Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups

Written by Salter-Townshend & Myers (2019)
Presented by Seth Temple

https://www.genetics.org/content/212/3/869

Context for ancestry inference

- Li and Stephens (2003)
- STRUCTURE
- ChromoPainter & FineSTRUCTURE
- GLOBETROTTER
- HapMix
- RFMix

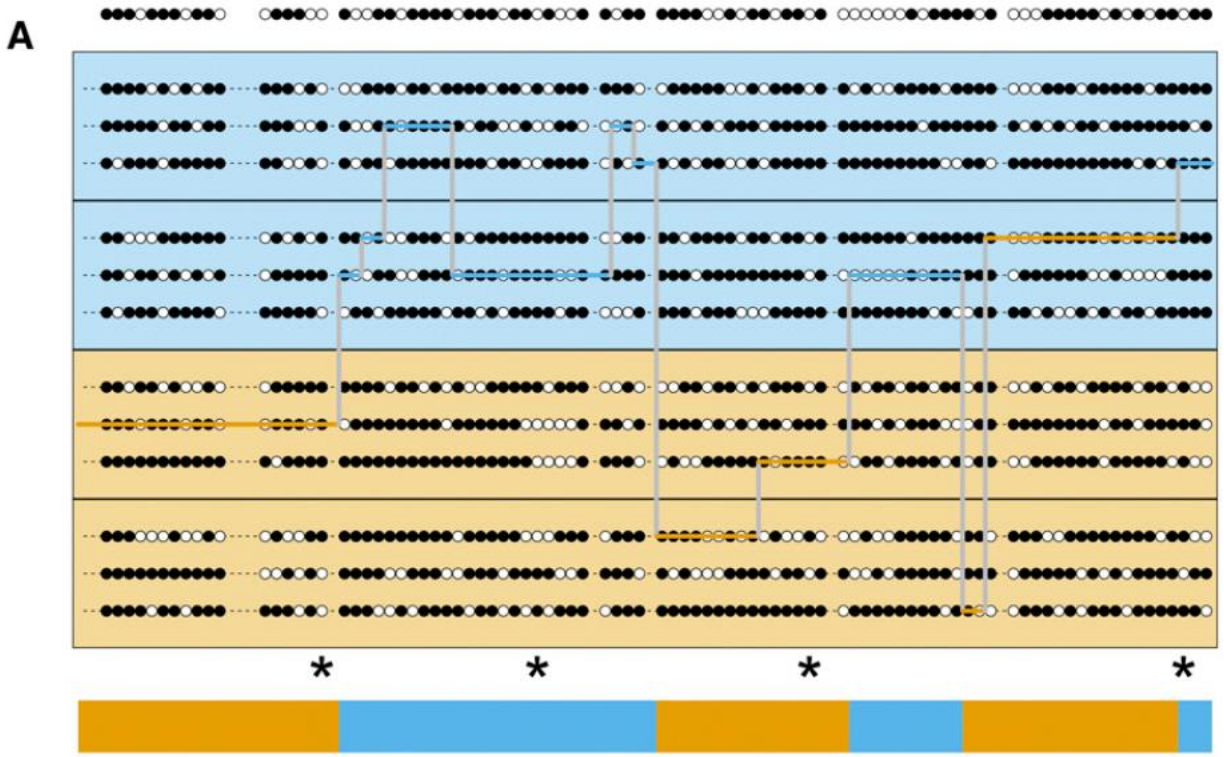
And now presenting . . . MOSAIC

Highlight(s)

- *Does not require knowledge of relationship between labelled surrogates and ancestral source groups*
- Incorporates many ideas from aforementioned papers: exploiting LD info, iteratively updating haplotype phase, etc.
- Provides same or more outputs: local ancestry estimates, fitted coancestry curves, etc.

Another hidden Markov model

> “The key difference is that our method builds these relationships directly into our HMM, which uncovers accurate local ancestry estimates along the genome, whereas GLOBETROTTER fits a mixture model to the output of an ancestry unaware HMM.”



