Estimating **L**ocal **A**ncestry in Ad**m**ixed **P**opulations (LAMP)

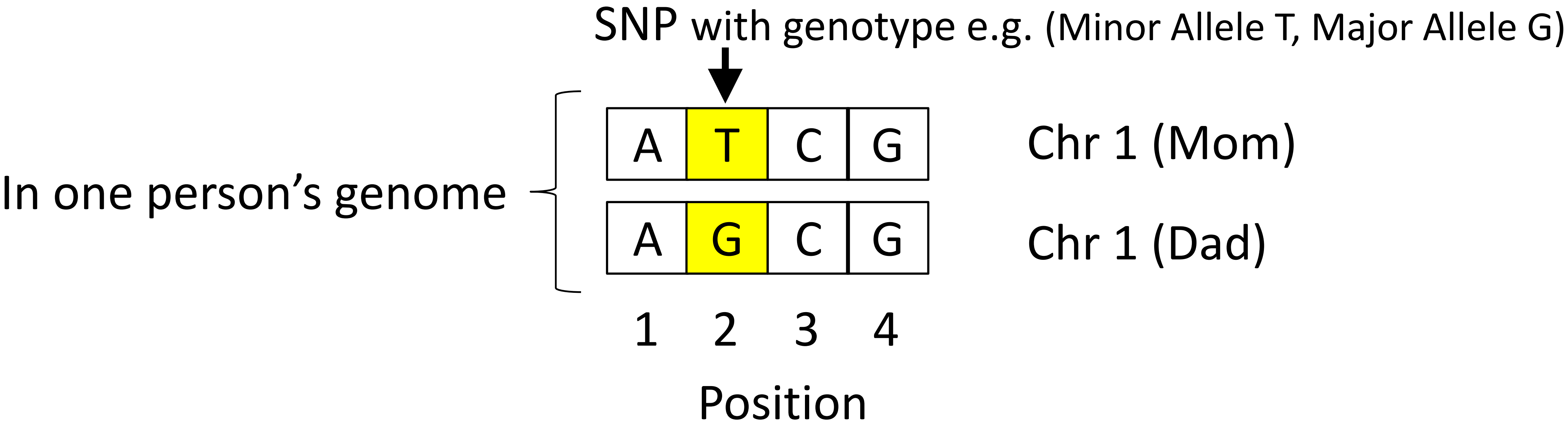
QIAN ZHANG

572 INTRO TALK

Outline

1. **Scientific and Statistical Motivation**
2. How LAMP Works
3. LAMP vs Other Methods
4. What’s Next?

- Human genome = 23 Chromosomes (Chr 1 – Chr 23)
- Each Chr comes as a pair (Mom, Dad)
- Each Chr has positions, where there are variants A, T, C, or G
- SNP: A position where people have different variants
- Most SNPs have 2 variants: Minor Allele, Major Allele
- Genotype = Pair of alleles at a SNP



SNP with genotype e.g. (Minor Allele T, Major Allele G)

In one person's genome

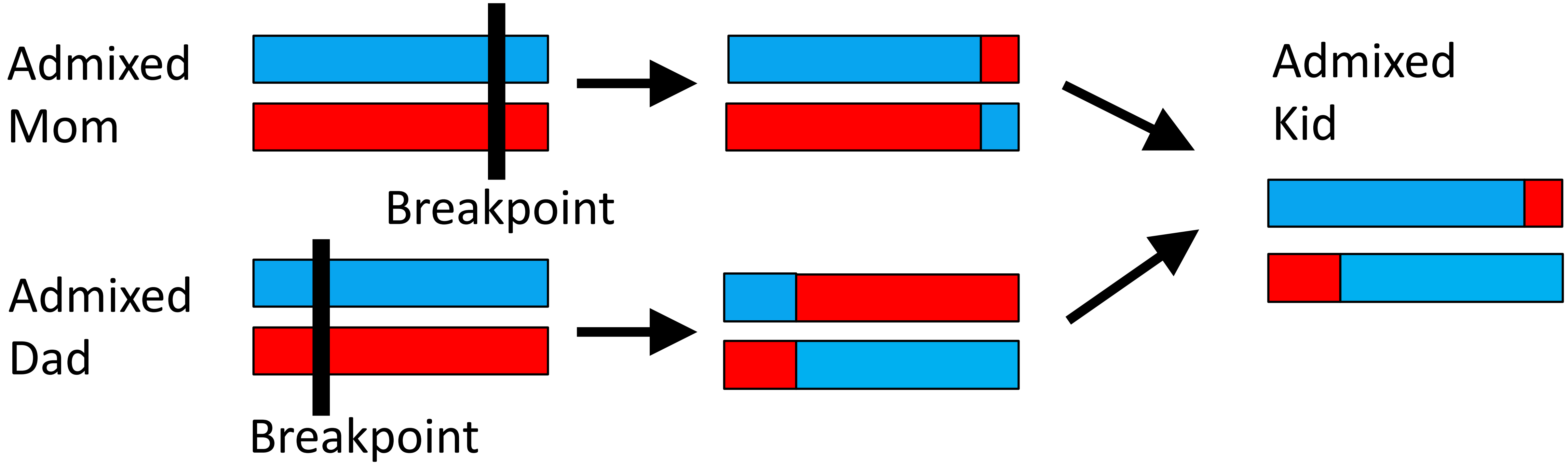
	1	2	3	4
Chr 1 (Mom)	A	T	C	G
Chr 1 (Dad)	A	G	C	G

Position

🟦 = Chromosome from Population 1 (e.g. Europeans)

🟥 = Chromosome from Population 2 (e.g. Africans)

Admixed Person: Person who has DNA from different ancestral populations (e.g. African American)



