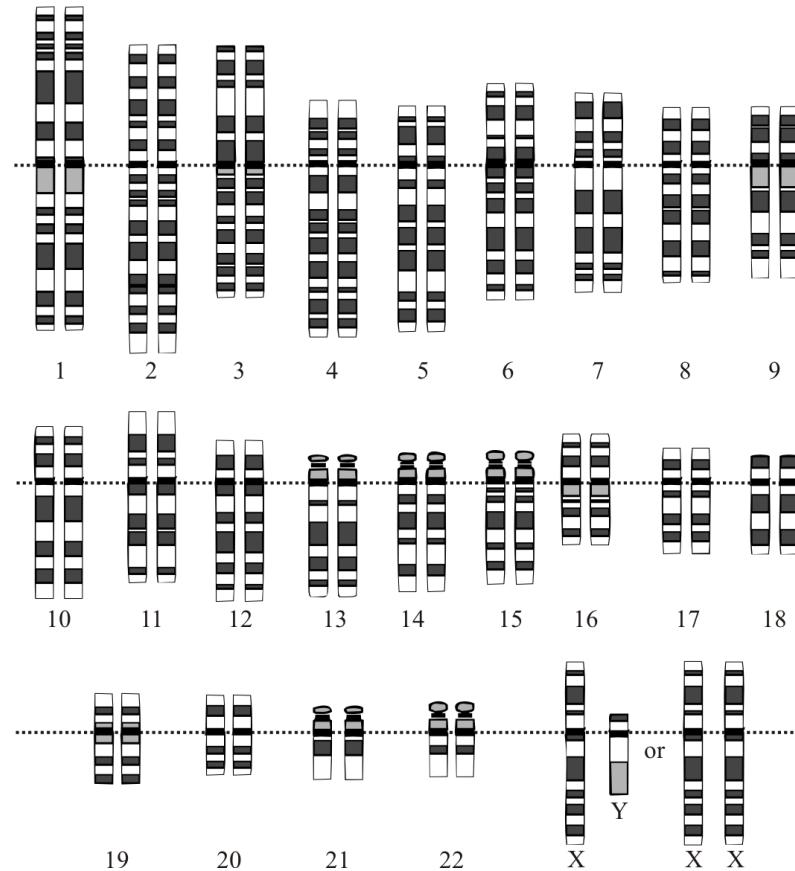
presuppose

reckon

speculate

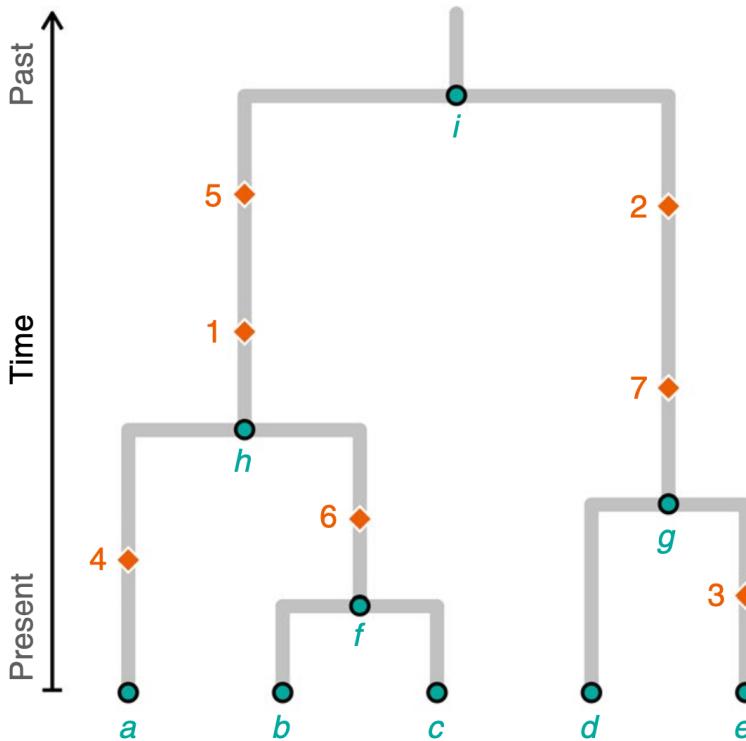
source: thesaurus.com

Inferring whole-genome histories in large population datasets



source: National Human Genome Research Institute

Inferring whole-genome histories in large population datasets



source: Kelleher et al., 2019

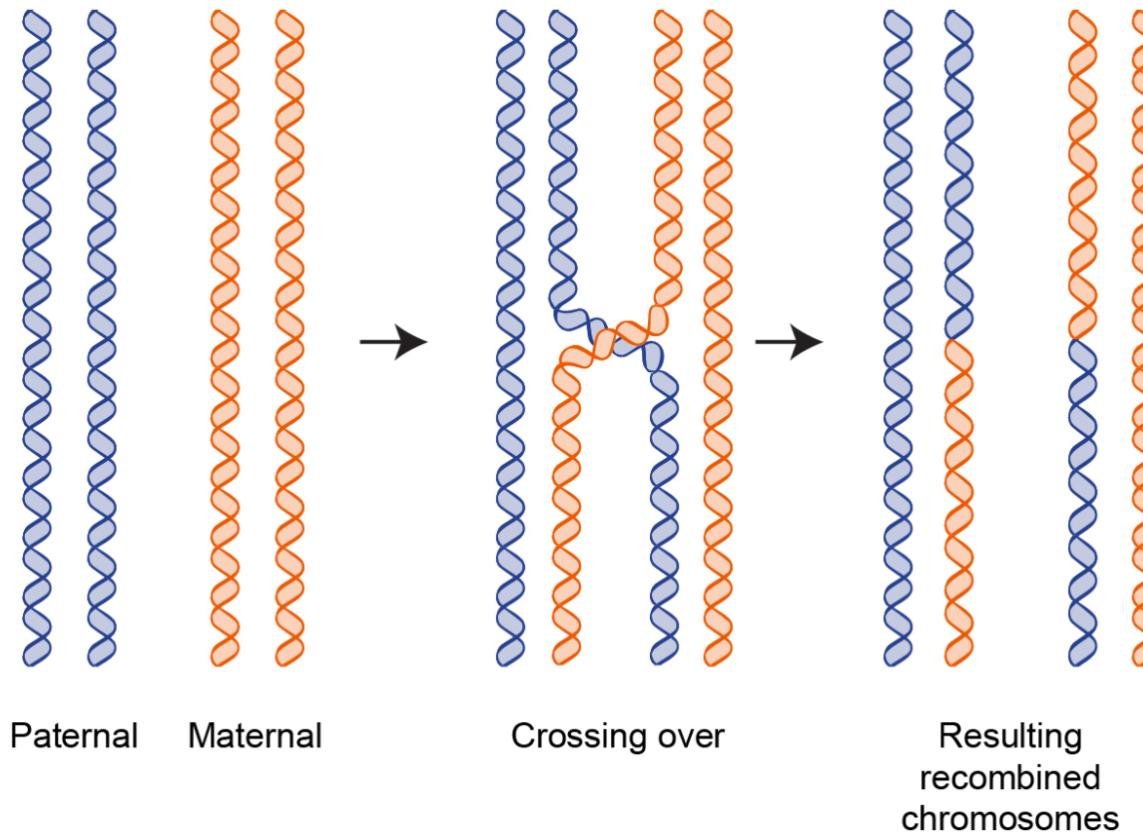
