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

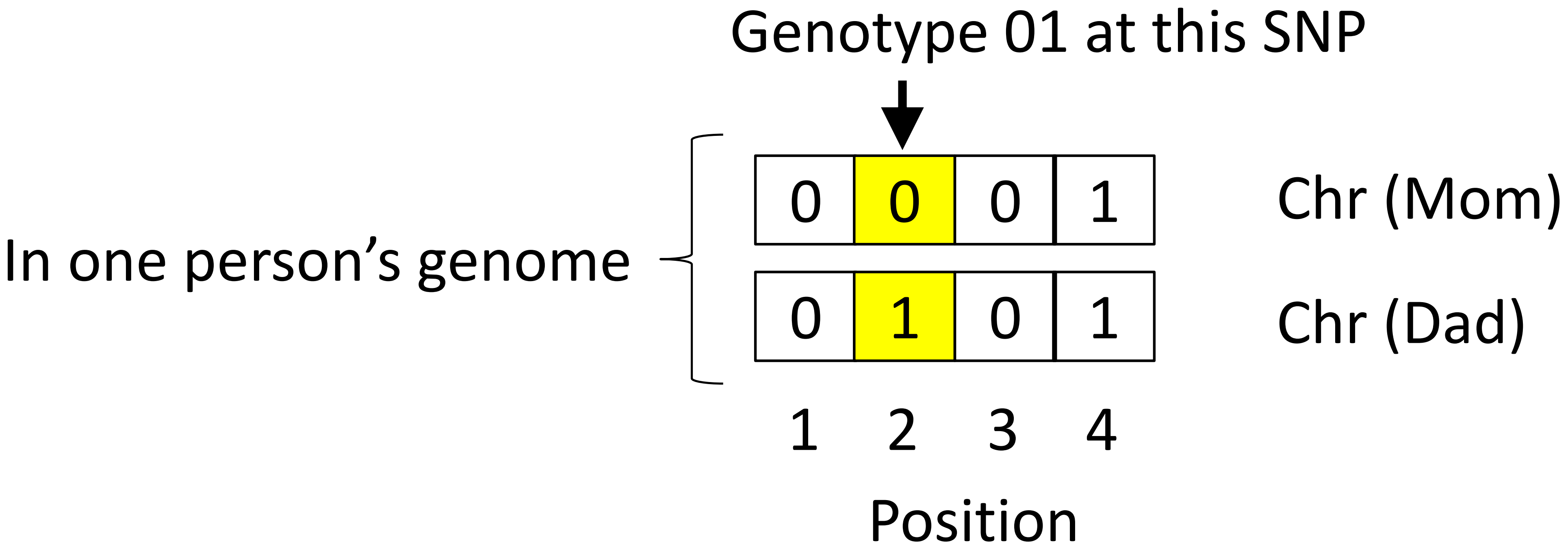
QIAN ZHANG

572 TALK 2

Outline

- **Overview**
- 1) Choose a window length
- 2) MAXVAR: Initial ancestry estimates
- 3) EM within an ICM
- Issues

- SNP: A position where people have different alleles
- Most SNPs have 2 alleles: Minor Allele (0), Major Allele (1)



