Updating an efficient pipeline for local ancestry inference

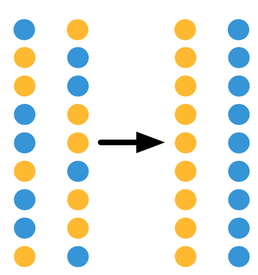
E.M. Jewett*, G.D. Poznik, K.F. McManus, K. Bryc, J.L. Mountain, A. Auton, and the 23andMe Research Team
23andMe, Inc. Mountain View, CA - research.23andme.com *ejewett@23andme.com

Abstract

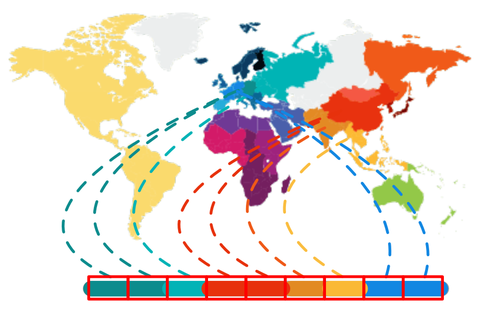
The goal of local ancestry inference is to identify the population of origin at each base in the genome of an individual. At 23andMe, inference is performed by an efficient pipeline known as Ancestry Composition. Ancestry Composition employs support vector machines (SVM) to assign local ancestries to windows along the genome, as well as a hidden Markov model to combine information across window-level SVM calls. The Ancestry Composition pipeline is computationally efficient and has demonstrated low error rates on both simulated and real data. Here, we explore methods for improving the window-level classification component of the pipeline by introducing a method based on the positional Burrows-Wheeler transform (PBWT). This adaptation of Ancestry Composition enables window-free local ancestry estimation and improves computational efficiency.

The Ancestry Composition pipeline

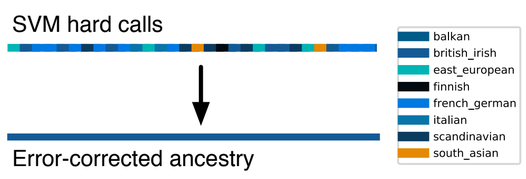
The ancestry composition (AC) pipeline (Durand et al., 2014) provides chromosome paintings through a four-step process:



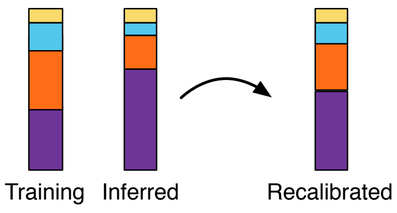
1. **Haplotype phasing**: Haplotypes are phased using Eagle (Loh et al 2017).



2. **SVM classification:** Chromosomes are divided into windows of fixed length. An SVM classification approach is used to obtain an initial ancestry estimate in each window.



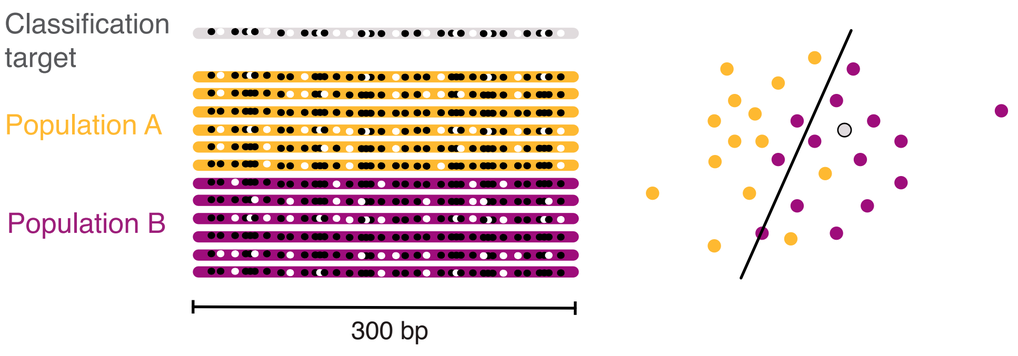
3. **Error correction:** An HMM is applied to reduce noise in the SVM classifications.



4. **Recalibration:** Posterior ancestry estimates in each window are weighted to reflect the observed distribution of ancestry.

SVMs for Ancestry Composition

One SVM model is trained for each pair of populations in each window. Training has complexity $O(WK^2)$ for K populations and W windows.

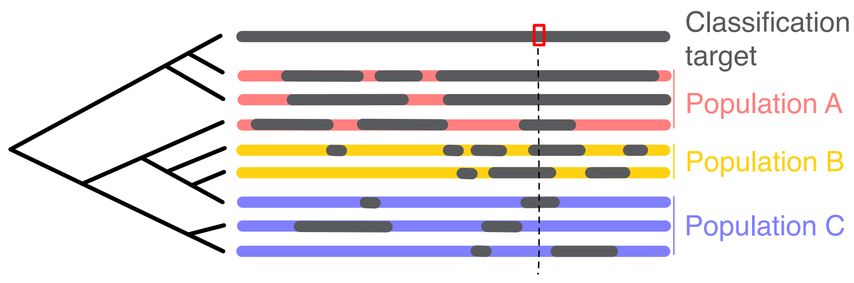


Motivation for PBWT-based estimation

- Facilitates expansion of a reference panel to many populations.
- Obviates the need for window-based estimates.
- Facilitates estimation in reference panels of admixed individuals.