Inferring whole-genome histories in large population datasets

	Position	1	2	3	4	5	6	7	8	9	10
Sample haplotype	P1	A	T	A	A	C	G	G	G	C	A
	P2	T	T	G	G	C	G	G	G	C	A
	M1	T	T	A	A	C	G	G	G	C	A
	M2	T	T	A	A	C	G	G	C	C	A
	C1	A	T	A	G	C	G	G	G	C	A
	C2	T	T	G	G	C	T	G	C	C	A

Paternal
Maternal
Child

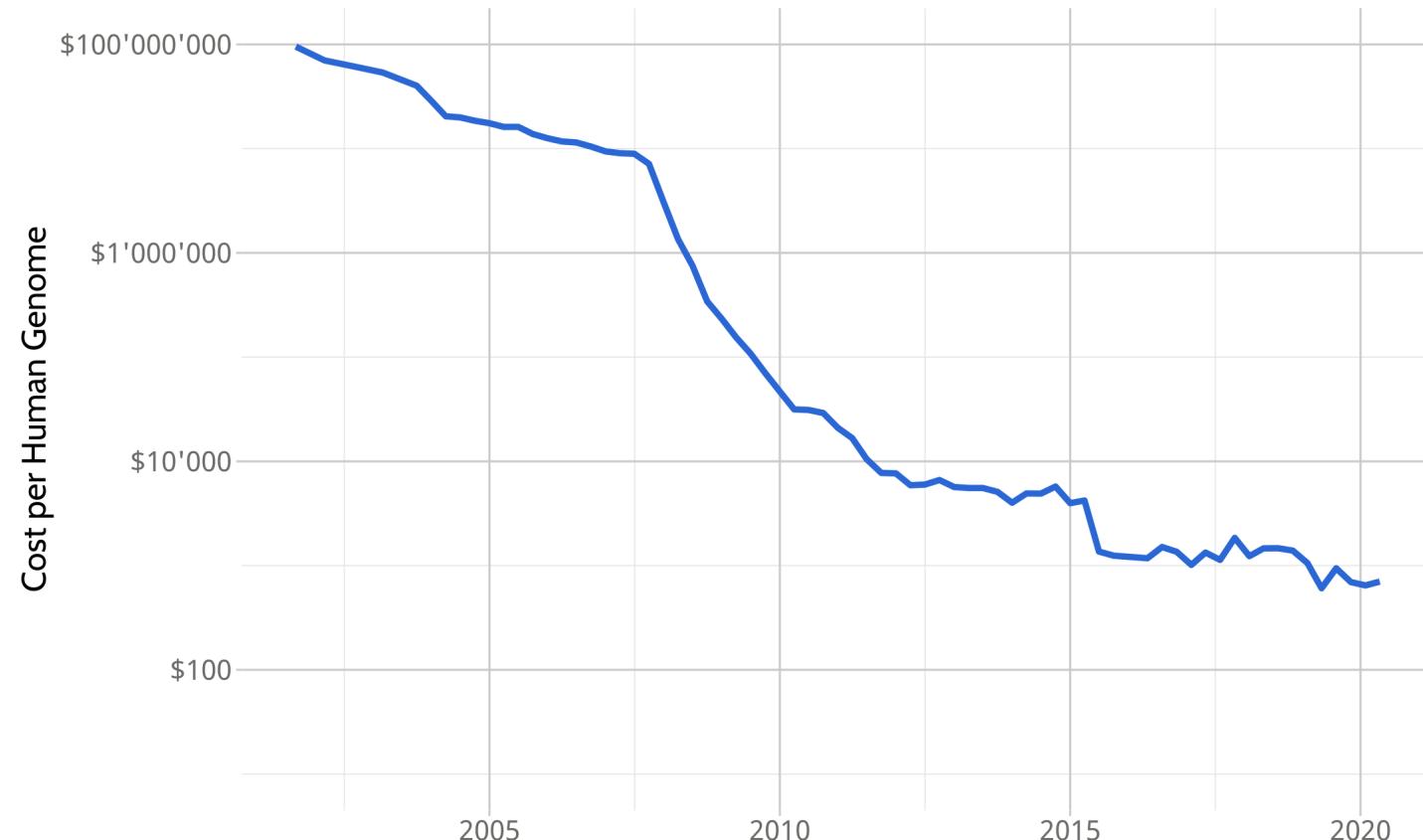
{A, T, G, C} ancestral
{A, T, G, C} derived