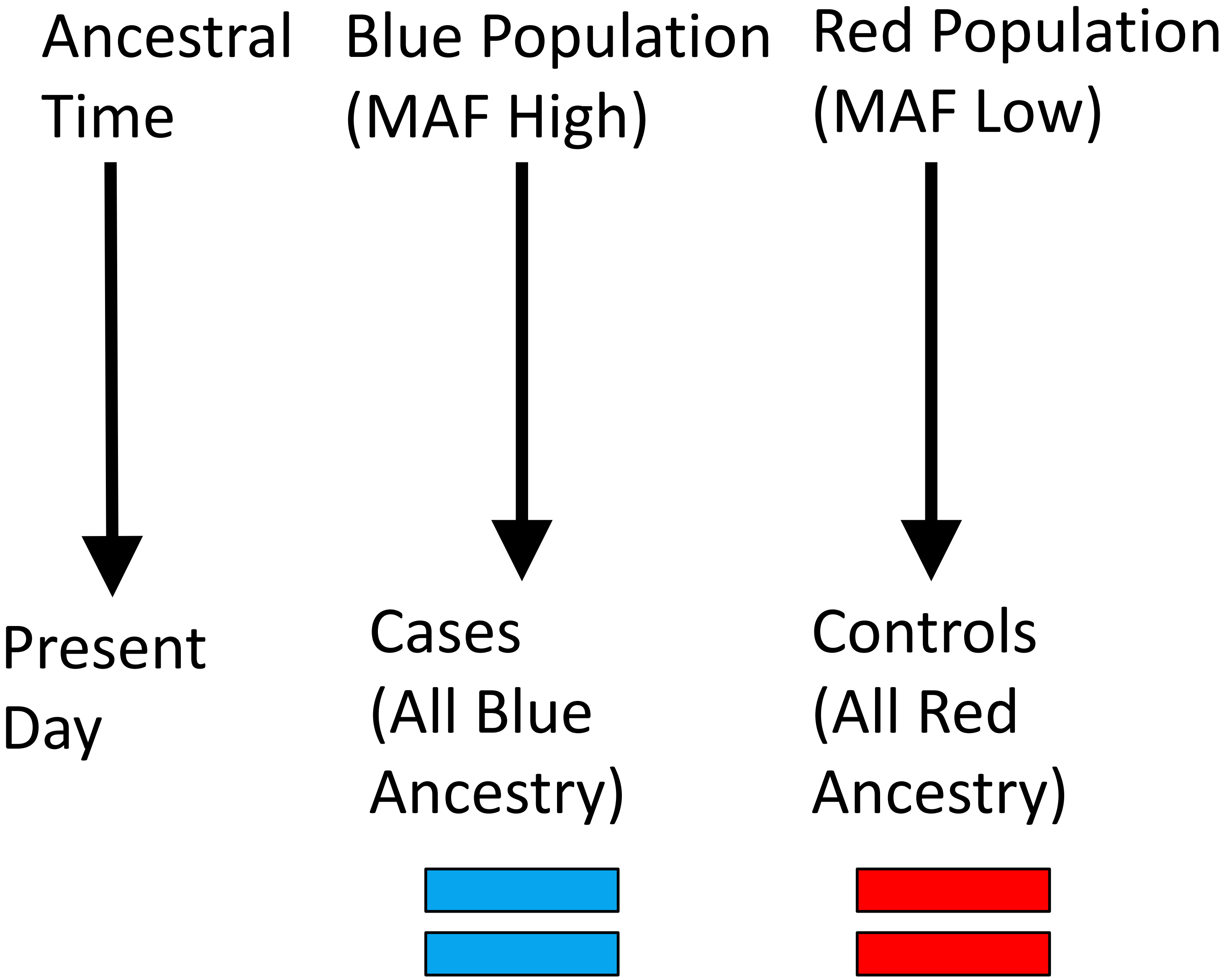
Genome-wide Association Study

- At each SNP, test H0:
  The minor allele is not associated with a disease-related outcome
  (e.g. case/control status, LDL level)

$$\mathrm{logit}(1[\mathrm{case}]) = B_0 + B_1[\#\ \mathrm{copies\ of\ minor\ allele}] + B_2 X + \epsilon$$

- Is the minor allele more frequent in cases than in controls?
- X = Population history is a confounder because

Different Population history ⇒ Different minor allele frequency (MAF)



Ancestral Time

Present Day

Blue Population (MAF High)

