CS 229 Project

Local Ancestry Inference in Admixed Populations

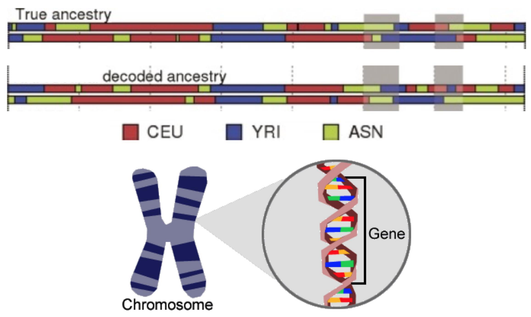
Naveen Arivazhagan, Hye Ji Kim, Edwin Yuan

Computer Science, Electrical Engineering, Applied Physics, Stanford Universiy

STANFORD UNIVERSITY

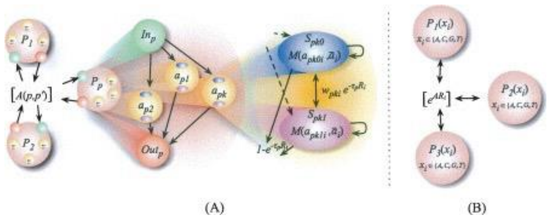
Introduction

Local Ancestry Inference Problem

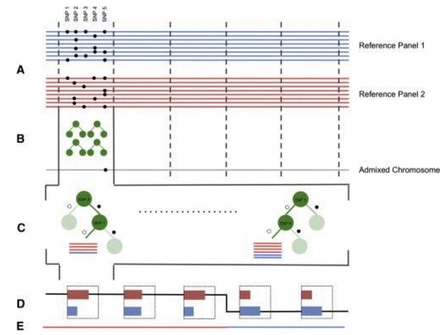


Data:
The 1000 Genomes Projects Phase I – A dataset of 1092 individuals from 14 populations (2013)

HAPAA algorithm:
• Step 1: Divide the genome into continguous windows of SNP's (single neucleotide polymorphisms)
• Step 2: Assign ancestries for genes using a **Hidden Markov Model based** clustering algorithm



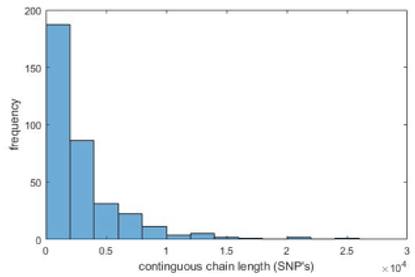
RFMix algorithm:
• Step 1: Segment an input strand of DNA into contiguous windows of SNPs(single nucleotide polymorphisms)
• Step 2: Assign ancestries for genes using a **Conditional random fields (CRF) algorithm** based on randomforests



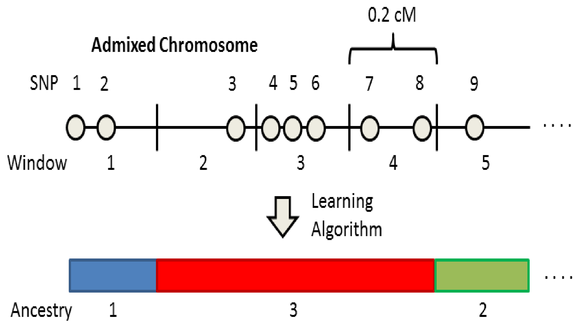
Methods

Data:
• We used pre-processed data that the authors of the RFMix paper provided.
• 51213 SNP's from both chromosome one's of 362 individuals (bi-allelic).
• Test set:10 admixed, Latino individuals, whose genomes were created using a Wright-Fischer simulation to sample 12 generations after admixture.
• The simulated Latino genomes have 45% Native American, 50% European, and 5% African ancestry.
• Training set: 170 Native American, 194 European, and 340 African (Simulated samples were used)

Our scheme:
• Step 1: Segment into windows of fixed size in **centi-morgans**
-The histogram shows distribution of contiguous window sizes in terms of number of SNP's

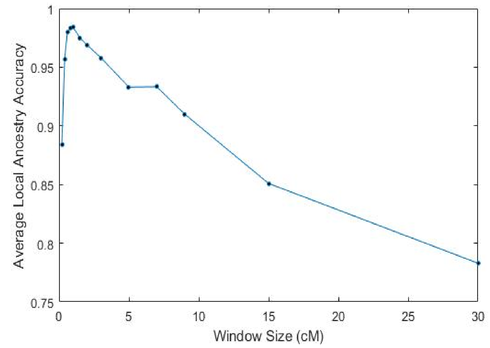


• Step 2: Measure the **Manhattan distances** between the references and thetest sequence in each window
- **Manhattan distance**: counting the number of replacements needed to get rom one window to another.
• Step 3: We then use a voting scheme where the ancestry of the admixed window is assigned the reference population in which it has largest number of to he highest similarity values. (All SNPs in a window are assigned to the same ancestry.)



Results

• By default, using the same test and training set, RFMix uses windows of 0.2 cM and achieves an accuracy of 97.5% averaged over the 10 admixed individuals.



• The results suggest **very high performance**, compared to RFMix, peaking at around 98.4% accuracy for a window size of 1.0 cM.
• For the same window size RFMix uses, of 0.2 cM, the accuracy is 88.4%.
• As window size is increased, the accuracy peaks and then falls rapidly.

Why does ours perform so well?
• We have an extreme abundance of reference panels on which to train in the data set.

What if we only have a small number of references?
• 30 reference(training)examples for each population
Ours: ~97%
RFMix: 95.6%
• 3 reference(training) examples for each population
Ours: 76%
RFMix: 87.8%, 93.2% after one iteration of EM

• RFMix outperforms our simple voting scheme, because of its ability to **construct new reference data from the existing admixed samples**

Way to improve our scheme

• Relax the independence assumptions between the ancestry labels of nearby windows

• Model the recombination process to deal with the few reference panels case

Conclusions

• We seek to better understand how to construct models for local ancestry inference. There are multiple steps forward:

• Reconstructing the recombination process:

- HAPAA utilizes a specific biological model of recombination
- We will construct a model of recombination that allows us simulate more admixed data
- observe how the recombination model affects LAI performance

• Investigate alternative non-biological models of LAI

- We already have created a simple model of LAI using the Manhattan metric that performs well with abundance of training data
- Implement more complex, non-biological HMM based algorithm for sequence recognition
- compare performance of biological with non-biological models
- Implement EM into our non-biological models

Acknowledgement

We thank Irene for guiding us the directions, and Volodymyr for introducing us the problem and answering questions. Hyeji Kim thanks Jason Junjie for guiding us with how to get the data and helping us to understand the genetics.

Bibliography

1.Maples BK, Gravel S, Kenny EE, and Bustamante CD. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. 93, 278-288
2. Sundquist A, Fratkin E, Do CB, Batzoglou S. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Research. 2008;18(4):676-682. doi:10.1101/gr.072850.107.