Inferring whole-genome histories in large population datasets



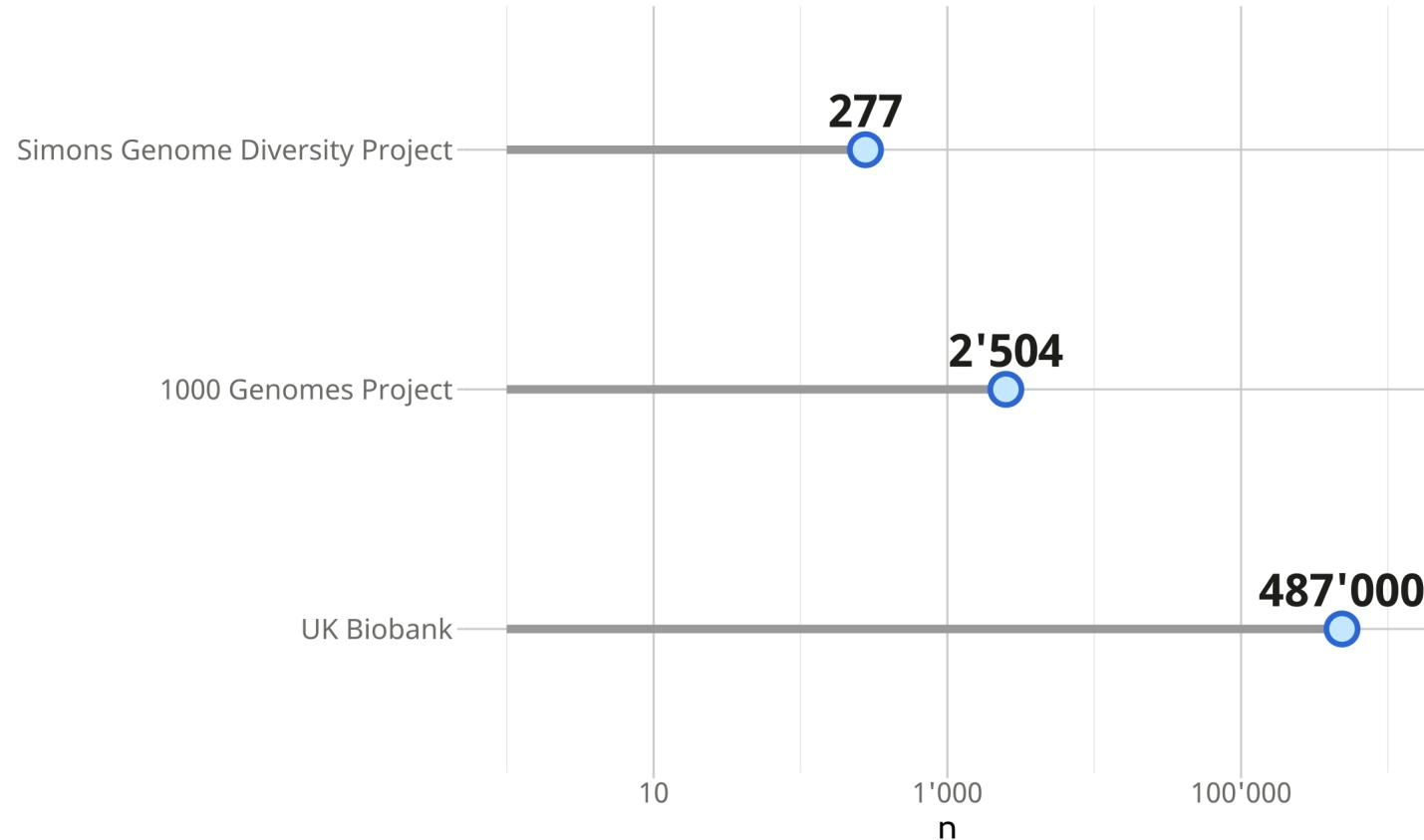
source: genome.gov

Inferring whole-genome histories in large population datasets



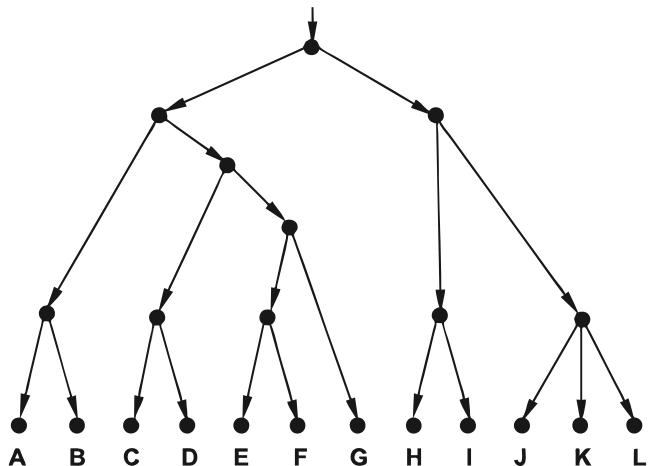
source: genome.gov/sequencingcosts

Inferring whole-genome histories in large population datasets



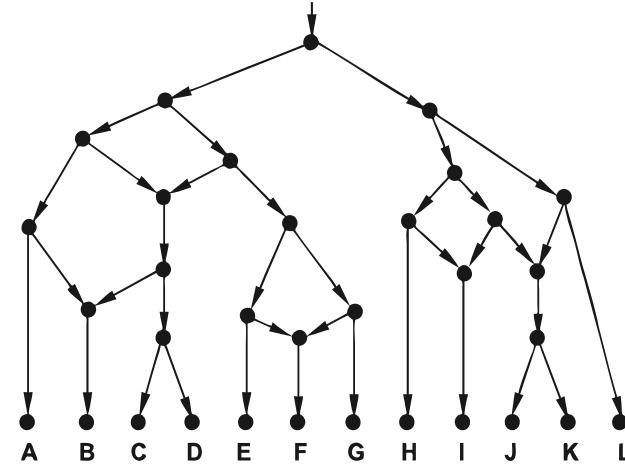
Current situation

Tree