Genotype 01 at this SNP

In one person's genome

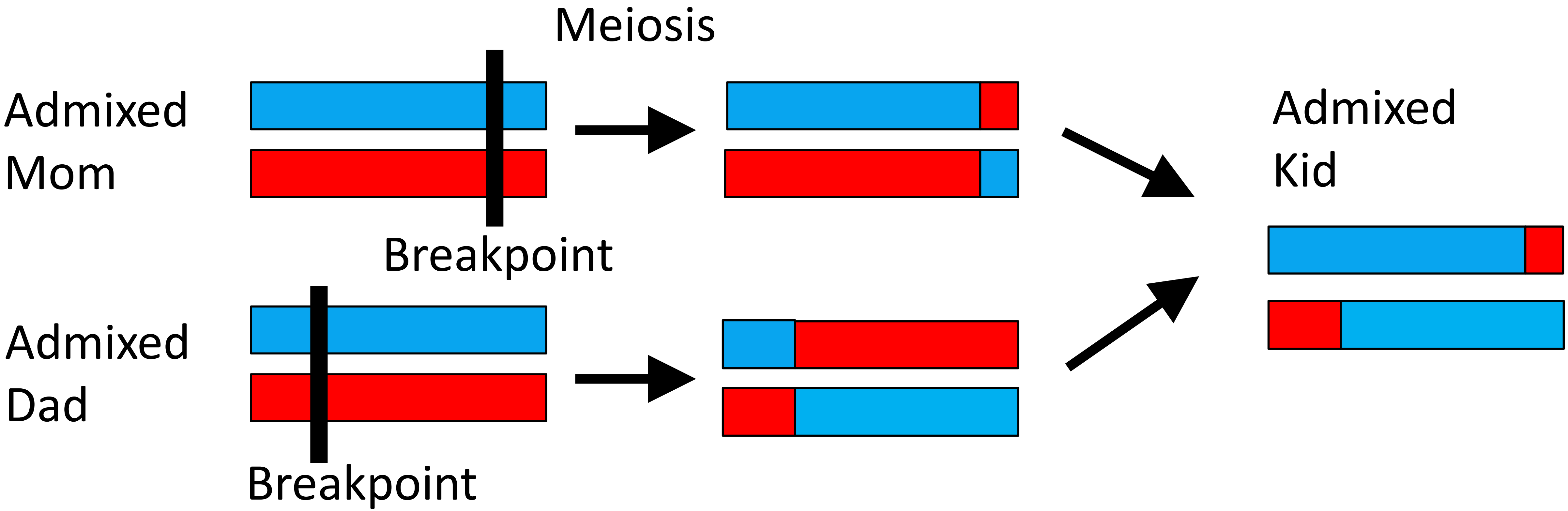
0	0	0	1	Chr (Mom)
0	1	0	1	Chr (Dad)
1	2	3	4	

Position

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

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

Admixed: Having DNA from different ancestral populations.



Meiosis

Admixed Mom

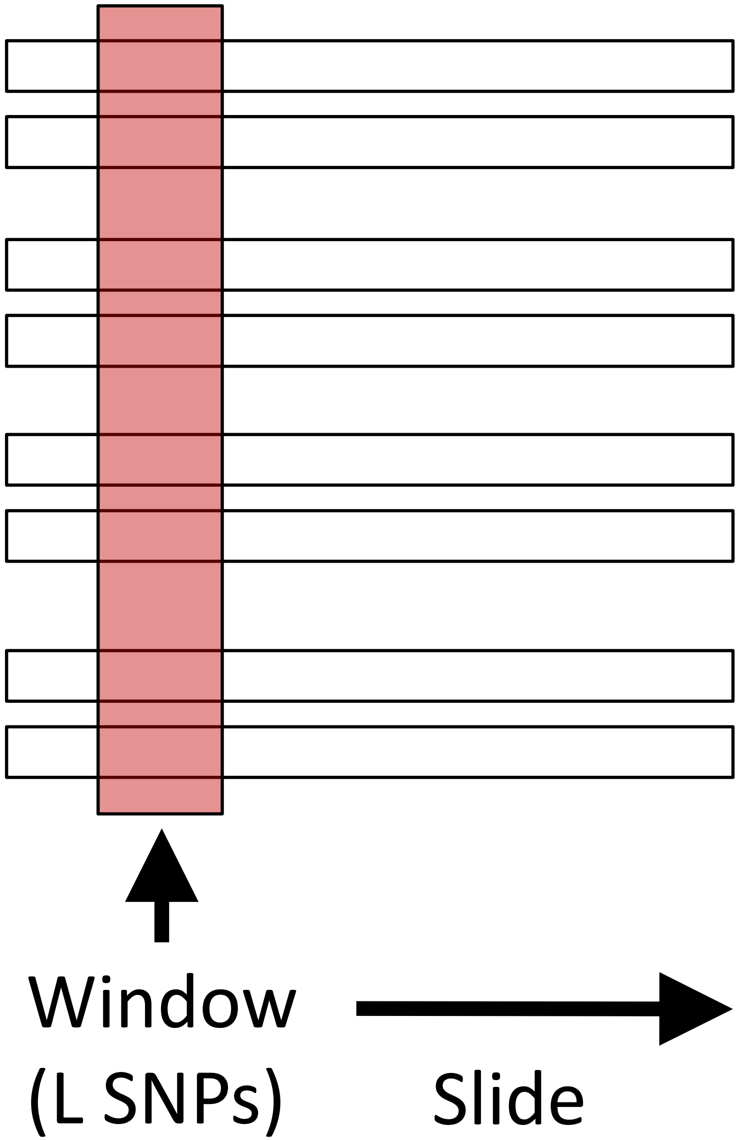
Breakpoint

Admixed Dad

Breakpoint

Admixed Kid

LAMP: Overview



0) Toss some SNPs: Get independent SNPs

1) Window of L SNPs:

- Assume L small, so 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

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

From Mom

From Dad

Number of Ancestral Populations

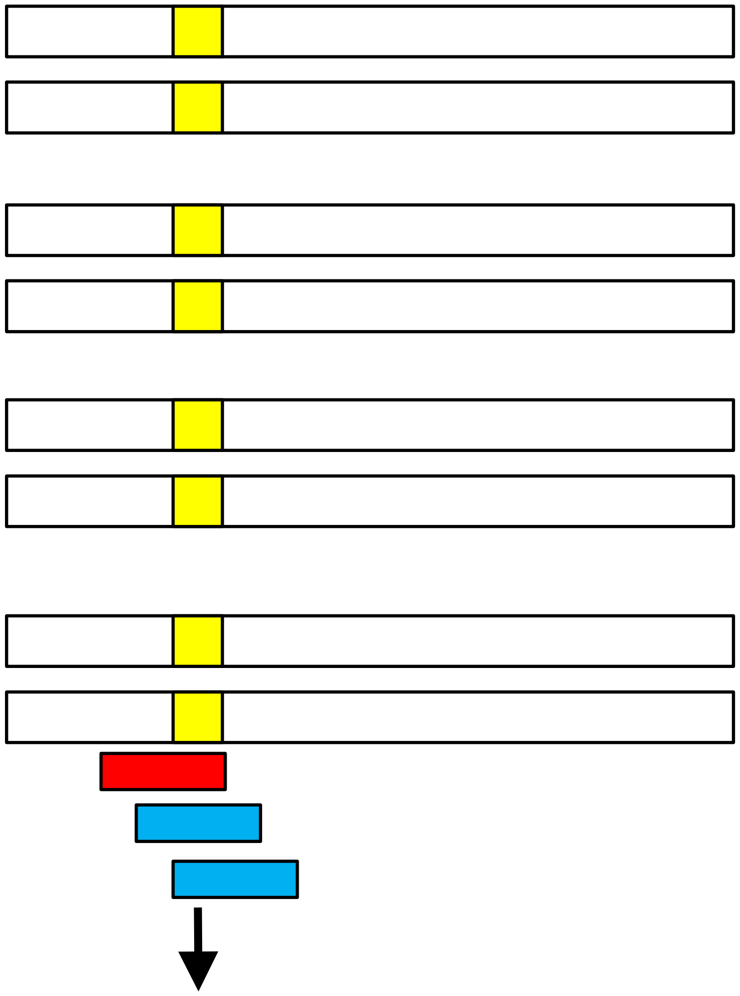
for individuals i = 1, …, m

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 2-3.

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

LAMP: Overview



SNP from Blue Population

0) Toss some SNPs: Get independent SNPs

1) Window of L SNPs:

- Assume L small, so 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

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

↑ From Mom ↑ From Dad ↑ Number of Ancestral Populations

for individuals i = 1, …, m

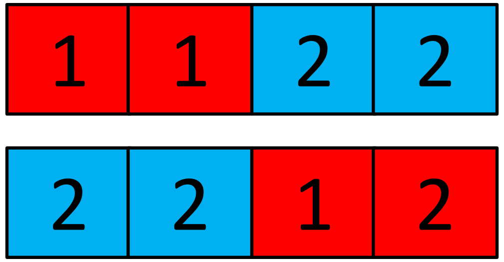
3) Iterative clustering algorithm: Get final ancestry estimates $\theta(i)$ in the window

4) Slide window over to get a new window overlapping the old one. Repeat 2-3.