The algorithm

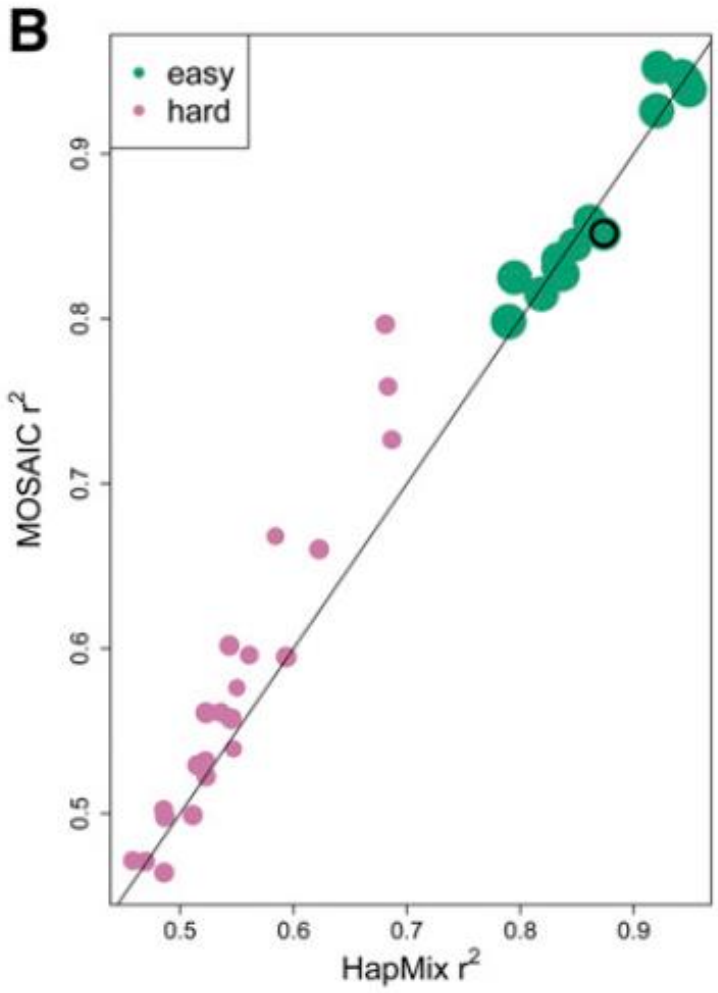
> Uses grid points to speed up computation

1. Initialization
2. Phase convergence
   i. Thinning
      - Restricting to < 100 donors
   ii. Rephasing (phase-hunting)
      - Marginally updating phase based on likelihood
   iii. 10 Baum-Welch EM iterations
3. Final run
4. Dating admixture events & calculating statistics

Outputs

- Expected r^2 for inferred ancestral groups
- F_{st} & R_{st}, treating posterior assignments as "unadmixed partial haploid genomes"
  - Weir and Cockerham (1984)
  - A figure argues that R_{st} can be considered to determine when surrogates are poor
- Dating admixture events by fitting coancestry curves

2-way admixture (simulation study)



MOSAIC outperforms HapMix. This is notable because HapMix is specially designed for studying two-way admixture.