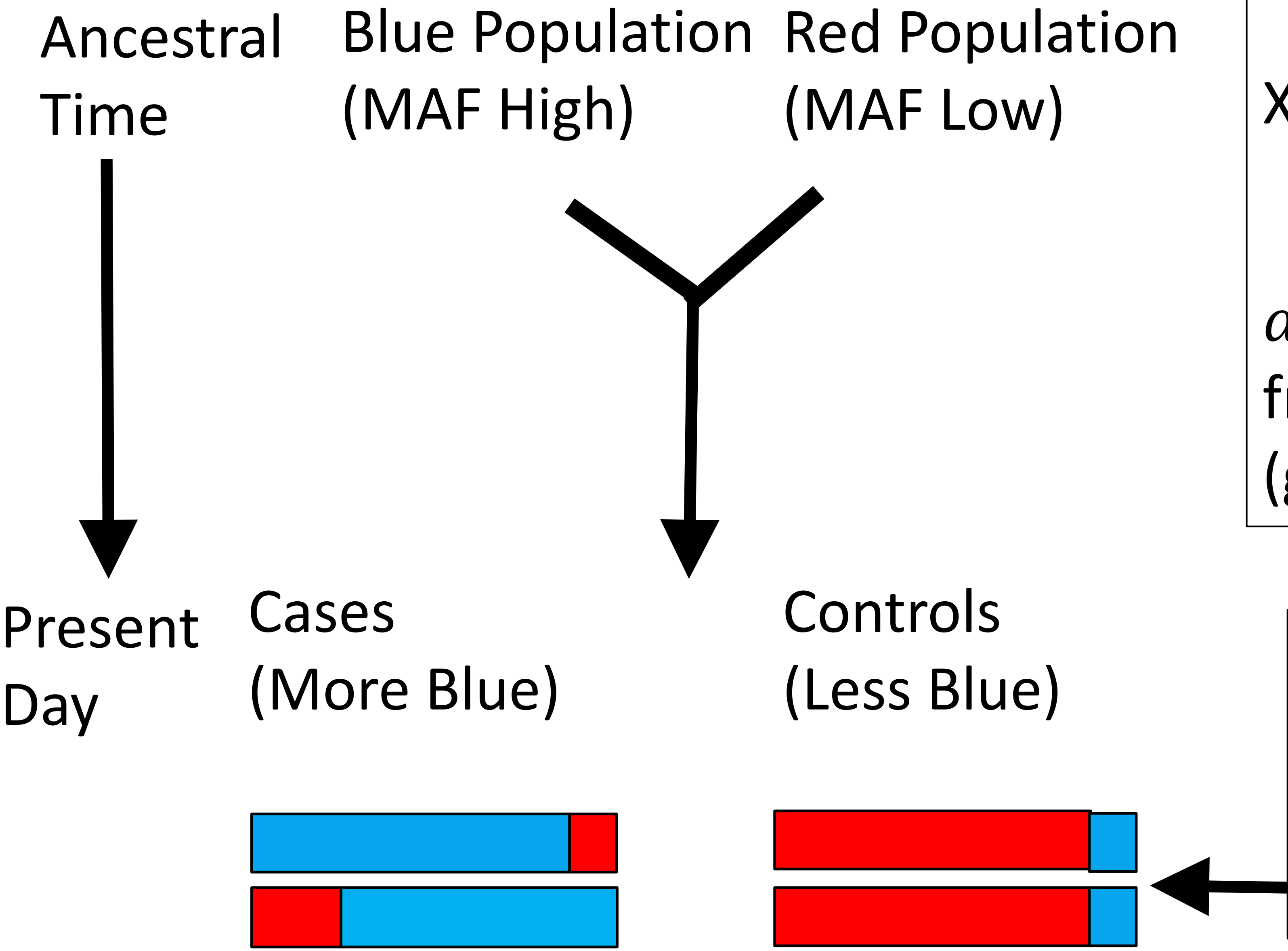
Cases (All Blue Ancestry)

Red Population (MAF Low)

Controls (All Red Ancestry)

(# m of people x 1) vector

$$X = 1[\text{Blue ancestry}] = \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix}$$



Ancestral Time

Blue Population (MAF High)

Red Population (MAF Low)

Present Day

Cases (More Blue)

Controls (Less Blue)

(# m of people x 1) vector

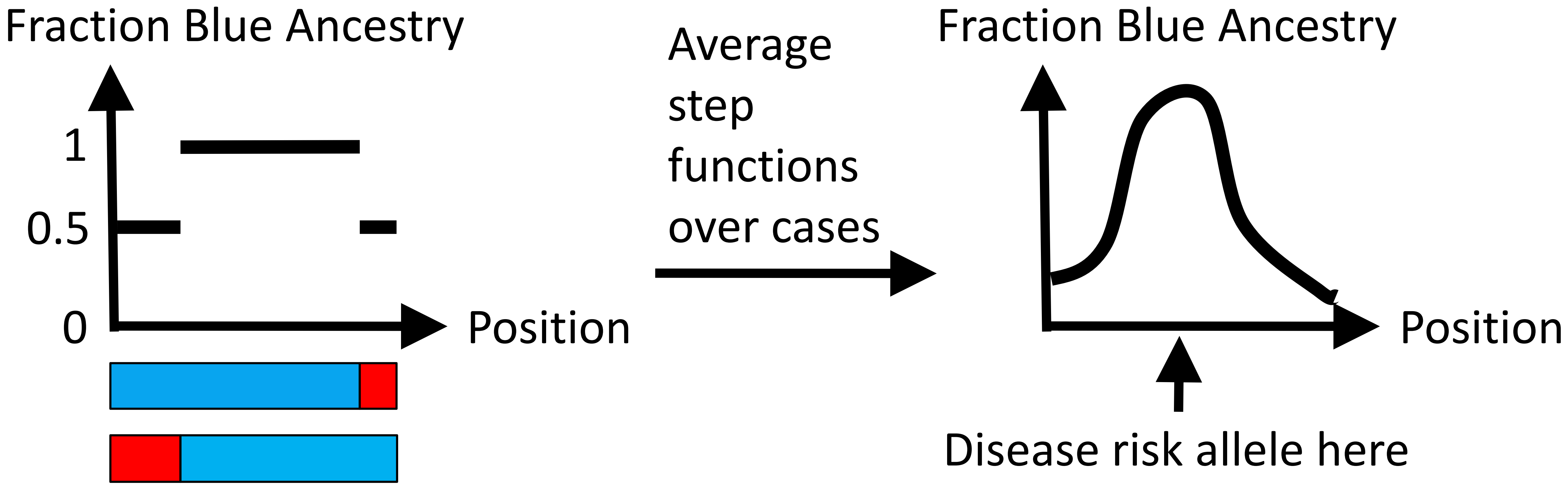
$$X = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

α_i = Fraction of genome from Blue Population (global ancestry)

Local Ancestry Inference: Classifies ancestral origin of each variant at at a position

Mapping by Admixture Linkage Disequilibrium

- Disease-risk allele more frequent in Blue Population than Red Population
- Admixed cases



Average step functions over cases

Disease risk allele here

Summary of Uses of Local Ancestry Inference (the classification of variants to ancestral populations):

