High Resolution Ancestry Deconvolution for Next Generation Sequencing

Helgi Hilmarsson[1,3], Arvind Kumar[1,3], Richa Rastogi[2], Daniel Mas Montserrat[2], Carlos D. Bustamante[2], Alexander G. Ioannidis[1,2,*]

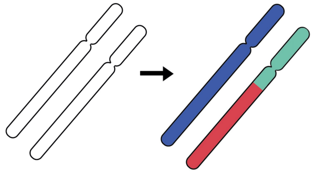
[1]Institute for Computational and Mathematical Engineering, Stanford University

[2]Department of Biomedical Data Science, Stanford University, Stanford

[3]Authors contributed equally to this work

[*]Correspondence: ioannidis@stanford.edu

Local Ancestry Inference



Genomic medicine promises increased resolution for accurate diagnosis, personalized treatment and identification of population-wide health burdens.

Local Ancestry Inference (LAI) refers to the ancestry classification task for short segments of DNA sequences and can be used to detect and account for genetic inheritance and population structure when it comes to genomic medicine.

Gnomix

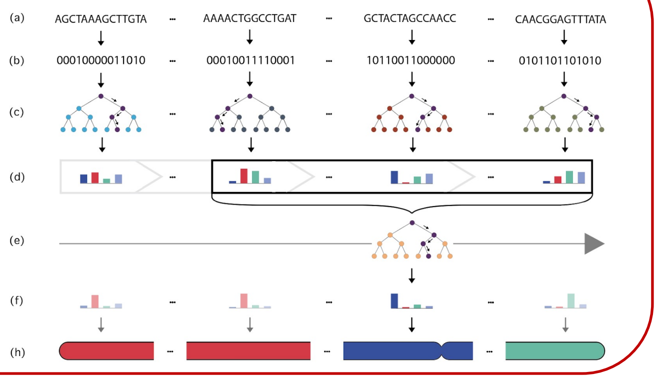
a) Encode DNA sequence variants as 0s and 1s

b) Split sequence into windows and learn a base ML model for each window.

c) Get ancestry estimates for each window.

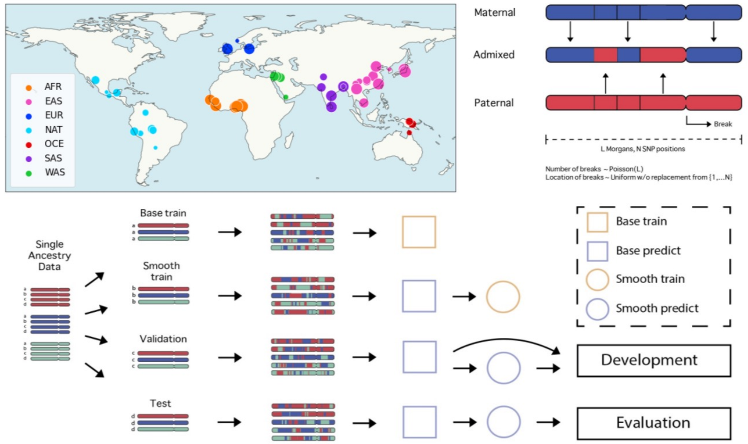
d-e) Sliding window ML model (smoother) learns to aggregate base model probabilities

f) Smoothing module outputs final ancestry prediction along the segmented DNA.



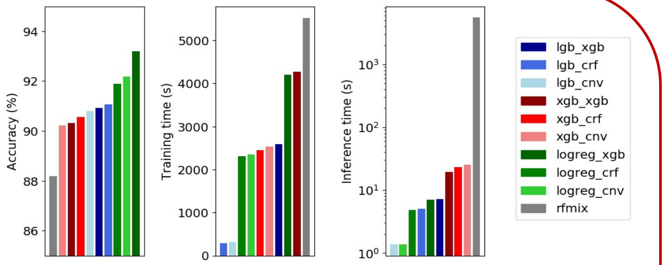
Data Generation and Learning Procedure

To know the true ancestry of a given individual's DNA segment is hard. Our solution is to simulate admixed (of many ancestries) individuals from a set of single ancestry individuals and keep track of their local ancestry. Since both modules (base and smoothing) are data-driven, we split the training data into two disjoint set, one for each module.

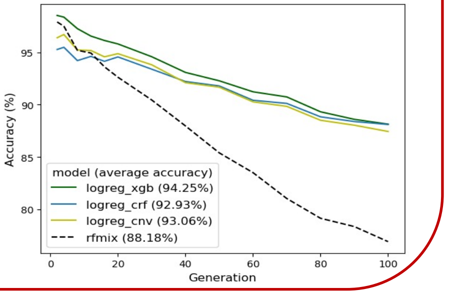


Results

- More accurate inference and faster training than current state of the art (rfmix)



- Robustness to generation (time since admixture) and high accuracy (~90%) for high generations (~100) on dataset with 7 different populations.


Accuracy (%)
Generation
model (average accuracy)
logreg_xgb (94.25%)
logreg_crf (92.93%)
logreg_cnv (93.06%)
rfmix (88.18%)