4-way admixture

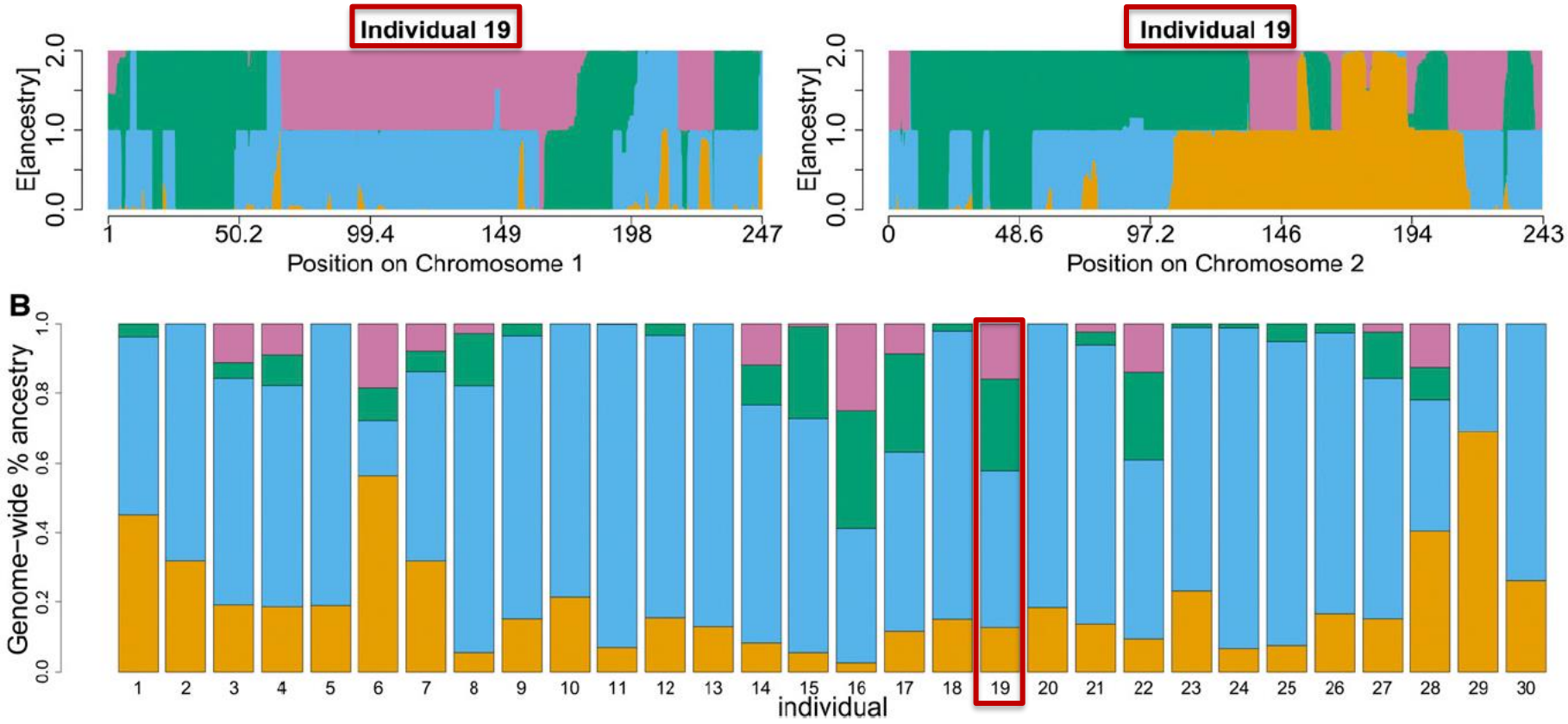


Figure 7 Details of San-Khomani four-way admixture model fit. Each color represents one of four latent ancestries, which in this case correspond to four different ethnicities. The orange source is Bantu-like, blue is San, green is European, and purple is Asian (see Table 5 for details). The colors in these plots are