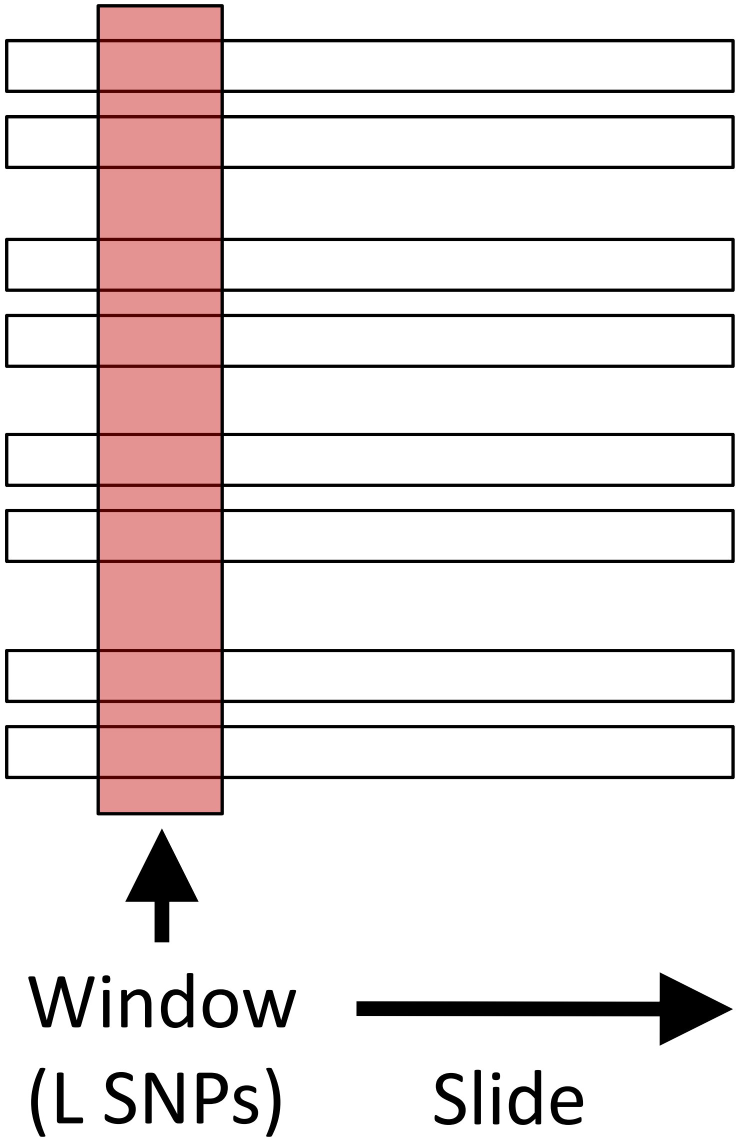
1) Calculate global ancestry → Control for population history in genome-wide association studies

2) Perform mapping by admixture linkage disequilibrium

Outline

1. Scientific and Statistical Motivation
2. **How LAMP Works**
3. LAMP vs Other Methods
4. What’s Next?

LAMP: Overview



0) Remove 1 of each pair of correlated SNPs

→ Work with smaller set of independent SNPs

1) Consider a window of L SNPs:

Assume L so small that nobody has a breakpoint in the window, so in the window, a Chr is all from one ancestry

2) MAXVAR: Get initial ancestry estimates in the window

$$\theta(i) = (\underbrace{\theta_1(i)}_{\text{From Mom}}, \underbrace{\theta_2(i)}_{\text{From Dad}}) \in \{1, \dots, \underbrace{K}_{\text{Number of Ancestral Populations}}\} \times \{1, \dots, K\}$$

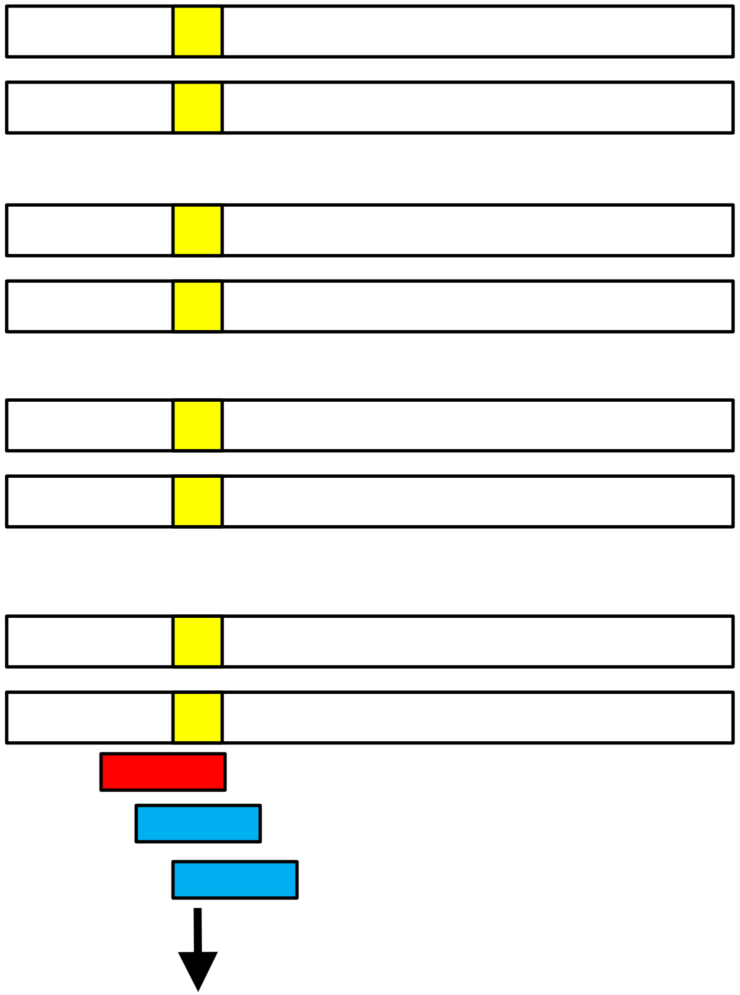
for individuals i = 1, …, m

3) Iterative clustering algorithm: Get final ancestry estimates $\theta(i)$ in the window

4) Slide window over to get a new window overlapping the old one. Repeat 1-3.

