Inferring whole-genome histories in large population datasets
Kelleher et al. (2019)[1]

Summary

In this article, Kelleher and colleagues make use of a recently developed datastructure named tree sequences[2] for whole-genome ancestry inference.
Tree sequences is an alternative to the available but computationally complex ancestral recombination graphs (ARGs). While both methods allow us to represent histories of organisms which recombine in an accurate way, the tree sequence datastructure is much more efficient and scalable.
The main focus of this article is a tool which efficiently infers those tree sequences from genetic data (`tsinfer` available on github.com/tskit-dev/tsinfer).

Efficiency and scalability is much needed as the size of whole genome datasets steadily grow. In addition to demonstrate the practical use of `tsinfer` on three datasets (from $n = 277$ to $n = 487'000$), the authors also compare required storage space, accuracy and speed with state-of-the-art tools.
This benchmarking was done using simulated data (generated by msprime[3]). Not only does `tsinfer` perform as good as or better when compared to the state-of-the-art tools, it is shown that it would be in theory possible to infer tree sequences of the entire human population.

There are some limitations and assumptions made which is acknowledged by the authors. However solutions to some of those has already been worked on in the meantime.

Reflection

After reading the article for the first time, there were still many things unclear to me. The main reason for this was that the basis (data structure of tree sequences) was presented in a previous research article of the same first author. It is therefore not an independent work but some context is crucial for understanding. As a consequence of this, I had to read the previous article to better understand this one.
To comprehend the details of the method, consulting the supplementary information was also helpful as well as recorded presentations of two of the authors within the scope of the phyloseminar[4,5].

Since it's quite different to prepare slides using R Markdown compared to other tools (e.g. Microsoft PowerPoint, Keynote) it required a bit more time to create the final presentation.
The default beamer presentation is in my opinion not very exciting to look at, but I've found a very nice alternative in xaringan[6] and xaringanExtra[7]. Generally I think using R Markdown to create slides would highlight its strengths better if more code snippets are used. For this presentation I mainly just used titles and screenshots of figures from the article which does not tap the full potential of R Markdown.
Nevertheless I always like using R Markdown and took this opportunity to learn new methods to create presentations with it. I also learned, in addition to do version control with Git, how to publish the presentation online using GitHub Pages.

Presenting remotely using Zoom is still something to get used to. Difficult for me was mainly the lack of interaction before the talk as this usually calms my nerves a bit.
Regarding the content of the presentation, the main challenge was to narrow it down to the minimum required for understanding and not go over the time limit. I tried to achieve this by structuring the presentation a bit larger/longer than the target was and then gradually remove non-crucial information until it was met.
Also was it sometimes necessary to simplify certain things considering the target audience are not experts on this field. In these situations I prefer to rather leave something out or not explain a completely correct but a simplified version in order to remain understandable.

I'm generally satisfied with the outcome, everything worked from the technological aspect as I've tested it before on a separate Zoom meeting. From re-watching the recording I was surprised how slow I spoke. I however hope it was not too much to interrupt the flow of the presentation. Timewise, I think I still presented faster than planned and finished therefore a bit under the aimed 25 minutes.
Something I'm not happy about was the loss for words I had at one point. It took me quite a bit to gather myself until I could continue. Having the talk prepared a bit better might have helped to avoid this situation.

A nice experience were the many interesting questions I got during the discussion. Most of them where on topics I had not had time to go into detail during the main talk. This indicates that the audience could (hopefully) follow and understand the topic and its main strengths and gaps.

As mentioned at the end of the talk, I think this method could also be applied to viral organisms (the field I'm working in) as long as we have sampled a representative fraction of viral genome sequences. It has been shown that recombination or reassortment events also happen in certain viruses and in order to consider this for accurately inferring histories, appropriate methods would need to be applied.[8,9]
The difficulties with that might not be to apply this method but more to prepare the sequences. `Tsinfer` requires phased and aligned sequences which can be, using the current sequencing techniques, quite a difficult task as most viral genomes are very diverse within a host.

Final thoughts

Generally it was a very nice experience reading this well written article but some basic knowledge on this topic was required beforehand to understand the context.
I think this article is a very good model on how to publish a computational tool. It's not only important to understand the algorithm (which is described in the article) but also how the tools (`tskit` and `tsinfer`) can actually be used. In order to achieve this, they document all their tools very well using readthedocs and have the source code open on GitHub.
In addition to that, the analysis done for the article can even be replicated on your own using code which is also published on GitHub (github.com/mcveanlab/treeseq-inference). This provides full transparency and is hopefully the path to follow for publishing any scientific work.

References

1. Kelleher, J. *et al.* Inferring whole-genome histories in large population datasets. *Nature Genetics* **51**, 1330–1338 (2019).
2. Kelleher, J., Thornton, K. R., Ashander, J. & Ralph, P. L. Efficient pedigree recording for fast population genetics simulation. *PLOS Computational Biology* **14**, e1006581 (2018).
3. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* **12**, e1004842 (2016).
4. Phyloseminar #96: Wilder Wohns (Oxford). (2020).
5. Phyloseminar #97: Yan Wong (Oxford). (2020).
6. Xie, Y. *Xaringan: Presentation ninja.* (2020).
7. Aden-Buie, G. *xaringanExtra: Extras and extensions for xaringan slides.* (2020).
8. Neches, R. Y., McGee, M. D. & Kyrpides, N. C. Recombination should not be an afterthought. *Nature Reviews Microbiology* 1–1 (2020) doi:10.1038/s41579-020-00451-1.
9. Müller, N. F., Stolz, U., Dudas, G., Stadler, T. & Vaughan, T. G. Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences* **117**, 17104–17111 (2020).