19/06/2025 19:08:19

Compare Results

Old File:		New File:
old_main.pdf	versus	**paper.pdf**
4 pages (813 KB)		**5 pages (821 KB)**
28/04/2025 19:58:59		19/06/2025 17:29:45

Total Changes

119

Content

28 Replacements

44 Insertions

31 Deletions

Styling and Annotations

3 Styling

13 Annotations

Go to First Change (page 1)

Preprint, 2025, pp. 1–4

doi: DOI HERE
Paper

PAPER

Tsbrowse: an interactive browser for Ancestral Recombination Graphs

Savita Karthikeyan [1,*] Ben Jeffery [1] Duncan Mbuli-Robertson [1] and Jerome Kelleher [1]

[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Old Road Campus, Oxford OX3 7LF, United Kingdom

*Corresponding author. savita.karthikeyan@st-hughs.ox.ac.uk

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Ancestral Recombination Graphs (ARGs) represent the interwoven paths of genetic ancestry of a set of recombining sequences. The ability to capture the evolutionary history of samples makes ARGs valuable in a wide range of applications in population and statistical genetics. ARG-based approaches are increasingly becoming a part of genetic data analysis pipelines due to breakthroughs enabling ARG inference at biobank-scale. However, there is a lack of visualisation tools, which are crucial for validating inferences and generating hypotheses. We present `tsbrowse`, an open-source Python web-app for the interactive visualisation of the fundamental building-blocks of ARGs, i.e., nodes, edges and mutations. We demonstrate the application of `tsbrowse` to various data sources and scenarios, and highlight its key features of browsability along the genome, user interactivity, and scalability to very large sample sizes.

Availability:

Python package: `https://pypi.org/project/tsbrowse/`,
Development version: `https://github.com/tskit-dev/tsbrowse`,
Documentation: `https://tskit.dev/tsbrowse/docs/`,
DOI: `https://doi.org/10.5281/zenodo.15683039`

Key words: Ancestral Recombination Graph, HoloViz, data visualisation, interactive browser

Introduction

Ancestral recombination graphs (ARGs) describe how a set of sample sequences relate to each other at each position along the genome in a recombining species, and are currently the subject of intense research interest (Brandt *et al.*, 2024; Lewanski *et al.*, 2024; Nielsen *et al.*, 2024; Wong *et al.*, 2024). ARGs are a fundamental object in population genetics, and although they have been of theoretical interest for decades (Hudson, 1983; Griffiths and Marjoram, 1996, 1997) it is only with recent breakthroughs in inference methods (Rasmussen *et al.*, 2014; Speidel *et al.*, 2019; Kelleher *et al.*, 2019; Wohns *et al.*, 2022; Zhang *et al.*, 2023; Gunnarsson *et al.*, 2024; Deng *et al.*, 2024) that widespread application has become possible. Varied applications have been proposed, such as inferring selection (Stern *et al.*, 2019; Hejase *et al.*, 2022) and the spatial location of genetic ancestors (Osmond and Coop, 2024; Deraje *et al.*, 2024; Grundler *et al.*, 2025), more powerful approaches to quantifying genetic relatedness (Fan *et al.*, 2022; Zhang *et al.*, 2023; Gunnarsson *et al.*, 2024; Lehmann *et al.*, 2025) and other methodological improvements for genome wide association studies (Nowbandegani *et al.*, 2023; Link *et al.*, 2023), and the development of machine learning methods using inferred ARGs as input (Hejase *et al.*, 2022; Pearson and Durbin, 2023; Korfmann *et al.*, 2024; Whitehouse *et al.*, 2024). While these developments are exciting, the performance of these new methods depends critically on the accuracy of the inferred ARGs. Although studies benchmarking the various inference methods on simulated data have emerged (Brandt *et al.*, 2022; Deng *et al.*, 2024; Peng *et al.*, 2024), the practicalities of applying ARG inference to real data are understudied. In particular, there is a critical lack of software infrastructure to support evaluation and quality control of inferred ARGs.