Selection for African ancestry

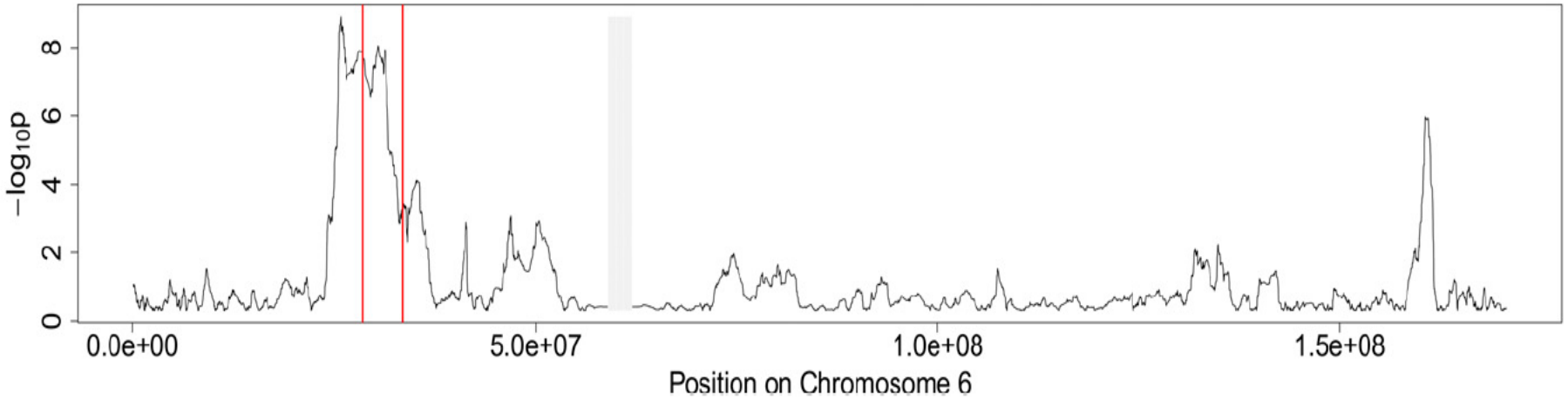


Figure 8 Mean African Ancestry and -$\log_{10}$p of mean ancestry across all 220 individuals in North Africa plotted against: (a) genome position (b) Chromosome 6 position. There is a high and wide spike at the HLA (marked by two vertical red lines) on Chromosome 6 at the HLA. Note that we have blocked out (in light gray) all 1 Mb regions with <10 markers; this includes centromeres with low recombination rates and few SNPs.

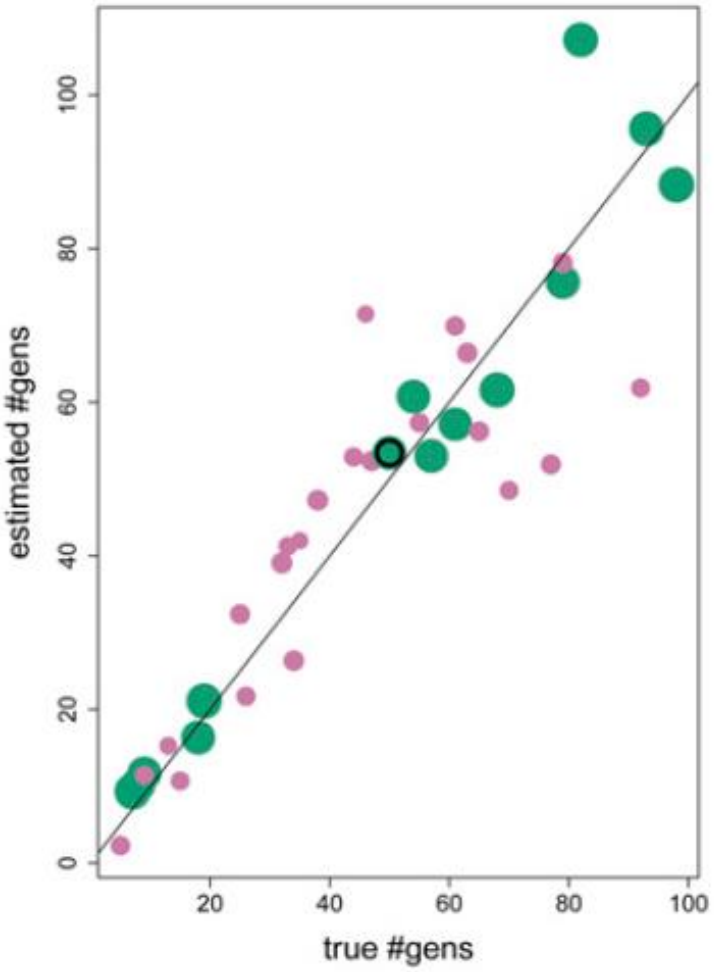
Questions

- How would/could our lab address the problem of local ancestry inference using IBD segments?

- Must we specify # of latent ancestries beforehand?
  - Yes, I believe so.
- Does MOSAIC compute r^2 for > 1 ancestral group?
  - Yes, see tables.
- Is the statistic R_{st} computed each time for MOSAIC? The paper's description of the statistic seems focused on two-way admixture.