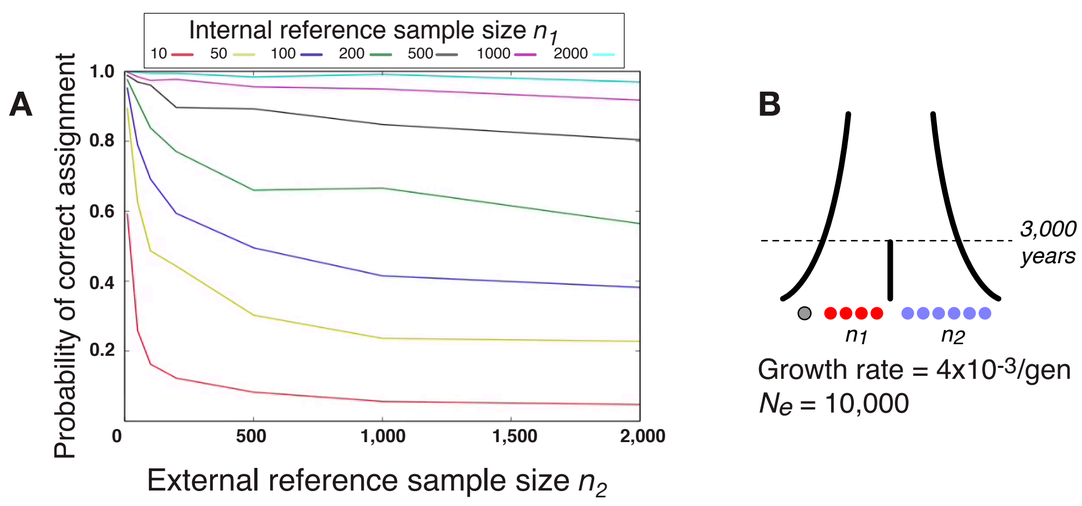
Positional Burrows-Wheeler Transforms for Ancestry Composition

Positional Burrows-Wheeler Transforms (PBWTs) allow fast matching of a target sequence to haplotypes in a reference panel of many populations (Durbin, 2014). The ancestry of the longest PBWT match at a given position provides an estimate of the ancestry at that position.



The accuracy of PBWT-based estimates

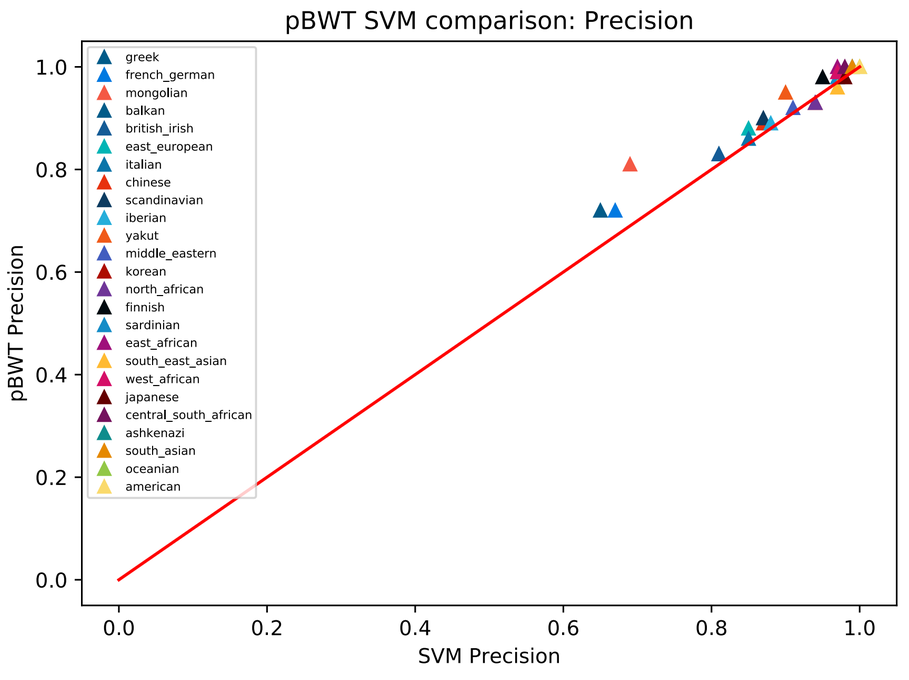
The reference haplotype with the longest PBWT match at a given position provides an increasingly accurate estimate of local ancestry as the size of the sample from the true population increases.



Panel A of the figure shows the probability that the PBWT-based ancestry estimate in an Ancestry Composition window matches the true ancestry. Probabilities were simulated under the model in Panel B. The reference panel was composed of n_1 haplotypes from the true ancestral population (red dots in Panel B) and n_2 haplotypes from a related population (blue dots in Panel B).

Comparison of final AC estimates for SVM and PBWT

Replacing the SVM module of the AC pipeline with a module based on the PBWT produces estimates that have similar accuracy, while reducing the computational complexity of training.



The figure shows precision, computed as the fraction of Ancestry Composition windows in which the inferred ancestry matches the true ancestry in holdout samples from the 23andMe Ancestry Composition reference panel.

References

1. E Durand et al. (2014). Ancestry Composition: A novel, efficient pipeline for ancestry deconvolution. bioRxiv 010512; doi: https://doi.org/10.1101/010512.
2. PR Loh et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*. 48(11):1443-1448
3. R Durbin (2014). Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*. 30(9):1266–1272.

Acknowledgements

We thank the employees and research participants of 23andMe who made this research possible.