5) Majority vote (over windows) to call a SNP’s ancestry

LAMP: Overview



SNP from Blue Population

0) Remove 1 of each pair of correlated SNPs

→ Work with smaller set of independent SNPs

1) Consider a window of L SNPs:

Assume L so small that nobody has a breakpoint in the window, so in the window, a Chr is all from one ancestry

2) MAXVAR: Get initial ancestry estimates in the window

$$\theta(i) = (\underset{\substack{\uparrow \\ \text{From Mom}}}{\theta_1(i)}, \underset{\substack{\uparrow \\ \text{From Dad}}}{\theta_2(i)}) \in \{1, \dots, \underset{\substack{\uparrow \\ \text{Number of Ancestral} \\ \text{Populations}}}{K}\} \text{ x } \{1, \dots, K\}$$

for individuals i = 1, …, m

3) Iterative clustering algorithm: Get final ancestry estimates $\theta(i)$ in the window

4) Slide window over to get a new window overlapping the old one. Repeat 1-3.

5) Majority vote (over windows) to call a SNP’s ancestry

2) MAXVAR: Get initial ancestry estimates in the window

Define a similarity score between individuals i1 and i2

$$S(i_1, i_2) = \frac{\sum_{j=1}^{n}(G_{i1,j} \; - u_j)(G_{i2,j} \; - u_j)}{\sigma_j^2}$$

which is like a sample correlation$(\mathrm{G}_{\mathrm{i}1}, \mathrm{G}_{\mathrm{i}2})$.

$G_{i,j}$ is the Genotype (# of minor alleles) at position j in individual i.

$u_j = \sum_{i=1}^{m} \frac{G_{ij}}{m}$ is the Genotype at position j averaged over m individuals i.

$\sigma_j^2 = \sum_{i=1}^{m} \frac{\left(G_{ij} - u_j\right)^2}{m}$ is the variance in Genotype

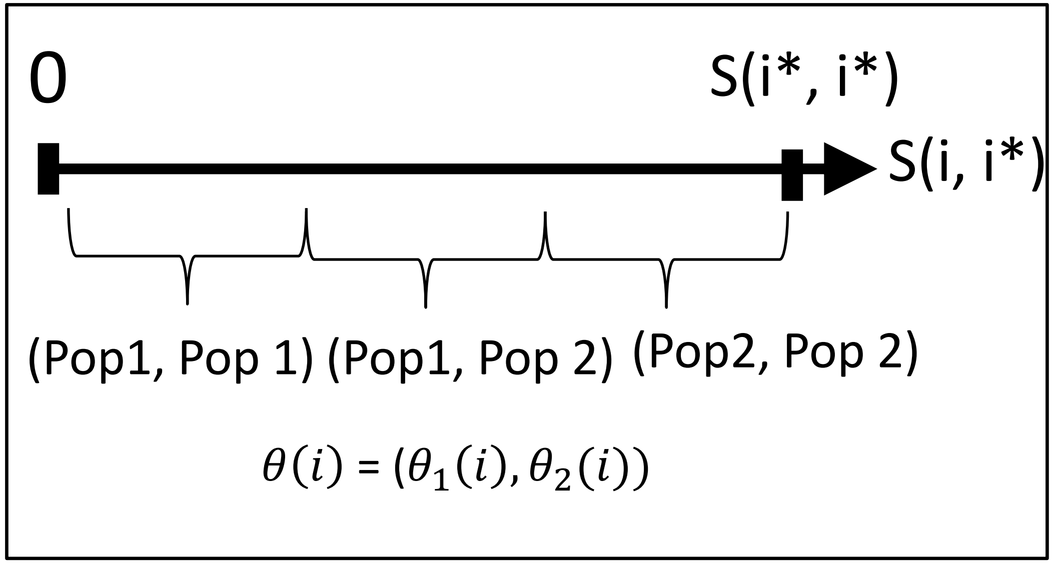
2) MAXVAR: Get initial ancestry estimates in the window

A) For each individual i, Var(i) = $\sum_{\{i':i'\neq i\}} S(i,i')^2$
measures how genetically similar i is to everyone else (the other m - 1 people).

B) Find the person i* who is most genetically similar to everyone else:
i* = $\text{argmax}_i \left(Var(i)\right)$.

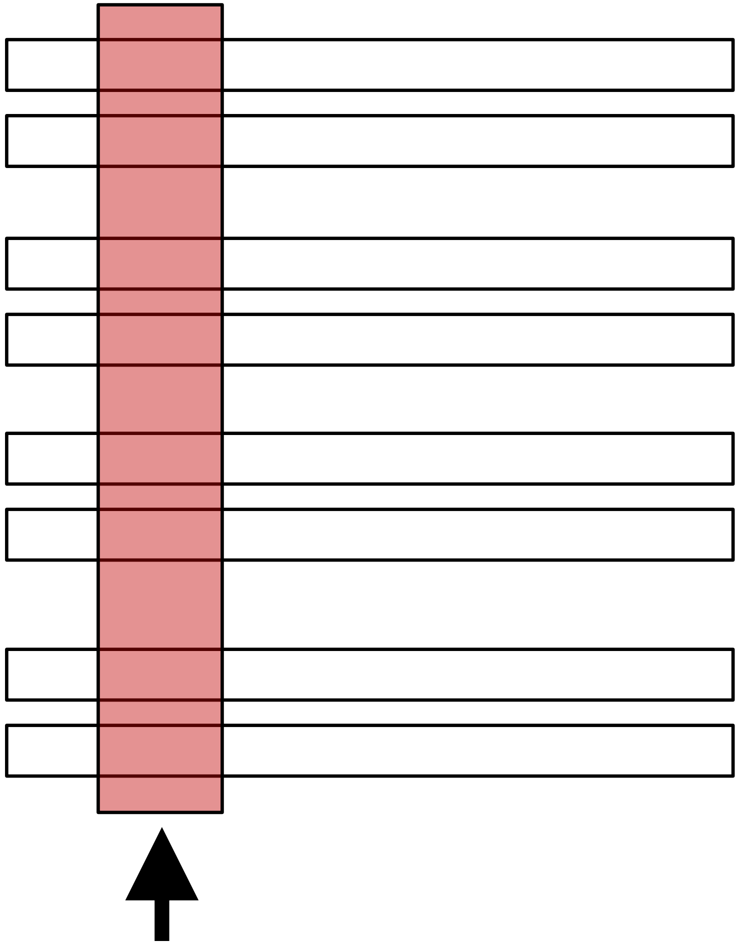
C) Assign initial ancestry estimates of $\theta(i)$ based on how similar i is to i*:

- Assume
  - K = 2 ancestral populations, Pop 1 and Pop 2
  - i* has ancestry (Pop 2, Pop 2)
  - Know α: fraction of variants from Pop 1
- For each individual i, find how similar i is to i*: S(i, i*)
  - Lowest $(1-\alpha)^2 n$ scores $S(i,i^*)$ → (Pop 1, Pop 1)
  - Highest $\alpha^2 n$ scores $S(i,i^*)$ → (Pop 2, Pop 2)
  - Everyone else → (Pop 1, Pop 2)



$\theta(i) = (\theta_1(i), \theta_2(i))$

3) Iterative clustering algorithm: Get final ancestry estimates $\theta(i)$



Window
(L SNPs)

Goal: For each individual i, finally assign to each Chr in the window an ancestry

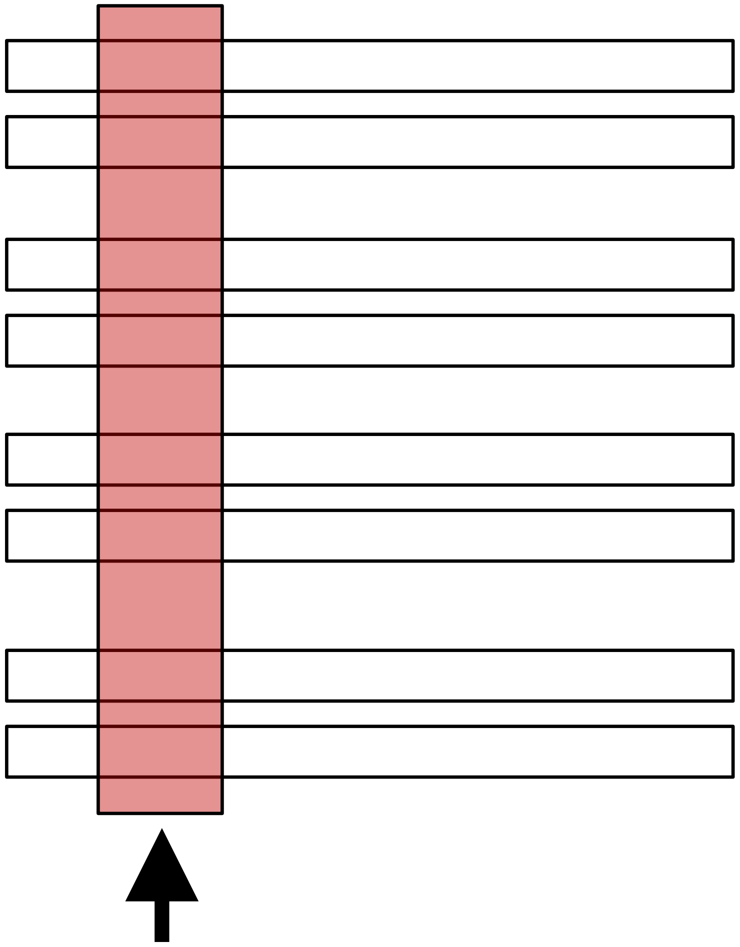
$$\theta(i) = (\theta_1(i), \theta_2(i)) \in \{1, \dots, K\} \text{ x } \{1, \dots, K\}$$

From Mom

From Dad

Number of Ancestral Populations

3) Iterative clustering algorithm: Get final ancestry estimates $\theta(i)$



Window (L SNPs)

If know:

- MAFs $\vec{f_1}, \dots, \vec{f_K}$ of K ancestral populations
- Genotypes $G_1, \dots, G_m$ of m individuals

Then estimate individual i's ancestries

$\theta(i) = (\theta_1(i), \theta_2(i)) \in \{1, \dots, K\} \text{ x } \{1, \dots, K\}$ in the window as