Visualisation is fundamentally important to data analysis. Many specialised tools exist to aid the visual analysis and quality control of genome assembly (Wick *et al.*, 2015; Challis *et al.*, 2020), read mapping (Robinson *et al.*, 2011), and variant calling (Robinson *et al.*, 2017; Tollefson *et al.*, 2019; König *et al.*, 2023), for example. At every stage of a bioinformatics pipeline, it is important to visualise results to avoid artefacts and aid understanding of the data. Genome browsers such as IGV (Robinson *et al.*, 2011) and the UCSC Genome Browser (Nassar *et al.*, 2023) integrate many different data modalities, and are vital infrastructure for the field.

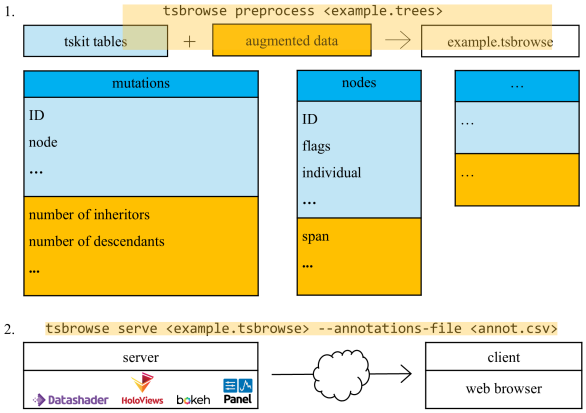

Fig. 1. Overview of tsbrowse architecture. In tskit, ARGs are defined by a collection of tables. Exemplar table names are denoted in dark blue and columns that describe the property in light blue. In the pre-process step, tables are augmented with additional information computed for each property (yellow). The output from this pre-processing step is stored as a `.tsbrowse` file. Next the serve step renders the visualisation in a web browser by leveraging tools in the Holoviz ecosystem. Optional annotations are provided as an input file, allowing users to overlay contextual information about genes or other sequence features.

There is currently no straightforward means of visually summarising ARGs, presenting a significant stumbling block for the nascent field of practical ARG inference. Inferred ARGs are essentially opaque, with only the most basic numerical summaries (such as numbers of nodes, mutations etc) or high-level statistics (Ralph *et al.*, 2020) available. While tools for visualising the local tree topologies exist, they are difficult to interpret and do not scale well to large sample sizes. To address this gap, we present **tsbrowse**, a client-server application providing genome browser-like functionality for ARGs. It provides interactive visualisations of the information structure of ARGs, smoothly scrolling from chromosome-level views down to individual nodes, edges and mutations. Supporting very large ARGs is a particular focus for **tsbrowse**, as millions of genome sequences have been sampled for several species (Cesarani *et al.*, 2022; Stark *et al.*, 2024; Hunt *et al.*, 2024) and ARGs of this scale are a particular focus of ongoing research (Kelleher *et al.*, 2019; Zhang *et al.*, 2023; Zhan *et al.*, 2023; Anderson-Trocmé *et al.*, 2023; Gunnarsson *et al.*, 2024).

Results

Data model

Tsbrowse uses the "succinct tree sequence" encoding of ARGs (Wong *et al.*, 2024). This efficient ARG encoding is implemented by the **tskit** library (Ralph *et al.*, 2020) and supported by most modern ARG simulation (Kelleher *et al.*, 2016, 2018; Haller *et al.*, 2019; Baumdicker *et al.*, 2022; Adrion *et al.*, 2020; Lauterbur *et al.*, 2023; Korfmann *et al.*, 2023; Tsambos *et al.*, 2023; Tagami *et al.*, 2024), inference (Kelleher *et al.*, 2019; Speidel *et al.*, 2019; Wohns *et al.*, 2022; Mahmoudi *et al.*, 2022; Zhan *et al.*, 2023; Zhang *et al.*, 2023; Deng *et al.*, 2025), and processing methods (Fan *et al.*, 2022; Nowbandegani *et al.*, 2023). In **tskit**, ARGs are encoded as a collection of tables, storing information about the nodes and edges that describe the graph topology and the sites and mutations that encode the sequence variation (Fig. 1).