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

INPUTS

- $K = 2$ populations
- SNP positions (measured)
- Generation $g = 7$ of admixing
- α_i: Frequency of alleles from population i in admixed population (STRUCTURE)



1	1	2	2
2	2	1	2

$$\alpha_1 = \frac{1}{2} \qquad \alpha_2 = \frac{1}{2}$$

- r: Recombination rate

(# of breakpoints per meiosis per generation)

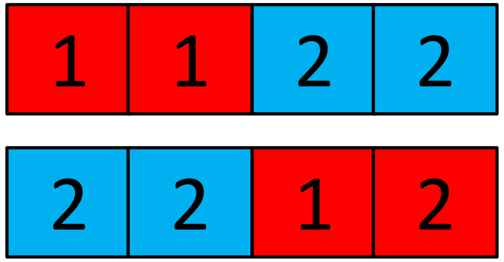
- Offset: Fraction of window's bp length by which to shift the window

- Genotypes Gij (ind i, position j)

0	0	0	1
0	1	0	1

OUTPUT

Ancestries



1	1	2	2
2	2	1	2

Outline

- Overview
- **1) Choose a window length**
- 2) MAXVAR: Initial ancestry estimates
- 3) EM within an ICM
- Issues

- Want window length small:

No one has a breakpoint in the window, so it makes sense to call the entire window of a Chromosome a single ancestry

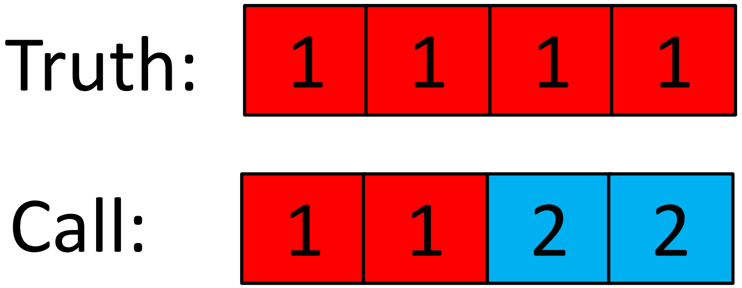
- Want window length big:

Enough SNPs in the window to tell apart ancestries

- Pick the largest L such that

$$L \leq \frac{\epsilon}{(g-1)r\sum_{i<j}\alpha_i\alpha_j}$$

- ϵ is the expected fraction of bp positions in a window of a Chr that have incorrectly called ancestry


Truth: 1 1 1 1

Call: 1 1 2 2

$$\epsilon = \frac{1}{2}$$

Outline

- Overview
- 1) Choose a window length
- **2) MAXVAR: Initial ancestry estimates**
- 3) EM within an ICM
- Issues

Similarity score between individuals i1 and i2

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

which is like a sample correlation$(\mathrm{G}_{\mathrm{i1}}, \mathrm{G}_{\mathrm{i2}})$. Higher score, more similar.

$G_{i,j}$ 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.

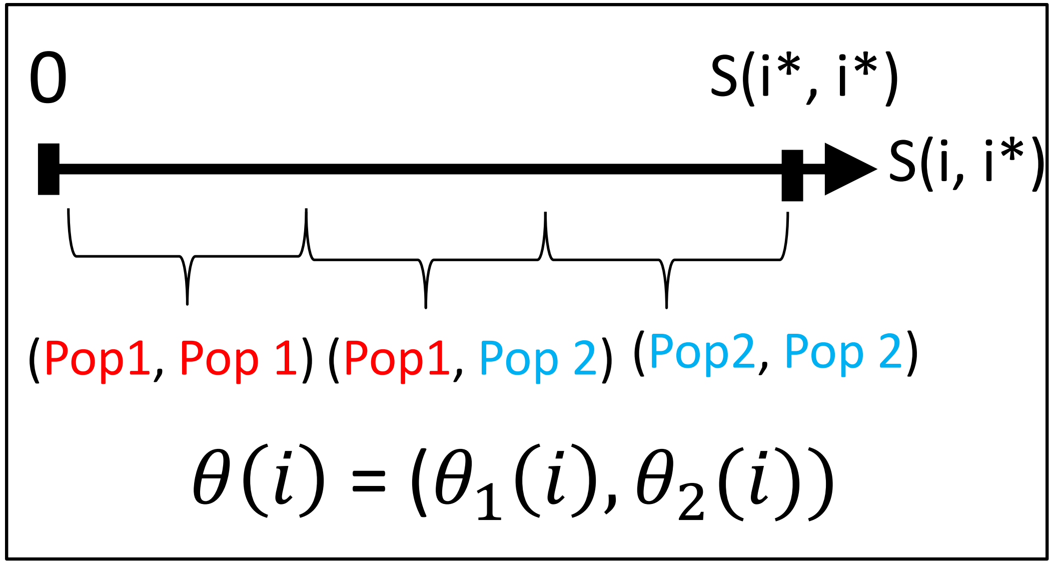
$\sigma_j^2 = \sum_{i=1}^{m} \frac{\left(G_{ij} - u_j\right)^2}{m}$ is the variance in Genotype

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* = $\mathrm{argmax}_{\mathrm{i}} \left(Var(i)\right)$.