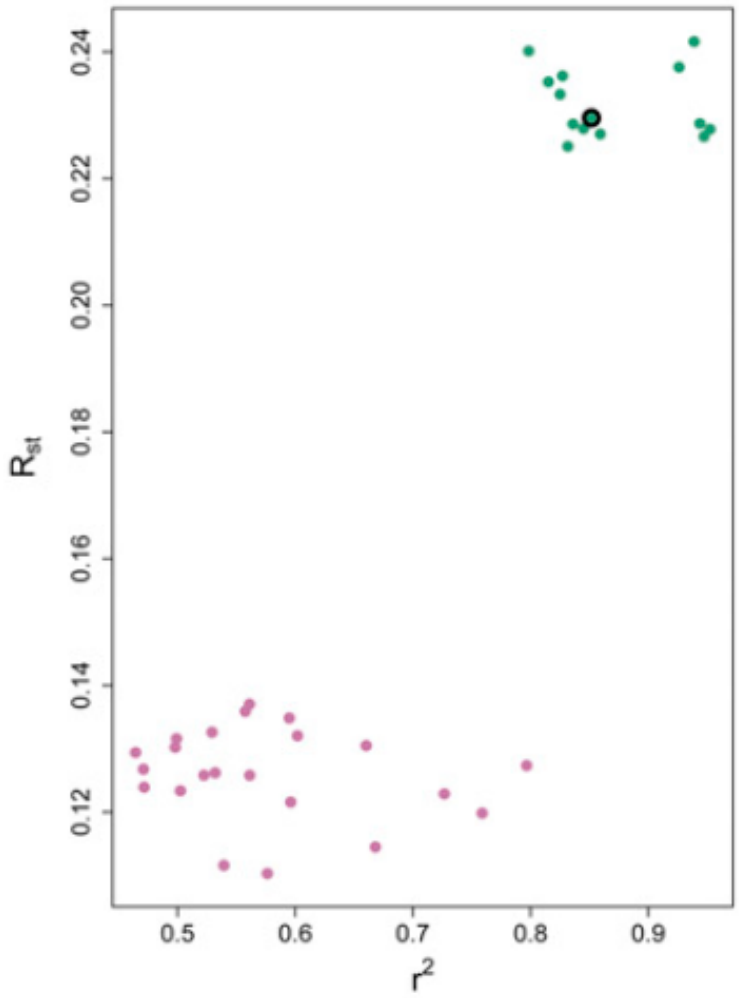
Appendix

2-way admixture (simulation study)



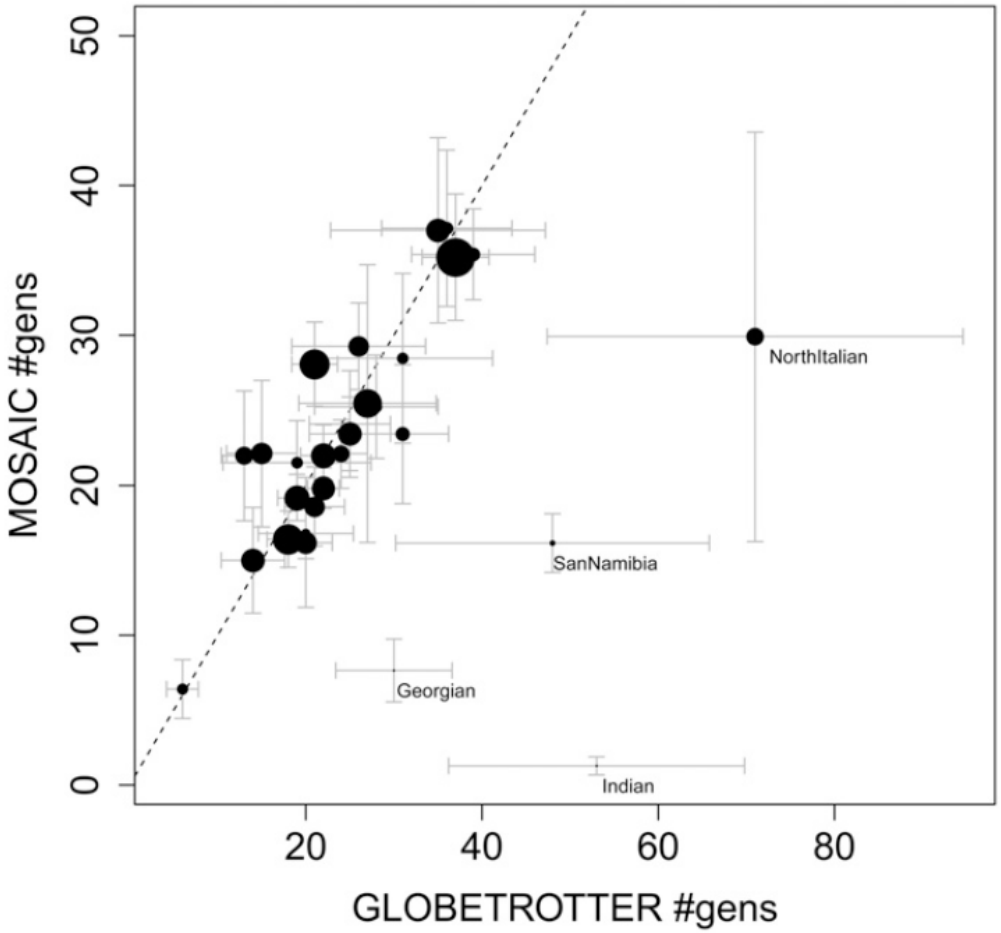
MOSAIC does a good job dating admixture events.