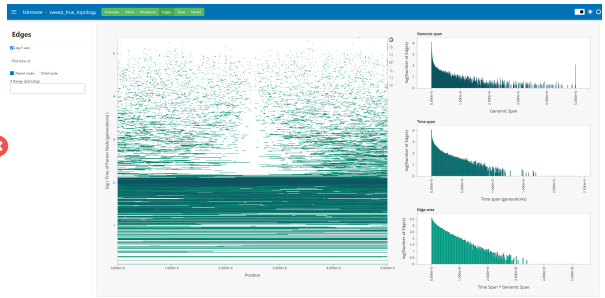

Fig. 2. Screenshot of the Edges view in **tsbrowse**. Visualisation of the edges in a simulated ARG with a strong selective sweep in the middle of the genome (see Supplement for details). Edges are shown in the main browser pane on the left, with additional histograms summarising the edges on the right. The effect of the sweep can be seen by the lack of edges crossing the centre of the simulated chromosome due to an excess of recent coalescent events. Moving away from the focal site, the oldest edges are the first to rejoin the ARG, followed by more recent edges, resulting in a wedge-like pattern of missing edges in the centre.

Architecture

Tsbrowse is a modular client-server application written in Python, optimised for ARGs with millions of samples (Fig. 1). The application has two basic commands: **preprocess** and **serve**. The **preprocess** command takes an input ARG file and augments the **tskit** tables with additional columns, precomputing all the information required for visualisation, and storing it as a compressed `.tsbrowse` file. To visualise an ARG we then run **tsbrowse serve**, which by default will open a browser window on the local machine, but also supports running the server on a remote machine across the network. This client-server architecture has some important advantages over a monolithic single-machine approach. Most importantly, the ARG being visualised remains in-situ and does not need to be downloaded from the server.

The goal of **tsbrowse** is to provide interpretable and interactive visual summaries of ARGs containing millions of nodes, edges and mutations. Mutations, for example, have a clear interpretation when plotted on genome coordinate vs time axes, but at this scale the data density is far too high to simply plot each mutation as a point. We overcome this problem by using Datashader and the wider Holoviz ecosystem (Holoviz developers, 2024), which efficiently rasterises large datasets at the requested resolution on the server, sends the image to the web browser for display, and dynamically updates as the user interactively navigates the ARG. This approach allows us to summarise very large ARGs interactively; Fig. S1 for example, shows a screenshot of **tsbrowse** summarising the 1.9 million mutations in a SARS-CoV-2 ARG with around 2.7 million nodes and edges.

User Interface

The user interface is presented as a dashboard, with views to describe various aspects of the ARG. Fig. 2 demonstrates the user interface using the Edges view as an example. The plot on the left is an overview of the 33,929 edges in a simulated ARG with a strong selective sweep (see Supplement for details). Each edge in the ARG is depicted as a horizontal line connecting the genomic coordinates of the parent and child nodes on the x-axis, and the y-axis shows the time of either the parent or the child, as chosen by the user. The user can interact with the main "genome browser" window using a set of controls on the top-right corner provided by Bokeh (Bokeh Development Team, 2018), allowing them to pan and zoom as required. The histograms to the right then summarise the edges as depicted in the browser window. A similar browser

interface is provided for mutations (e.g., Fig. S1) and nodes (e.g., Fig. S2). `tsbrowse` also provides an interactive table viewer with flexible searching and sorting utilities, which is a valuable debugging utility for developers.

Applications

The purpose of `tsbrowse` is to provide an interactive view of the `tskit` ARG data model to guide intuition, improve inference quality control and facilitate debugging. An example of how we can deepen our understanding using the genome-browser-like perspective provided by `tsbrowse` is given in the simulated ARG of Fig. 2, where the gap in the Edges view corresponds to the characteristic dip in diversity of a selective sweep. Comparing this ground truth to the ARGs inferred by four different inference methods in Fig. S3, we can see that there are substantial qualitative differences between the results. These differences in ancestral haplotypes illustrated by the Edges view are unlikely to be captured by the tree-by-tree distance metrics usually used to evaluate inference methods (e.g. Kelleher *et al.*, 2019; Zhang *et al.*, 2023), providing further motivation for new and improved ways to compare simulated and inferred ARGs (Fritze *et al.*, 2024).