source: Adapted from Morrison, 2016

Network



source: Adapted from Morrison, 2016

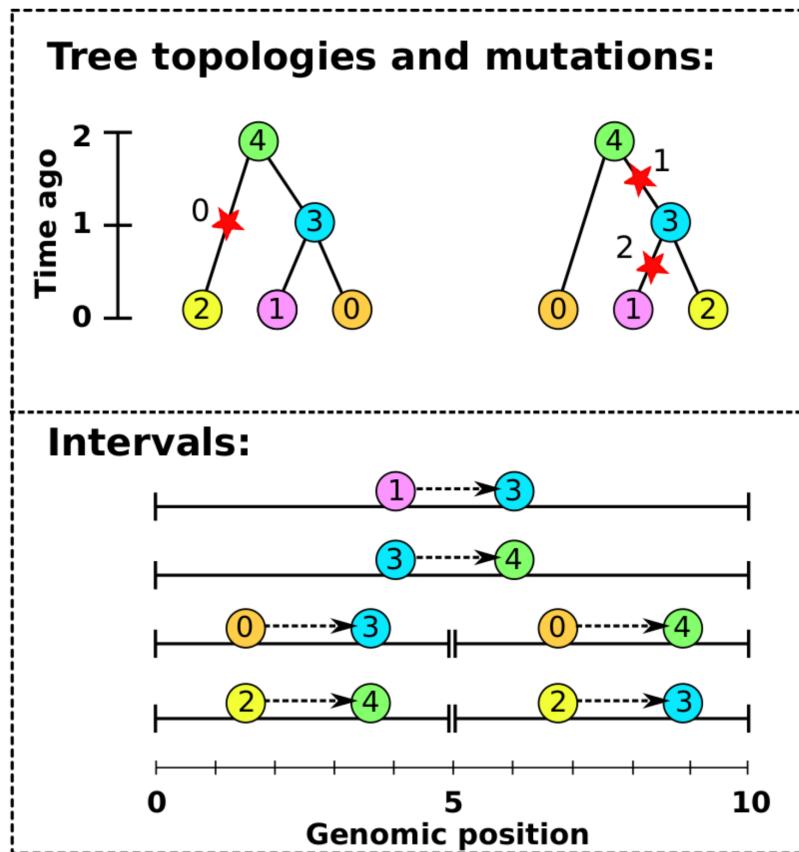
What's the Author's proposed Solution?



source: github.com/tskit-dev

tsinfer

tree sequence



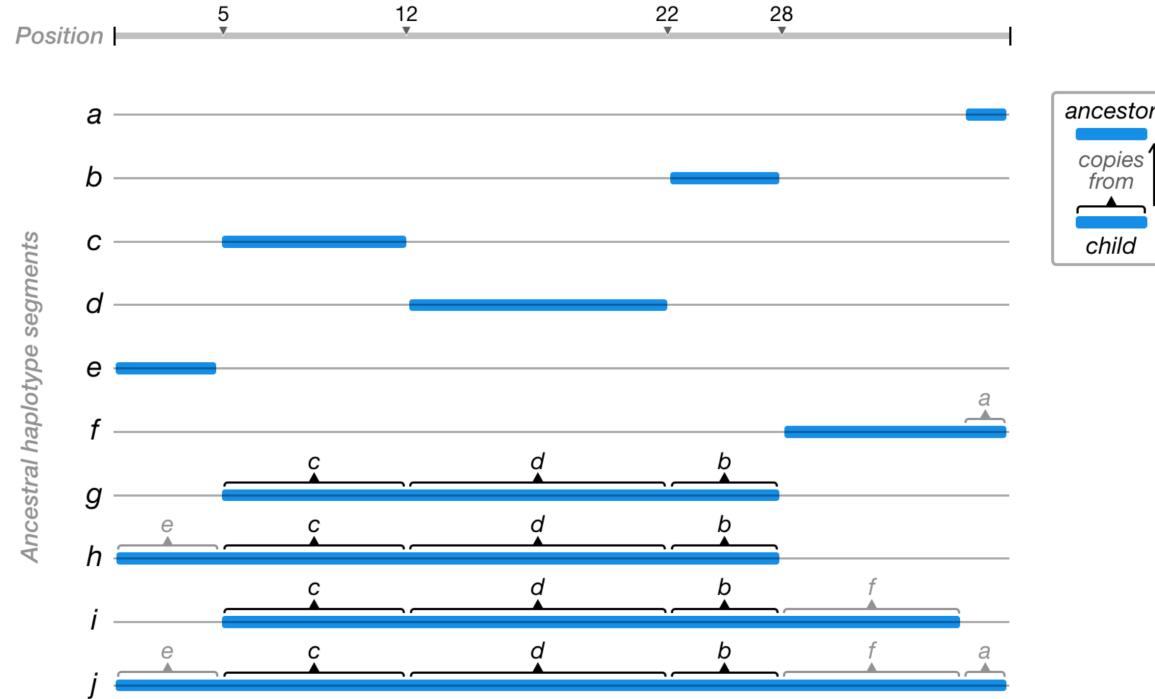
source: Kelleher et al., 2018

Nodes:		Edges:			
ID	time	left	right	parent	child
0	0.0	0	10	3	1
1	0.0	0	10	4	3
2	0.0	0	5	3	0
3	1.0	0	5	4	2
4	2.0	5	10	3	2
		5	10	4	0