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

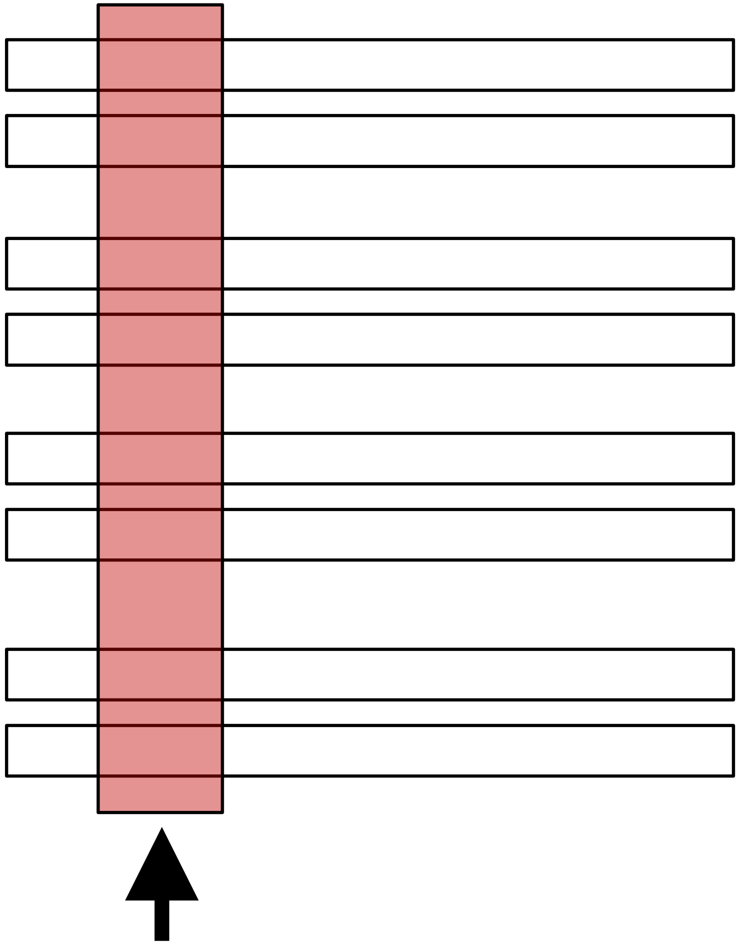
- Assume
    - i* has ancestry (2, 2)
- For each individual i, assign ancestry to i:
    - Lowest $(1-\alpha)^2 n$ scores $S(i,i^*)$ → (1, 1)
    - Highest $\alpha^2 n$ scores $S(i,i^*)$ → (2, 2)
    - Everyone else → (1, 2)



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

Outline

- Overview
- 1) Choose a window length
- 2) MAXVAR: Initial ancestry estimates
- **3) EM within an ICM**
- Issues



Window
(L SNPs)