Interest in ARG-based methods is burgeoning, but the methods are new and practical guidance on applying inferences to real data is lacking. Data filtering is essential, and the effects of the choices that must be made along any bioinformatics pipeline on the final ARG are hard to predict and quantify. `Tsbrowse` was primarily developed as a way to quickly visualise the effects of such filtering choices on ARGs inferred by `tsinfer`, and it is now an indispensable element of the inference pipeline. Fig. S4 shows a region of the 1000 Genomes data with gaps in site density that are spanned by exceptionally long edges, which are likely to bias downstream statistics. These interactive visualisations have also helped diagnose issues with the `tsinfer` inference algorithm. Fig. S2 shows the genomic spans of ancestral nodes in an ARG inferred with `tsinfer`, demonstrating a clear excess of long haplotypes in the very ancient past. These insights have helped guide development and may lead to significant improvements in performance.

Discussion

Visualisation of tree topology is a central task in phylogenetics, and although numerous tools exist (e.g. Huson *et al.*, 2007; Vaughan, 2017), the methods can typically only handle a few hundred nodes (but see e.g., Hadfield *et al.* (2018)). Visualisation of large-scale tree topologies with millions of nodes requires much more sophisticated approaches to capture topological features at different scales, and is an active research area (Wong and Rosindell, 2022; Kramer *et al.*, 2023). Adapting such methods, and integrating them into `tsbrowse` to provide a local tree viewer that operates at the million-node scale is an important direction for future work. An interactive viewer for the entire ARG topology, capturing semantic properties of the graph at a range of scales, is an even more ambitious goal, and would be a major asset for the field.

Conflict of interest

No competing interest is declared.

Funding

SK acknowledges full support and funding from Novo Nordisk Research Centre Oxford, and funding from the Biotechnology and Biological Sciences Research Council (UKRI-BBSRC) [grant number BB/T008784/1]. DMR is funded by a studentship from the Wellcome Programme in Genomic Medicine and Statistics. JK acknowledges EPSRC (research grant EP/X024881/1), NIH (research grants HG011395 and HG012473) and the Robertson Foundation.

References

Adrion, J. R., Cole, C. C., Dukler N., *et al.* (2020). A community-maintained standard library of population genetic models. *eLife*, **9**.

Anderson-Trocmé L., Nelson D., Zabad S., *et al.* (2023). On the genes, genealogies, and geographies of Quebec. *Science*, **380**.

Baumdicker F., Bisschop G., Goldstein D., *et al.* (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, **220**.

Bokeh Development Team (2018). Bokeh: Python library for interactive visualization.

Brandt D., Wei X., Deng Y., Vaughn A. H., and Nielsen R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, **221**(1), iyac044.

Brandt D. Y., Huber C. D., Chiang C. W., and Ortega-Del Vecchyo D. (2024). The promise of inferring the past using the ancestral recombination graph. *Genome Biology and Evolution*, **16**(2), evae005.

Cesarani A., Lourenco D., Tsuruta S., Legarra A., Nicolazzi E., VanRaden P., and Misztal I. (2022). Multibreed genomic evaluation for production traits of dairy cattle in the united states using single-step genomic best linear unbiased predictor. *Journal of Dairy Science*, **105**(6), 5141–5152.

Challis R., Richards E., Rajan J., Cochrane G., and Blaxter M. (2020). Blobtoolkit–interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*, **10**(4), 1361–1374.

Deng Y., Nielsen R., and Song Y. S. (2024). Robust and Accurate Bayesian Inference of Genome-Wide Genealogies for Large Samples. *bioRxiv*.

Deng Y., Song Y. S., and Nielsen R. (2025). A general framework for branch length estimation in ancestral recombination graphs. *bioRxiv*, pages 2025–02.

Deraje P., Kitchens J., Coop G., and Osmond M. M. (2024). Inferring the geographic history of recombinant lineages using the full ancestral recombination graph. *bioRxiv*, pages 2024–04.

Fan C., Mancuso N., and Chiang C. W. K. (2022). A genealogical estimate of genetic relationships. *The American Journal of Human Genetics*, **109**(5), 812–824.

