UNIVERSITY OF COPENHAGEN
SECTION OF FORENSIC GENETICS
DEPARTMENT OF FORENSIC MEDICINE

SNPtotree

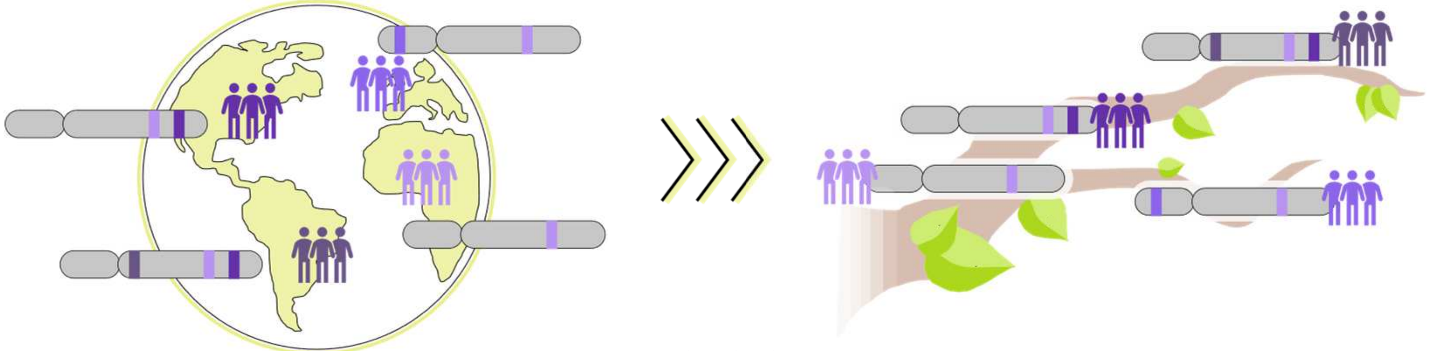
Resolving the phylogeny of variants on non-recombining DNA

Zehra Köksal[1], Claus Børsting[1], Leonor Gusmão[2], Vania Pereira[1]

[1]Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark; [2]DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil;

CONCLUSIONS

SNPtotree is the only available tool to automatically generate a phylogenetic tree representing the hierarchy of biallelic variants on non-recombining DNA. It also assigns the samples carrying the variants to the respective branches and gives statistical support values for each branch. SNPtotree was validated using human Y-chromosomal variants, supporting known phylogenies and establishing previously unknown phylogenies even for sequencing data with high fractions of missing data.