To finally assign each individual i's chromosome in this window an ancestry

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

From Mom

From Dad

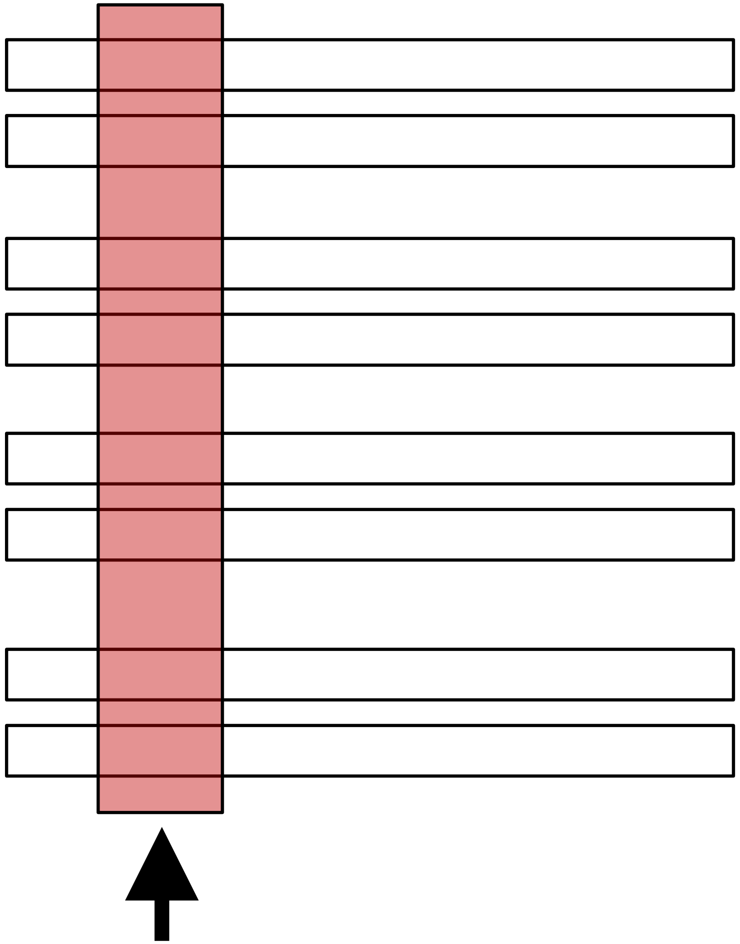
Number of Ancestral Populations

- If data phased:

Iterate equations (1) and (2)

- If data unphased:

Iterate equations (1) and (3), with EM to solve (3)



Window
(L SNPs)

If know:

- MAFs $\overrightarrow{f_1}, \dots, \overrightarrow{f_K}$ $(f_{K1}, \dots, f_{Kn})$ of K ancestral populations
- Genotypes $G_1, \dots, G_m$ of m individuals

Then estimate individual i’s ancestries

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

From Mom

From Dad

Number of Ancestral Populations

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

Posterior mode, not mean, so Iterative Conditional Modes not EM:

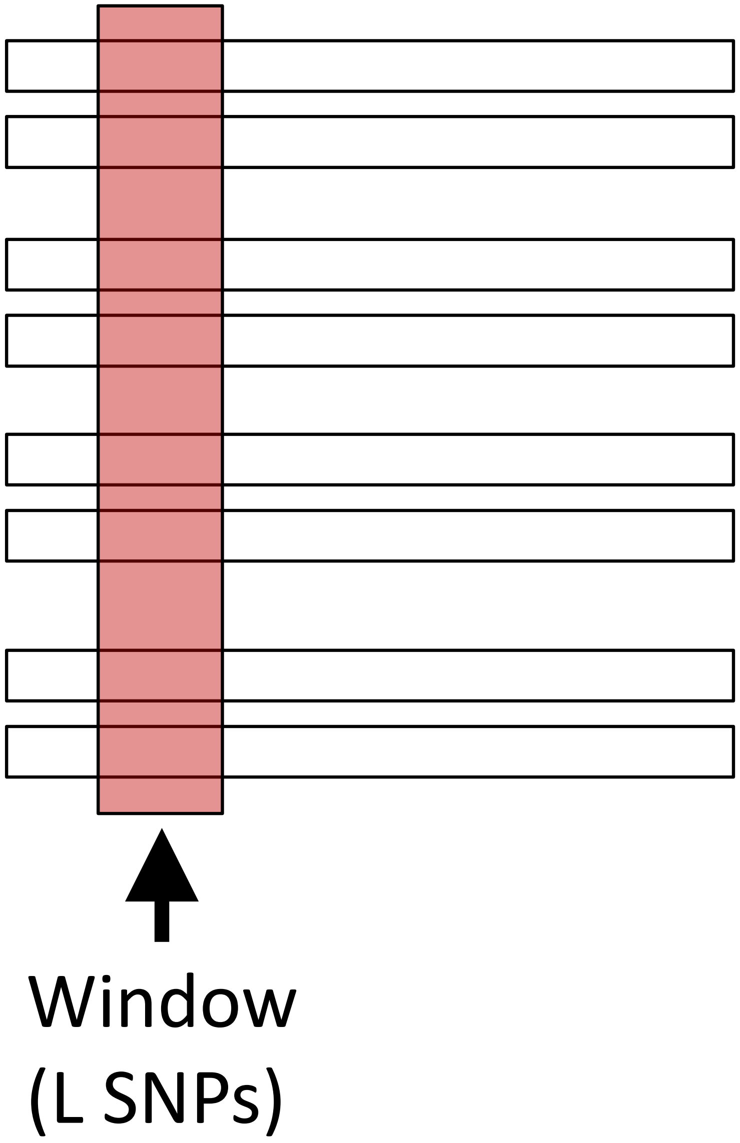
$\widehat{\theta(i)} = \ \mathrm{argmax}_{\theta(\mathrm{i})}\ \mathrm{P}(\theta(i)) \mid \overrightarrow{f_1}, \ldots, \overrightarrow{f_K}, G_i)$ (1), the "E-Step"