Fritze H., Pope N., Kelleher J., and Ralph P. (2024). A forest is more than its trees: haplotypes and inferred args. *bioRxiv*, pages 2024–11.

Griffiths R. and Marjoram P. (1997). An ancestral recombination graph. In *Progress in population genetics and human evolution*, pages 257 – 270. Springer. Conference date: 01-01-1997.

Griffiths R. C. and Marjoram P. (1996). Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology*, **3**(4), 479–502.

Grundler M. C., Terhorst J., and Bradburd G. S. (2025). A geographic history of human genetic ancestry. *Science*, **387**(6741), 1391–1397.

Gunnarsson Á. F., Zhu J., Zhang B. C., Tsangalidou Z., Allmont A., and Palamara P. F. (2024). A scalable approach for genome-wide inference of ancestral recombination graphs. *bioRxiv*, pages 2024–08.

Hadfield J., Megill C., Bell S. M., Huddleston J., Potter B., Callender C., Sagulenko P., Bedford T., and Neher R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, **34**(23), 4121–4123.

Haller B. C., Galloway J., Kelleher J., *et al.* (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, **19**, 552–566.

Hejase H. A., Mo Z., Campagna L., and Siepel A. (2022). A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Molecular Biology and Evolution*, **39**(1), msab332.

Holoviz developers (2024). High-level tools to simplify visualization in Python. `https://holoviz.org/`. Accessed: 2024-10-03.

Hudson R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**.

Hunt M., Hinrichs A. S., Anderson D., Karim L., Dearlove B. L., Knaggs J., Constantinides B., Fowler P. W., Rodger G., Street T., *et al.* (2024). Addressing pandemic-wide systematic errors in the sars-cov-2 phylogeny. *bioRxiv*.

Huson D. H., Richter D. C., Rausch C., Dezulian T., Franz M., and Rupp R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC bioinformatics*, **8**, 1–6.

Kelleher J., Etheridge A. M., and McVean G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLOS Computational Biology*, **12**(5), e1004842.

Kelleher J., Thornton K. R., Ashander J., and Ralph P. L. (2018). Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology*, **14**(11), e1006581.

Kelleher J., Wong Y., Wohns A. W., *et al.* (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, **51**(9), 1330–1338.

König P., Beier S., Mascher M., Stein N., Lange M., and Scholz U. (2023). DivBrowse–interactive visualization and exploratory data analysis of variant call matrices. *GigaScience*, **12**, giad025.

Korfmann K., Abu Awad D., and Tellier A. (2023). Weak seed banks influence the signature and detectability of selective sweeps. *Journal of Evolutionary Biology*, **36**(9), 1282–1294.

Korfmann K., Sellinger T. P. P., Freund F., Fumagalli M., and Tellier A. (2024). Simultaneous inference of past demography and selection from the ancestral recombination graph under the beta coalescent. *Peer Community Journal*, **4**.

Kramer A. M., Sanderson T., and Corbett-Detig R. (2023). Treenome Browser: co-visualization of enormous phylogenies and millions of genomes. *Bioinformatics*, **39**.

Lauterbur M. E., Cavassim M. I. A., Gladstein A. L., *et al.* (2023). Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *eLife*, **12**.

Lehmann B., Lee H., Anderson-Trocme L., Kelleher J., Gorjanc G., and Ralph P. L. (2025). On args, pedigrees, and genetic relatedness matrices. *bioRxiv*.

Lewanski A. L., Grundler M. C., and Bradburd G. S. (2024). The era of the arg: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *Plos Genetics*, **20**(1), e1011110.

Link V., Schraiber J. G., Fan C., *et al.* (2023). Tree-based QTL mapping with expected local genetic relatedness matrices. *The American Journal of Human Genetics*, **110**(12), 2077–2091.

Mahmoudi A., Koskela J., Kelleher J., Chan Y.-b., and Balding D. (2022). Bayesian inference of ancestral recombination graphs. *PLOS Computational Biology*, **18**(3), e1009960.

Nassar L. R., Barber G. P., Benet-Pagès A., *et al.* (2023). The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research*, **51**.