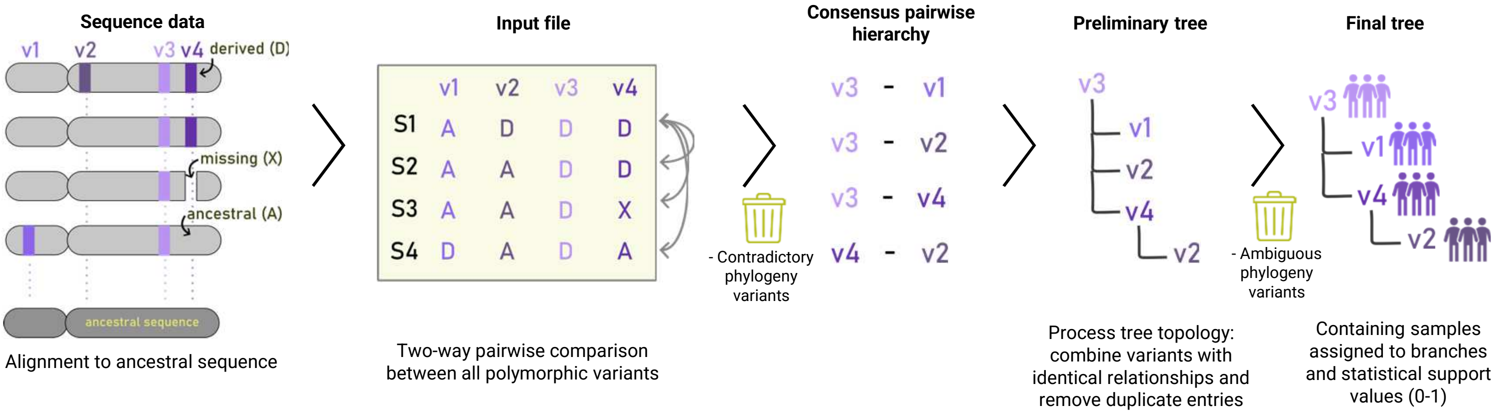


INTRODUCTION

Biallelic variants on non-recombining DNA (e.g. male-specific Y chromosome) are passed on over generations in a hierarchical order. This hierarchy can be reconstructed in phylogenetic trees, which have been utilized for decades in epidemiologic, evolutionary, forensic and population studies. These include determination of the genetic genealogy of individuals and investigation of the genetic evolution within species [1-3]. Previously, no tool existed to automatically establish this hierarchy. Traditional tree construction methods, such as maximum likelihood (ML) approaches, were utilized to generate phylogenetic trees with DNA sequences at the tree tips. With extensive manual work, variants exclusive to samples of certain clades could be identified to generate a phylogenetic tree of variants. Further, sequencing data with high fractions of missing information resulting from low quality samples or combinations of different datasets may result in erroneous or/and incomplete trees, where the relationships between the variants are not fully assessed or correctly represented [4-8]. Here, we present the publicly available software SNPtotree v1.0 [9] that may be used to organize biallelic variants from datasets with missing and low-quality data into phylogenetic trees without manual sorting. SNPtotree is applicable for generating or updating phylogenetic databases.

METHODS

SNPtotree is a command line tool written in Python (https://github.com/ZehraKoksal/SNPtotree) that uses the fundamental character-by-taxon data matrix of the allelic states of biallelic variants. It filters for phylogenetically informative variants and sorts them into a conservative phylogenetic tree, alongside the samples carrying the variants and a statistical support value per variant (0-1). SNPtotree was benchmarked using a combined dataset of the human Y-chromosomal haplogroup C in 46 samples with 4,349 variants and missing data ranging from 20 to 65% per sample [10,11]. For comparison, ML-based trees were constructed using the state-of-the-art software RAxML [12]. Only variants with reported phylogeny in the Y-SNP database ISOGG Y-DNA Haplogroup Tree 2019–2020 were included in the phylogenetic trees.



	v1	v2	v3	v4
S1	A	D	D	D
S2	A	A	D	D
S3	A	A	D	X
S4	D	A	D	A

Alignment to ancestral sequence

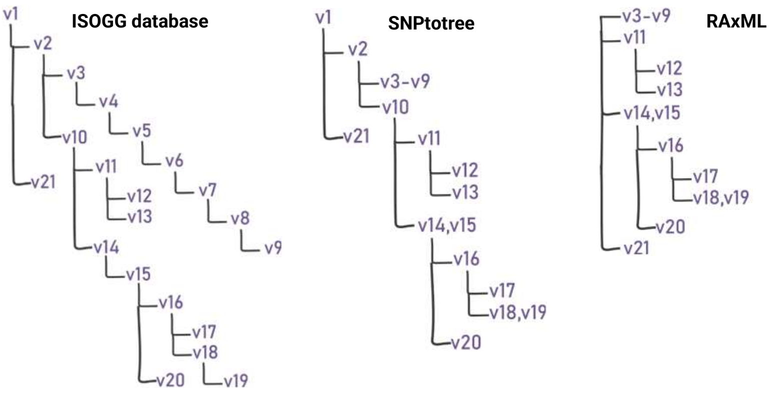
Two-way pairwise comparison between all polymorphic variants

Process tree topology: combine variants with identical relationships and remove duplicate entries

Containing samples assigned to branches and statistical support values (0-1)

RESULTS AND DISCUSSION

SNPtotree could sort 4,071 of the 4,349 variants into 81 branches. One subclade (C1b1) was selected for comparison between the different tools. Due to the thorough pairwise comparison of variants, the runtime of SNPtotree was 19 hours for 4,349 variants, while the RAxML runtime was 15 minutes. The phylogeny determined using SNPtotree was similar to the expected phylogeny from the database. The phylogeny obtained using RAxML was only comparable to that of SNPtotree for variants reported in samples with low amounts of missing data (~20%). SNPtotree outperformed RAxML in branch resolution for variants reported in sequences with high amounts of missing data (~80%).



REACH OUT!

If you are interested in investigating the Y chromosome, software for Y-chromosomal variant analysis, or have any questions, please reach out to me (Zehra.koksal@sund.ku.dk).
Collaborators are welcome!

Contact me!

Github

REFERENCE LIST

1. Underhill et al. (2007) Annu Rev Genet, 41:539–564., **2**.C. Guyeux et al. (2018) BMC Systems Biology, 12:100., **3**.N. Mizuno et al. (2010) Forensic Sci Int Genet, 4:73–79., **4**. J. J. Wiens et al. (2011) Systematic Biology, 60:719–731., **5**. K. A. Dunn et al. (2003) Mol Phylogenet Evol, 27:259–270., **6**. S. Hartmann et al. (2008) BMC Evolutionary Biology, 8:95., **7**. J. J. Wiens (2006) Journal of Biomedical Informatics, 39:34–42., **8.** D. Darriba et al. (2016) Bioinformatics, 32:1331–1337., **9**. Z. Köksal et al. (2023) Genes, 14:10., **10.** A. Bergström et al. (2016) Curr Biol, 26:809–813., **11**. M. Karmin et al. (2015) Genome Res, 25:459–466., **12**. A. Stamatakis (2014) Bioinformatics, 30:1312–1313.