Sites:			Mutations:	
ID	position	ancestral state	ID	site
0	2.5	A	0	0
1	7.5	G	1	1
			2	1

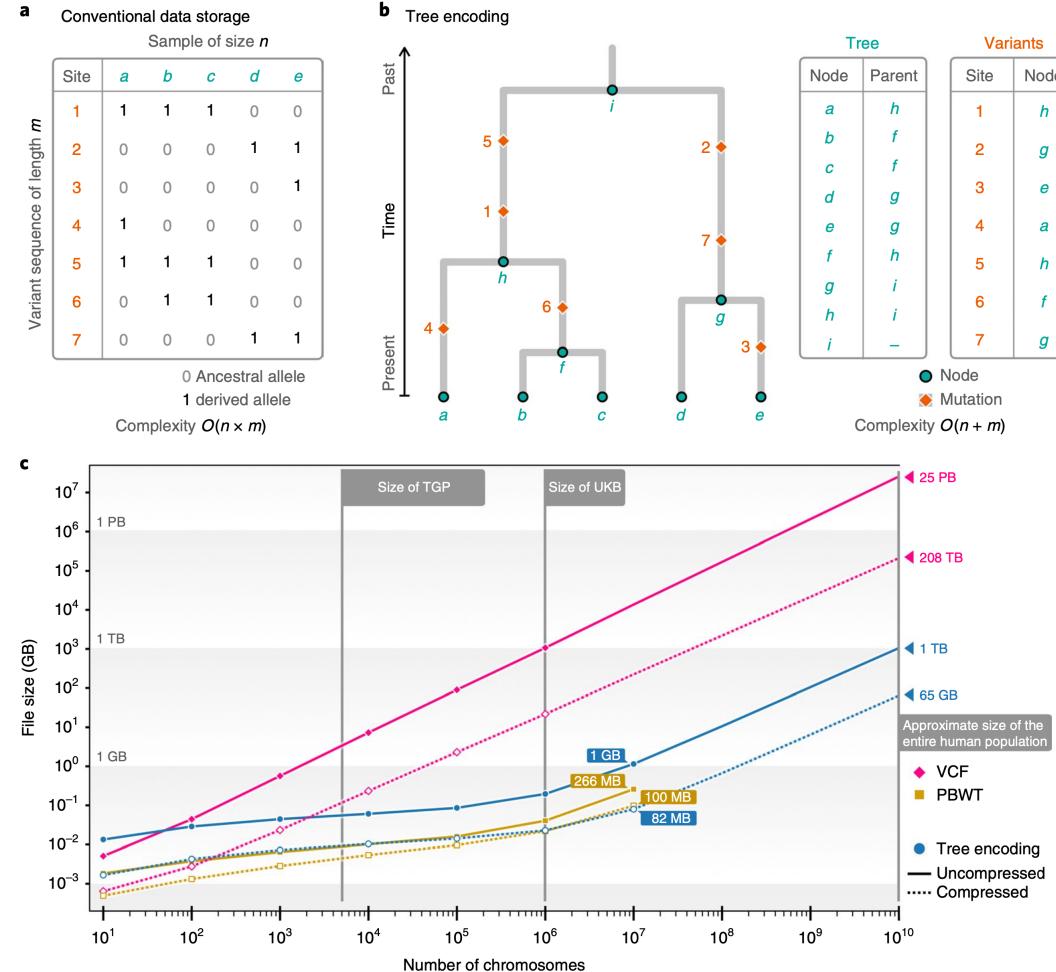
source: Kelleher et al., 2018

Ancestral haplotype inference



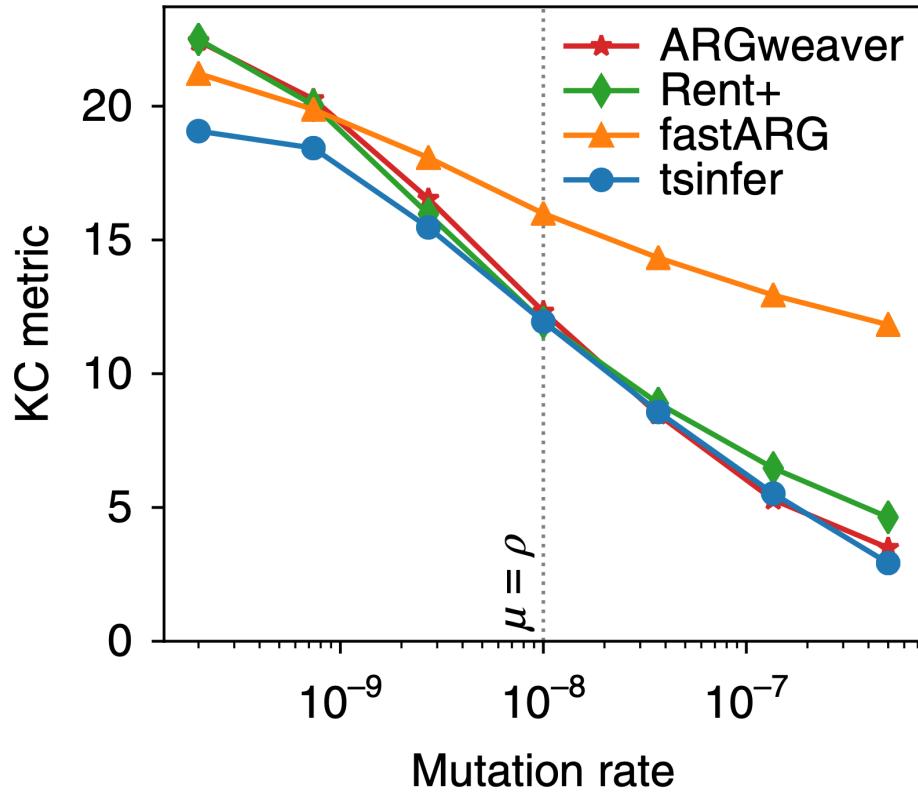
source: Kelleher et al., 2019

Comparison to state-of-the-art Storage Space



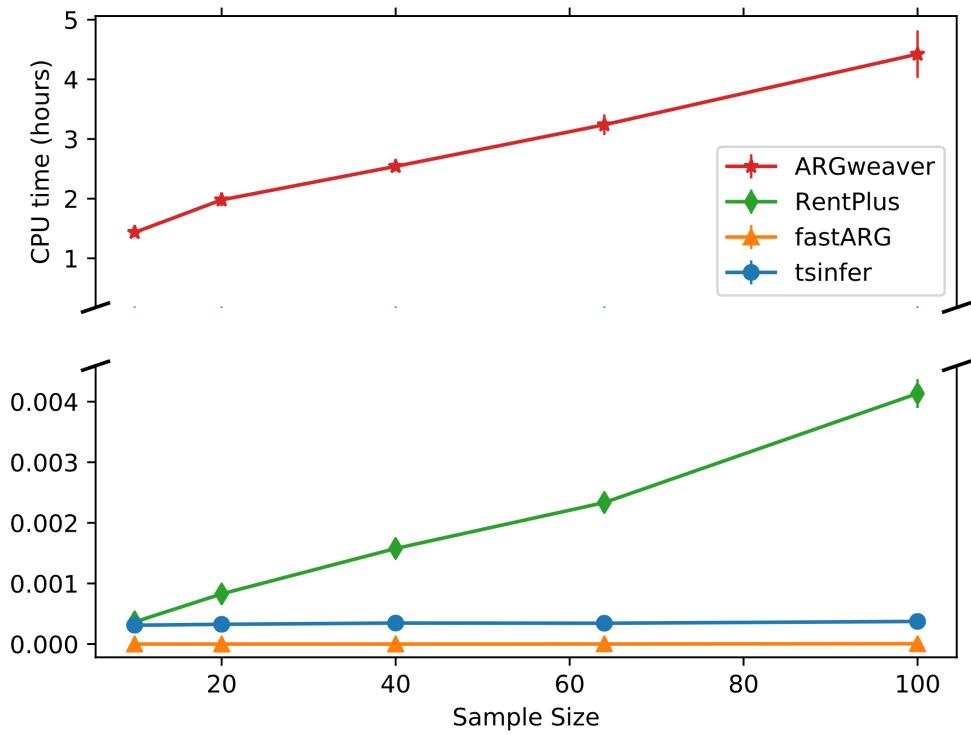
source: Kelleher et al., 2019

Comparison to state-of-the-art



source: Kelleher et al., 2019

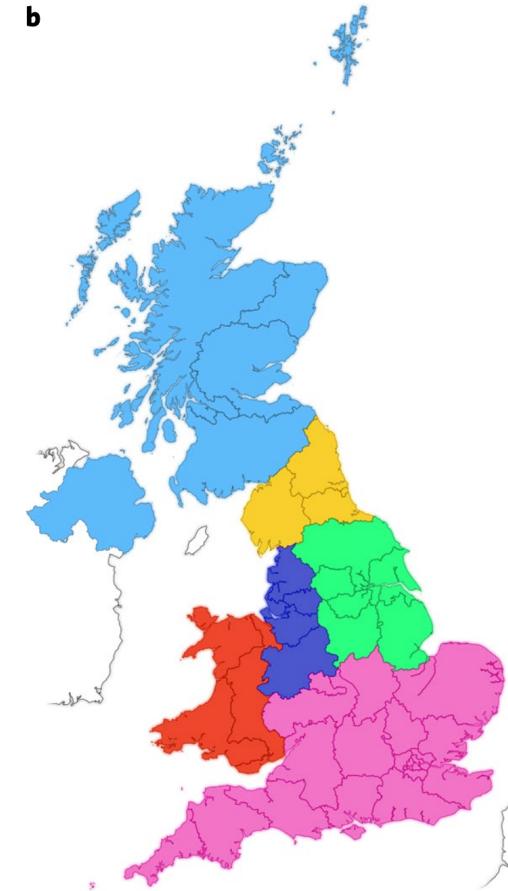
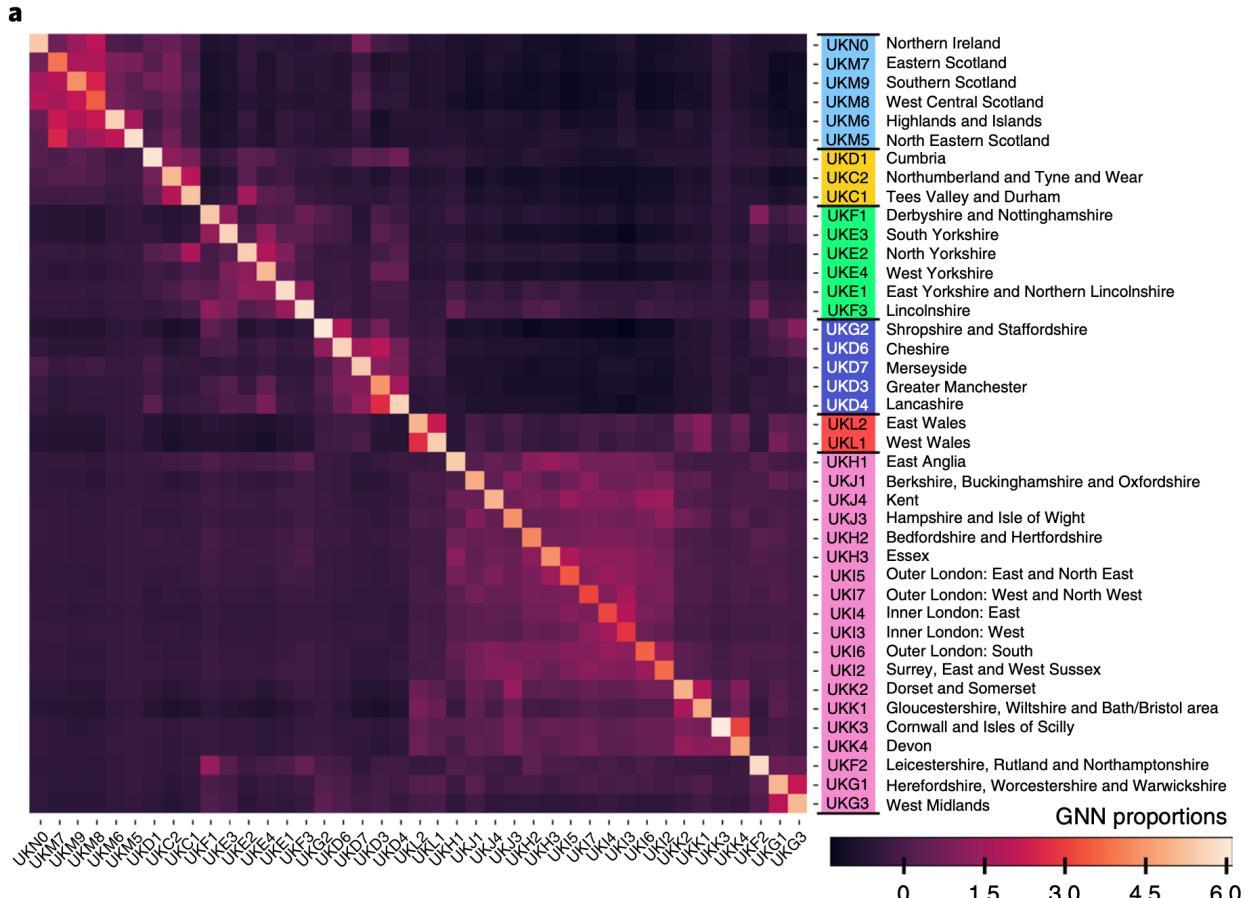
