Rapid Genotype Imputation of Biobank-Scale Whole Genome Sequence Data Using Tree Sequences

Shing Hei Zhan **Yan Wong** *Benjamin Jeffery* ***Jerome Kelleher***

Big Data Institute, Li Ka Shing Centre for Health Information and Discovery
University of Oxford, Oxford, UK

Yan! Help!!! 9:55 a.m.

Hello, Shing. Is everything alright? What is wrong? 9:56 a.m.

Storing and analyzing all the UK Biobank whole genome sequence data and doing ***genotype imputation*** on it using standard methods broke my piggy bank!!!!! 😭😭😭 9:57 am.

Have you heard of succinct tree sequences, or ***STS***? 9:57 a.m.

They are a *lossless*, *compressed* data structure that simultaneously allows for efficient and rapid processing and storage of genetic variants and genealogies across an entire genome. 9:58 a.m.

They scale very well to millions of human genomes. 9:58 a.m.

Woah, wuuuuuut? Can they do genotype imputation too? 10:00 am.

Yes, they can indeed. Because STS store all the variants and the genealogies, it should be possible to do genotype imputation very accurately when the genealogy is available or even inferred. 10:02 am.

See this **Result** comparing ***STS*** and ***BEAGLE***, which is a popular imputation method, on some simulated genotype data. 10:04 am.

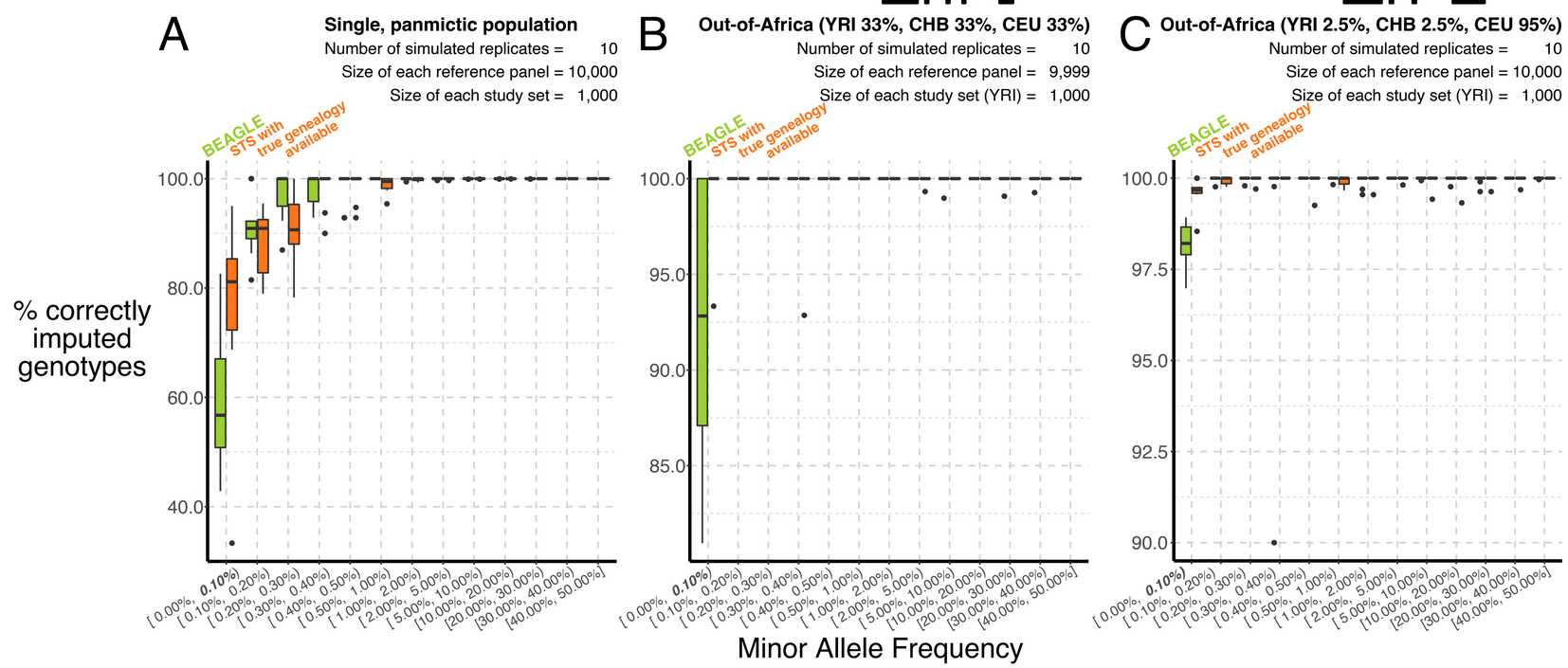
Yoooo, STS are so sick! They can impute rare genotypes that have stumped standard methods too?! 10:06 am.

OOOOOOMMMMMMGGGGGG!!!!!! I need to conda install it NOW!!! Thanks so much, Yan! 🤩🤩🤩 10:06 am.

You are welcome, Shing. You can also learn more about STS and the software ecosystem ***tskit*** here. 10:09 am.

tskit Kelleher et al. (2019) *Nature Genetics*

Result



Methods TL;DR

Software demes; msprime; tskit; tsinfer
Parameters sequence length = 1,000,000 bp; uniform recombination rate = 1 x 10-8; uniform mutation rate = 1 x 10-8; proportion of missing genotypes = 0.50

Acknowledgements

Robertson Foundation & Janssen Pharmaceuticals