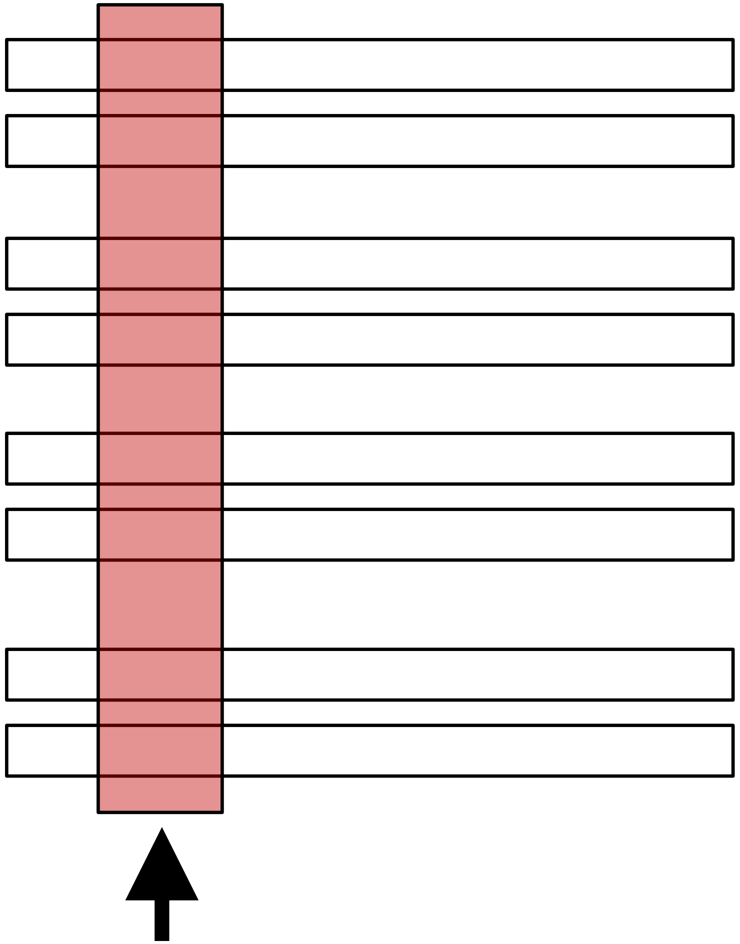
From Mom

From Dad

Number of Ancestral Populations

$$\boxed{\text{argmax}_{\theta(\text{i})}\ \text{P}(\theta(i))\ |\vec{f_1}, \dots, \vec{f_K}, G_i) \qquad (1)}$$

3) Iterative clustering algorithm: Get final ancestry estimates $\theta(i)$



If know:

- Individual i's ancestries
  $\theta(i) = (\theta_1(i), \theta_2(i)) \in \{1, \dots, K\} \text{ x } \{1, \dots, K\}$

  From Mom   From Dad   Number of Ancestral Populations

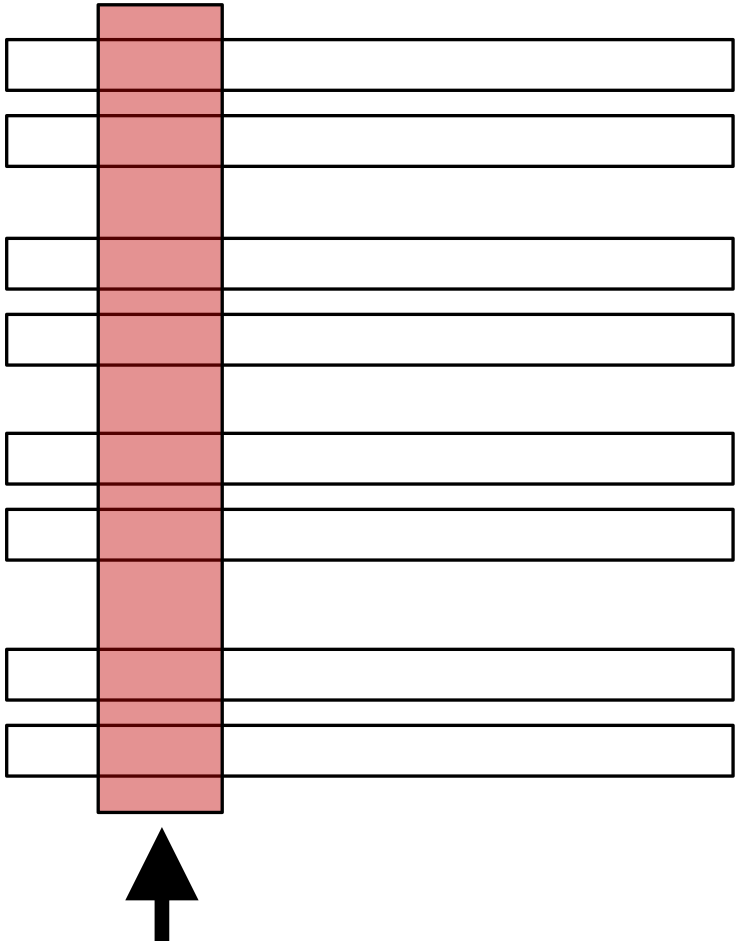
- Genotypes $G_1, \dots, G_m$ of m individuals

Then estimate ancestral MAFs $\overrightarrow{f_1}, \dots, \overrightarrow{f_K}$ as

$$\boxed{\operatorname{argmax}_{\overrightarrow{f_1}, \dots, \overrightarrow{f_K}} \prod_{i=1}^{m} \mathrm{P}(G_i | \overrightarrow{f_1}, \dots, \overrightarrow{f_K}, \theta(i)) \qquad (2)}$$

Start $\theta(i)$ from MAXVAR

3) Iterative clustering algorithm: Get final ancestry estimates $\theta(i)$



Window

(L SNPs)

Iterate equations (1) and (2) to finally estimate individual i's ancestries

$$\theta(i) = (\underset{\substack{\uparrow \\ \text{From Mom}}}{\theta_1(i)}, \underset{\substack{\uparrow \\ \text{From Dad}}}{\theta_2(i)}) \in \{1, \dots, \underset{\substack{\uparrow \\ \text{Number of Ancestral Populations}}}{K}\} \text{ x } \{1, \dots, K\}$$

Outline

1. Scientific and Statistical Motivation
2. How LAMP Works
3. **LAMP vs Other Methods**
4. What’s Next?

Methods that detect global (whole-genome ancestry proportion) or local ancestry (ancestry at particular positions):

- STRUCTURE (> 11,000 citations)
- SABER
- **LAMP (LAMP-ANC)**
- HAPMIX, RFMIX, FAST STRUCTURE, and others!