Accuracy and Speed



source: Kelleher et al., 2019

Application example

UK Biobank population structure



source: Kelleher et al., 2019

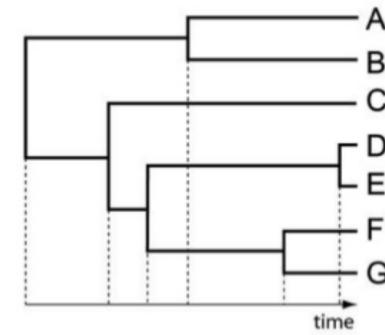
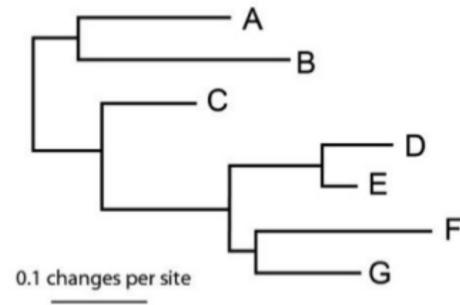
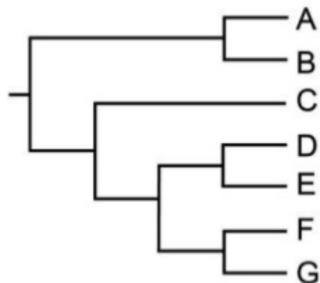
Drawbacks

Assumptions

- each variant has a single origin
- mutation/recombination ratio is sufficiently high
- haplotype frequency as proxy for relative age

Drawbacks

Cladogram vs. Phylogram vs. Chronogram



source: Riutort, 2016

Outlook

From topologies to branch lengths (tsdate)

Improved sequencing technologies

Possible application for genomes of other Species?

Genome Watch | Published: 21 September 2020

Recombination should not be an afterthought

Russell Y. Neches , Matthew D. McGee & Nikos C. Kyrpides

Nature Reviews Microbiology 18, 606(2020) | [Cite this article](#)

1073 Accesses | 17 Altmetric | [Metrics](#)

This month's Genome Watch highlights how the search for the origins of SARS-CoV-2 emphasizes the need for integrated phylogenetic methods.

