

# Inferring whole-genome histories in large population datasets

## Kelleher et al. (2019)

### Terms and Abbreviations

Genome .....	contains blueprint (genetic information) of an organism
DNA .....	molecule composed of four building blocks/nucleotides (adenine [A], thymine [T], cytosine [C] and guanine [G])
DNA sequencing .....	process of determining the order (sequence) of nucleotides of a DNA molecule
diploid .....	organism containing two sets of each chromosome
genealogy .....	study of history of organisms (how did they evolve, how are they related)
phylogenetic tree .....	method to visualise evolutionary relationships (see Figure 1)
topology .....	arrangement/shape of a phylogenetic tree
node .....	point/individual within a tree (see Figure 1)
edge .....	lines connecting nodes of a tree (see Figure 1)
DNA mutation .....	change in DNA sequence when compared to parent sequence
DNA recombination ..	when a DNA molecule gets parts of two different sources
ancestral sequence .....	original (past) DNA sequence of an ancestor
derived sequence .....	DNA sequence where changes happened over time in comparison to the ancestral sequence
tree sequence .....	set of multiple phylogenetic trees considering that different parts of a genome can have different histories
VCF .....	Variant Call Format, common format used in bioinformatics to store differences (variants) in DNA sequences
KC metric .....	Kendall-Colijn metric, compares topologies of trees (how close they are)
GNN .....	Genealogical Nearest Neighbors, composition of neighbors on the tree(s)
phasing .....	separate paternal and maternal DNA sequence when reconstructing a diploid genome
sequence alignment ...	arranging multiple sequences in a way that same regions are grouped
haplotype .....	part of the genome which was inherited from a single parent

### Context

DNA sequencing of genomes allows us to reconstruct the evolution of organisms. To do this, we need to focus on DNA mutations which occurred over time.

Histories of DNA sequences can be depicted by a phylogenetic tree using multiple established methods.

Instead of a tree, DNA evolution is often better represented as a network, since parts of the genomes can be exchanged by recombination events. There are models for such networks, they are however very computationally expensive and don't scale well with large amounts of data.

With the recent advances in high throughput sequencing techniques there's a need for a more efficient approach.

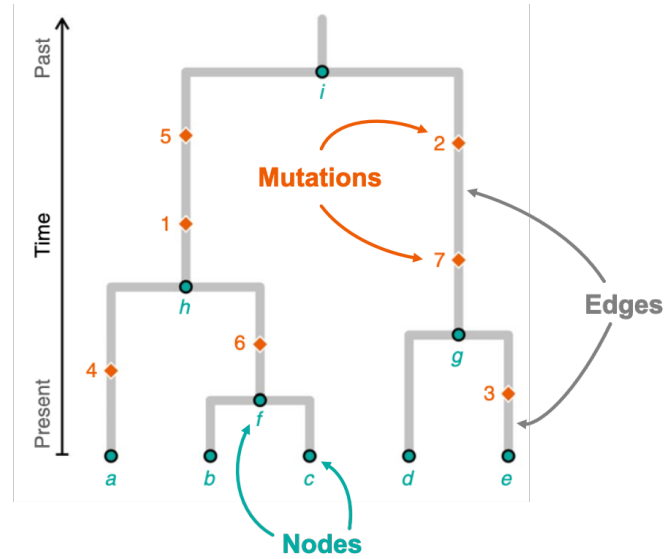


Figure 1: Phylogenetic tree, adapted from Kelleher et al. (2019)

## Main Points

Kelleher and colleagues present in this article the applicability of an efficient data structure, tree sequence, to infer genome histories without the restrictions of excluding recombination events.

In addition to being accurate, the method presented is also very fast and scalable compared to available state-of-the-art tools.

They demonstrate the functionality by characterizing data from three medium to large human genome datasets.

## Reference

Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., McVean, G., 2019. Inferring whole-genome histories in large population datasets. *Nature Genetics* 51, 1330–1338. <https://doi.org/10.1038/s41588-019-0483-y>