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 called tree sequences.[2]
Tree sequences is an alternative to the 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 grows. 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 performs `tsinfer` 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 living 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 authors. 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 strenghts better if more code snippets are used. For this presentation I mainly used a title and a screenshot of a figure from the article which does not tap the full potential of R Markdown.

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 down the content to the minimum required for understanding and not go over the time limit. Also was it sometimes necessary to simplify certain things considering the target audience are not experts on this field. I prefer to leave something out or not explain something 100% correct in order to remain understandable.

Generally glad with the outcome, everything worked from the technological aspect as I've tested it before on a separate Zoom meeting.

Loss for words at one point. Had to take some time to gather myself which might have interrupted the flow of the presentation. Presented a bit faster than planned and finished therefore a bit under the aimed 25 minutes.

Many good questions I had not time to go into detail during the main talk.

Very well written article, but some context knowledge needed as basis.

Very well documented codebase which can serve as a good example for own projects. Can be replicated using code published on GitHub (github.com/mcveanlab/treeseq-inference)

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).