RESEARCH ARTICLE

Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses

 Nicola F. Müller,  Ugné Stolz,  Gytis Dudas, Tanja Stadler, and Timothy G. Vaughan

PNAS July 21, 2020 117 (29) 17104–17111; first published July 6, 2020; <https://doi.org/10.1073/pnas.1918304117>

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved May 11, 2020 (received for review October 22, 2019)



Inferring whole-genome histories in large population datasets

Appendix

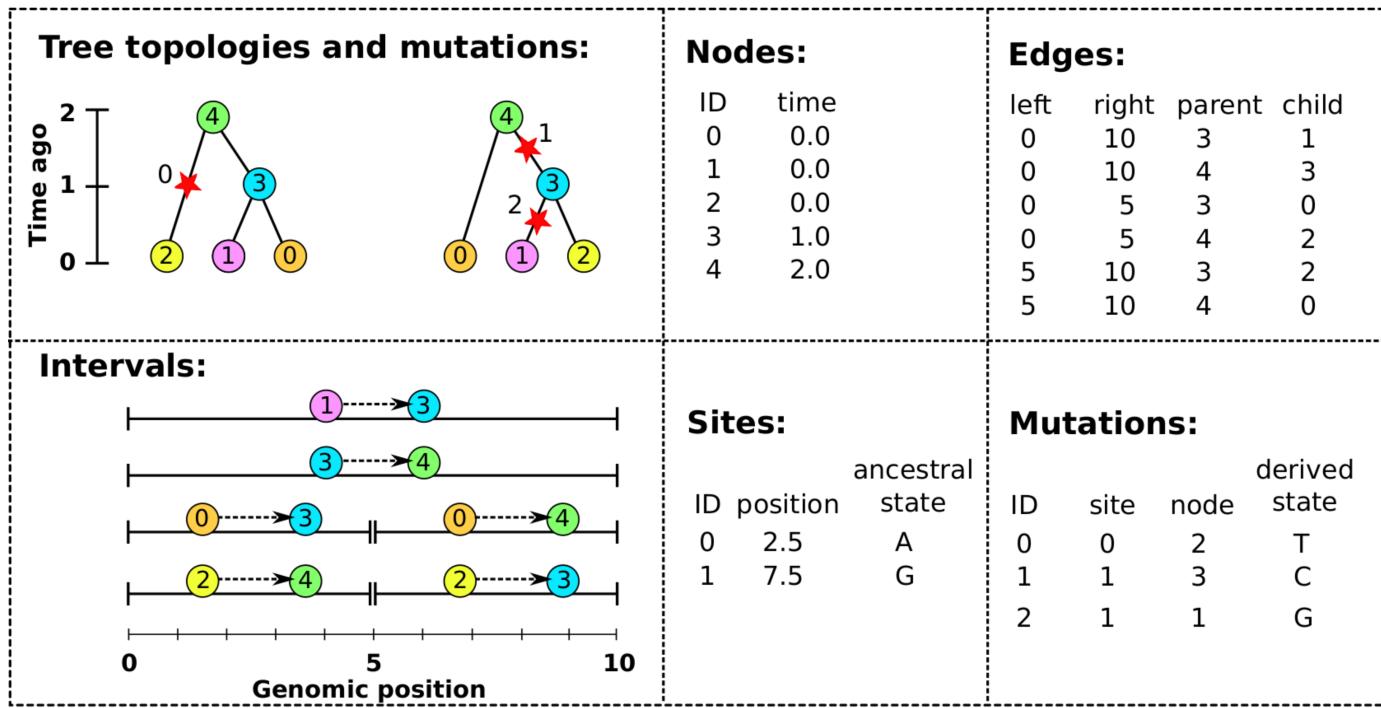


Fig 3. An example tree sequence with three samples over a chromosome of length 10. The leftmost panels show the tree sequence pictorially in two different ways: (top) a sequence of tree topologies; the first tree extends from genomic position 0 to 5, and the second from 5 to 10; and (bottom) the edges that define these topologies, displayed over their corresponding genomic segment (for instance, the edge from node 2 to node 4 is present only on the interval from 0 to 5). The remaining panels show the specific encoding of this tree sequence in the four tables (nodes, edges, sites and mutations).

source: Kelleher et al., 2018

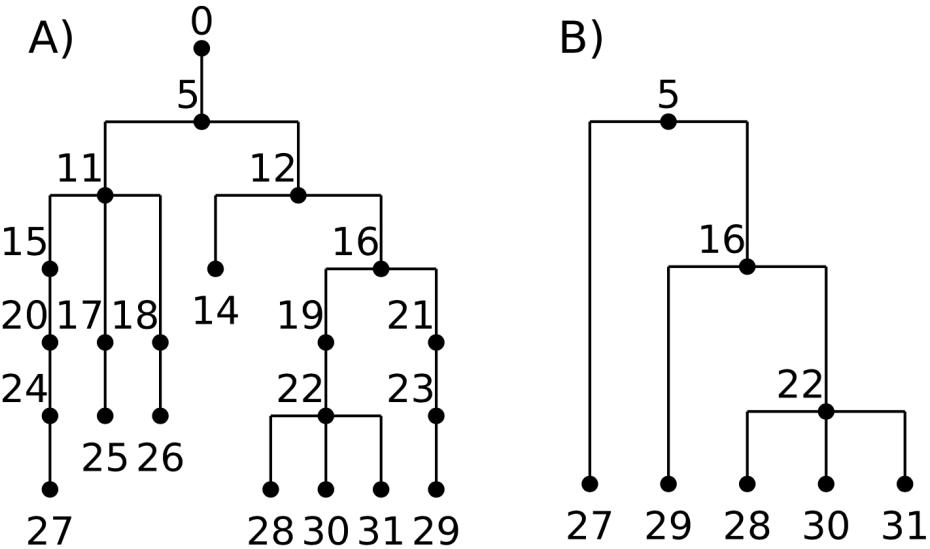


Fig 4. An example of a marginal genealogy from a Wright-Fisher simulation with $N = 5$. (A) the original tree including all intermediate nodes and dead-ends, and (B) the minimal tree relating all of the currently-alive individuals (27–31).

source: Kelleher et al., 2018