Nielsen R., Vaughn A. H., and Deng Y. (2024). Inference and applications of ancestral recombination graphs. *Nature Reviews Genetics*, pages 1–12.

Nowbandegani P. S., Wohns A. W., Ballard J. L., *et al.* (2023). Extremely sparse models of linkage disequilibrium in ancestrally diverse association studies. *Nature Genetics*, **55**, 1494–1502.

Osmond M. and Coop G. (2024). Estimating dispersal rates and locating genetic ancestors with genome-wide genealogies. *eLife*, **13**, e72177.

Pearson A. and Durbin R. (2023). Local ancestry inference for complex population histories. *bioRxiv*, pages 2023–03.

Peng D., Mulder O. J., and Edge M. D. (2024). Evaluating arg-estimation methods in the context of estimating population-mean polygenic score histories. *bioRxiv*.

Ralph P., Thornton K., and Kelleher J. (2020). Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics*, **215**(3), 779–797.

Rasmussen M. D., Hubisz M. J., Gronau I., and Siepel A. (2014). Genome-wide inference of ancestral recombination graphs. *PLOS Genetics*, **10**(5), e1004342.

Robinson J. T., Thorvaldsdóttir H., Winckler W., Guttman M., Lander E. S., Getz G., and Mesirov J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, **29**(1), 24–26.

Robinson J. T., Thorvaldsdóttir H., Wenger A. M., Zehir A., and Mesirov J. P. (2017). Variant review with the integrative genomics viewer. *Cancer research*, **77**(21), e31–e34.

Speidel L., Forest M., Shi S., *et al.* (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, **51**, 1321–1329.

Stark Z., Glazer D., Hofmann O., Rendon A., Marshall C. R., Ginsburg G. S., Lunt C., Allen N., Effingham M., Hastings Ward J., *et al.* (2024). A call to action to scale up research and clinical genomic data sharing. *Nature Reviews Genetics*, pages 1–7.

Stern A. J., Wilton P. R., and Nielsen R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, **15**(9), e1008384.

Tagami D., Bisschop G., and Kelleher J. (2024). tstrait: a quantitative trait simulator for ancestral recombination graphs. *Bioinformatics*, **40**(6), btae334.

Tollefson G. A., Schuster J., Gelin F., Agudelo A., Ragavendran A., Restrepo I., Stey P., Padbury J., and Uzun A. (2019). VIVA (VIsualization of VAriants): a VCF file visualization tool. *Scientific Reports*, **9**(1), 12648.

Tsambos G., Kelleher J., Ralph P., Leslie S., and Vukcevic D. (2023). Link-ancestors: fast simulation of local ancestry with tree sequence software. *Bioinformatics Advances*, **3**(1), vbad163.

Vaughan T. G. (2017). IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics*, **33**(15), 2392–2394.

Whitehouse L. S., Ray D. D., and Schrider D. R. (2024). Tree sequences as a general-purpose tool for population genetic inference. *Molecular Biology and Evolution*, **41**(11), msae223.

Wick R. R., Schultz M. B., Zobel J., and Holt K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**(20), 3350–3352.

Wohns A. W., Wong Y., Jeffery B., Akbari A., Mallick S., Pinhasi R., Patterson N., Reich D., Kelleher J., and McVean G. (2022). A unified genealogy of modern and ancient genomes. *Science*, **375**(6583), eabi8264.

Wong Y. and Rosindell J. (2022). Dynamic visualisation of million-tip trees: The OneZoom project. *Methods in Ecology and Evolution*, **13**(2), 303–313.

Wong Y., Ignatieva A., Koskela J., Gorjanc G., Wohns A. W., and Kelleher J. (2024). A general and efficient representation of ancestral recombination graphs. *Genetics*, **228**(1), iyae100.

Zhan S. H., Ignatieva A., Wong Y., Eaton K., Jeffery B., Palmer D. S., Murall C. L., Otto S. P., and Kelleher J. (2023). Towards pandemic-scale ancestral recombination graphs of sars-cov-2. *bioRxiv*.

Zhang B. C., Biddanda A., Gunnarsson Á. F., *et al.* (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nature Genetics*, **55**.