2-way admixture (simulation study)



R_{st} helps identify when poor performance is due to inappropriate surrogates.

Dating admixture events



MOSAIC generally offers tighter bootstrapped intervals. Moreover, MOSAIC has tighter confidence for some very recent admixture.

Transitions

- Notation:
  - $\mathbf{\Pi}^{n}$ denotes a matrix of individual-specific ancestry switches for the latent ancestries
    - e.g. $\mathbf{\Pi}^{n}_{ba}$ for probability of ancestry switch in an individual
    - Does not parameterize non-switches, so rows are not constrained to sum to 1
  - $\mathbf{1} - \mathbf{\Pi}^{n}_{a.}$ denotes a non-switch for an individual
  - $\boldsymbol{\mu}$ denotes copying probabilities for panels given ancestry
    - Columns sum to 1
    - Haplotype selection from within a panel is equiprobable
      - e.g. $\boldsymbol{\mu}_{pa} \div \boldsymbol{N}_{p}$ for copying a haplotype in panel given ancestry

Transitions

Translate these formulas aloud over and over again!

- We provide transition probabilities for (ancestry, haplotype pair within a panel) for (b, h_q) to (a, h_p)
- Think carefully about "redundant" ancestry switch $\boldsymbol{\Pi}^{\boldsymbol{n}}_{\boldsymbol{aa}}$

$$\begin{cases} \boldsymbol{\Pi}^{(n)}_{ba}\dfrac{\boldsymbol{\mu}_{pa}}{N_p} & a \neq b \\ \left(\left(1-\boldsymbol{\Pi}^{(n)}_{a\cdot}\right)\boldsymbol{\rho}+\boldsymbol{\Pi}^{(n)}_{aa}\right)\dfrac{\boldsymbol{\mu}_{pa}}{N_p} & a = b, h_p \neq h_q, \\ \left(\left(1-\boldsymbol{\Pi}^{(n)}_{a\cdot}\right)\boldsymbol{\rho}+\boldsymbol{\Pi}^{(n)}_{aa}\right)\dfrac{\boldsymbol{\mu}_{pa}}{N_p}+\left(1-\boldsymbol{\Pi}^{(n)}_{a\cdot}\right)(1-\boldsymbol{\rho}) & a = b, h_p = h_q, \end{cases}$$

Emissions

We deal with biallelic SNP data (denoted with a $\boldsymbol{Y}$), and we use θ to parameterize the emission probability of a 1 at locus l when copying donor haplotype h as

$$\theta(1 - Y_{lh}) + (1 - \theta)Y_{lh},$$

where $Y_{lh} = 1$ if donor haplotype h has biallelic SNP 1 at locus l, else it is 0. Thus, θ is the probability of a pointwise discrepancy between the allele of the haplotype being locally copied and the allele of the copying haplotype, *i.e.*, the miscopying rate. Note that, for notational simplicity, we have suppressed here the index of the panel from which that haplotype comes.

Fitting coancestry curves

- Based on Hellenthal et al. (2014)
- Estimate chance of ancestry $\mathbf{a}$ at a position and ancestry $\mathbf{b}$ at another position $\mathbf{d}$ away
- This probability is $\approx \boldsymbol{\delta_{ab}} \exp(-\mathbf{d}\boldsymbol{\lambda_{ab}}) + \boldsymbol{\tau_{ab}}$
- Optimize a squared difference of the above
  - Some interpretation of pairwise decay parameters
  - "we do not claim to have made a definitive contribution to the reconstruction of admixture histories based on local ancestry estimation"

References

- Li and Stephens (2003)
- STRUCTURE
- ChromoPainter & FineSTRUCTURE
- GLOBETROTTER
- Latin and Native Americans admixture
- HapMix
- RFMix
- MOSAIC
- HGDP Browser