Methods differ in:

- Estimating global vs local ancestry
- Can handle = 2 or >= 2 ancestral populations
- Input parameters to program

(SABER, LAMP-ANC assume ancestral MAFs known)

- Speed, Accuracy
- Whether they crash when given a lot of data (# SNPs, # people)

Paper in a Nutshell

- Simulate admixed genetic data from ancestral populations (African + European, European + Japanese, Japanese + Chinese)
- Apply STRUCTURE, SABER, LAMP (LAMP-ANC) to simulated data
- LAMP is faster, more accurate than STRUCTURE and SABER except with Japanese + Chinese
- LAMP-ANC is faster, more accurate than STRUCTURE and SABER in all cases
  - Accuracy measured as fraction of variants with correctly classified ancestry
    - Measure makes sense for SABER and LAMP (local ancestry)
    - Measure does not make sense for STRUCTURE (global ancestry)

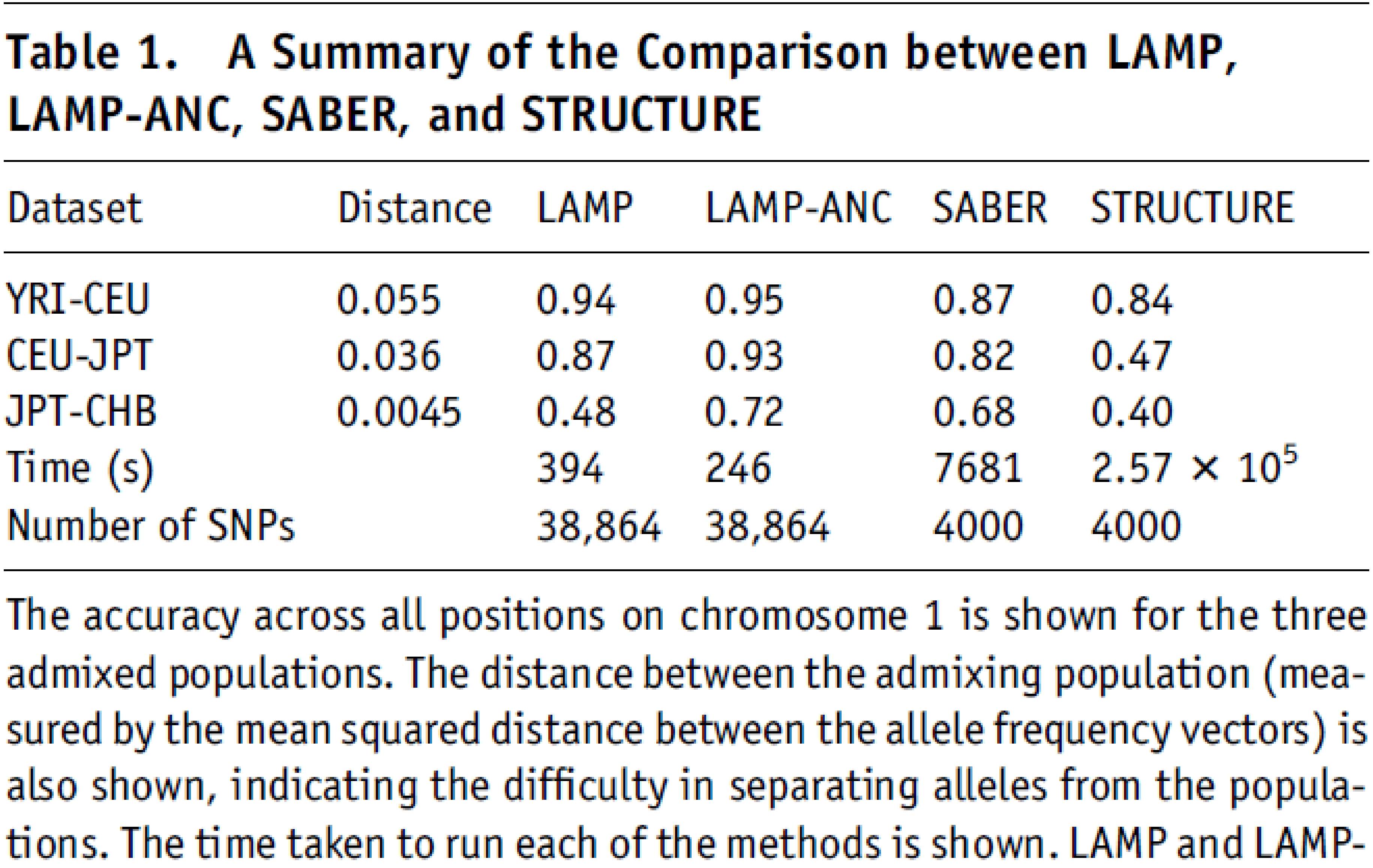


Table 1. A Summary of the Comparison between LAMP, LAMP-ANC, SABER, and STRUCTURE

Dataset	Distance	LAMP	LAMP-ANC	SABER	STRUCTURE
YRI-CEU	0.055	0.94	0.95	0.87	0.84
CEU-JPT	0.036	0.87	0.93	0.82	0.47
JPT-CHB	0.0045	0.48	0.72	0.68	0.40
Time (s)		394	246	7681	2.57 × 10^5
Number of SNPs		38,864	38,864	4000	4000

The accuracy across all positions on chromosome 1 is shown for the three admixed populations. The distance between the admixing population (measured by the mean squared distance between the allele frequency vectors) is also shown, indicating the difficulty in separating alleles from the populations. The time taken to run each of the methods is shown. LAMP and LAMP-

3. LAMP vs Other Methods

Maybe why LAMP is faster than STRUCTURE:

- STRUCTURE uses MCMC (computationally intensive) to generate draws from a posterior distribution
- LAMP does not use MCMC

Expect SABER to be more accurate than LAMP:

- SABER extends a hidden Markov model to account for linkage disequilibrium (correlation between closely positioned SNPs)
- LAMP removes 1 of each pair of highly correlated SNPs,
  uses reduced set of independent SNPs

Maybe why LAMP is more accurate than SABER:

- All the SNPs have only a little more info than the set of independent SNPs
- SABER estimates more parameters, so algorithm might not converge to global optimum

Outline

1. Scientific and Statistical Motivation
2. LAMP (LAMP-ANC) and Others
3. How LAMP Works
4. **What’s Next?**

What’s Next?

- Choice of window length
- Why MAXVAR is a reasonable thing to do
- Simulate data and run 3 methods to compare accuracy

OLGA (Omics in Latinos Genetic Analysis) at UW

will use local ancestry inference.