= $\mathrm{argmax}_{\theta(\mathrm{i})} \mathrm{P}(\mathrm{G_i} \mid \overrightarrow{f_1}, \ldots, \overrightarrow{f_K}, \theta(i)) * P(\theta(i) \mid \overrightarrow{f_1}, \ldots, \overrightarrow{f_K})$

$$\mathrm{P}(\mathrm{G_i} \mid \overrightarrow{f_1}, \ldots, \overrightarrow{f_K}, \theta(i) = (s,t)) = \prod_{j=1}^{n} (f_{sj} f_{tj})^{1[^{G_{ij}\ =\ 2}]} \prod_{j=1}^{n} ((1 - f_{sj})(1 - f_{tj}))^{1[^{G_{ij}\ =\ 0}]}$$

$$\prod_{j=1}^{n} (\left(f_{tj}\left(1 - f_{sj}\right) + f_{sj}\left(1 - f_{tj}\right)\right)^{1[^{G_{ij}\ =\ 1}]}$$

$$P(\theta(i) = (s,t) \mid \overrightarrow{f_1}, \ldots, \overrightarrow{f_K}) = \alpha_s \alpha_t 1[s = t] + 2\alpha_s \alpha_t 1[s \neq t]$$



If know:

- Individual i's ancestries

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

  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

Phase known:

0	1	1	0	Chr (Mom)
1	0	1	0	Chr (Dad)

If phase known,

MLEs of $\overrightarrow{f_1}, \dots, \overrightarrow{f_K}$ that maximize Equation (2):

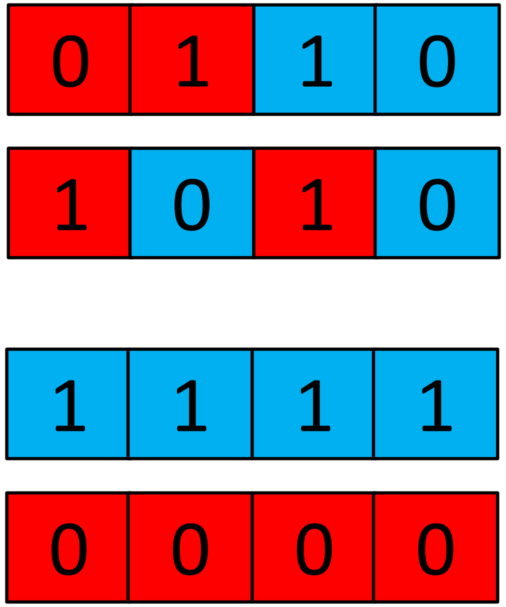
Count number of minor alleles from each population, and divide by number of alleles in the population.

Phase unknown:

1	1	2	0	Genotype

1	1	1	0
0	0	1	0

possible



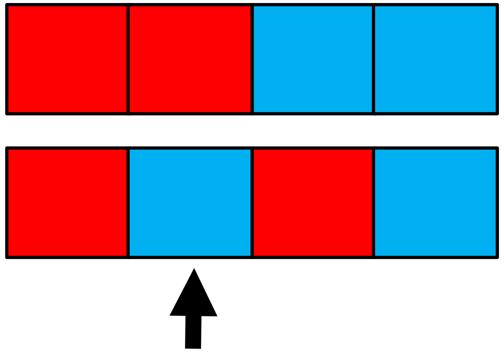
0	1	1	0
1	0	1	0

MLE of $\overrightarrow{f_1} = (\frac{1}{3}, \frac{1}{2}, \frac{1}{2}, 0)$

MLE of $\overrightarrow{f_2} = (1, \frac{1}{2}, 1, \frac{1}{3})$

1	1	1	1
0	0	0	0

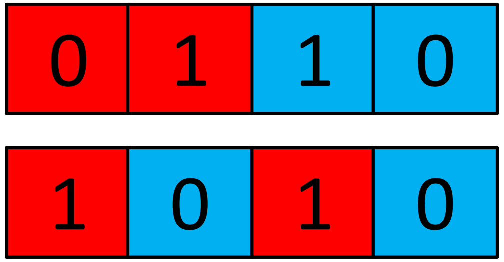
If phase unknown, cannot easily solve equation (2), because see:



and

1	1	2	0

not



0	1	1	0
1	0	1	0

Genotype 0/1 with ancestry 1/2: Count minor allele 0 to Pop 1 or Pop 2?

$\lambda_j(i)$ is a K-long vector with 1 in the s-th entry if the minor allele gets assigned to ancestry s:

$$P\Big(\lambda_j(i) = \vec{e}_{\theta_1(i)} \Big| f_{1j}, \ldots, f_{Kj}, \theta(i)) = f_{\theta_1(i)j}\big(1 - f_{\theta_2(i)j}\big)$$

$$P\Big(\lambda_j(i) = \vec{e}_{\theta_2(i)} \Big| f_{1j}, \ldots, f_{Kj}, \theta(i)) = f_{\theta_2(i)j}\big(1 - f_{\theta_1(i)j}\big)$$

$$P\Big(\lambda_j(i) = s \neq \vec{e}_{\theta_1(i)}, \vec{e}_{\theta_2(i)} \Big| f_{1j}, \ldots, f_{Kj}, \theta(i)) = 0$$

If phase unknown, re-write equation (2) as equation (3), and solve (3) via EM:

$$\text{argmax}_{\overrightarrow{f_1},\ldots,\overrightarrow{f_K}} \prod_{i=1}^{m} \left(\prod_{\text{j amb}} \mathrm{P}(G_i | \overrightarrow{f_1}, \ldots, \overrightarrow{f_K}, \theta(i)) \prod_{\text{j not amb}} \mathrm{P}(G_i | \overrightarrow{f_1}, \ldots, \overrightarrow{f_K}, \theta(i)) \right) \qquad (3)$$

For ambiguous j, $\mathrm{P}(G_i | \overrightarrow{f_1}, \ldots, \overrightarrow{f_K}, \theta(i))$

$=P\left(\lambda_j(i) = \vec{e}_{\theta_1(i)} \,\middle|\, f_{i1}, \ldots, f_{Kj}, \theta(i)\right) + P\left(\lambda_j(i) = \vec{e}_{\theta_2(i)} \middle| f_{i1}, \ldots, f_{Kj}, \theta(i)\right)$

Let $\lambda_{j,s}(i)$ = s-th coordinate of $\lambda_j(i)$, indicating whether the minor allele at position j in individual i gets assigned ancestry s.

Start $\theta(i)$ from MAXVAR

E-Step:

$$\hat{\lambda}_{j,s}(i) = E\left(\lambda_{j,s}(i) \mid f_{\theta_1(i)j}, f_{\theta_2(i)j}, \theta(i), G_{ij} = 1\right) =$$

$$\frac{f_{\theta_1(i)j}\left(1 - f_{\theta_2(i)j}\right)}{f_{\theta_1(i)j}\left(1 - f_{\theta_2(i)j}\right) + \left(1 - f_{\theta_1(i)j}\right) f_{\theta_2(i)j}}, \text{if } s = \theta_1(i)$$

$$\frac{f_{\theta_2(i)j}\left(1 - f_{\theta_1(i)j}\right)}{f_{\theta_1(i)j}\left(1 - f_{\theta_2(i)j}\right) + \left(1 - f_{\theta_1(i)j}\right) f_{\theta_2(i)j}}, \text{if } s = \theta_2(i)$$

M-Step: $$\widehat{f_{sj}} = \frac{2n_{2,2}^{sj} + n_{2,1}^{sj} + n_{1,2}^{sj} + \sum_{j\ ambiguous} \hat{\lambda}_{j,s}(i)}{2n_{2,2}^{sj} + 2n_{2,1}^{sj} + 2n_{2,0}^{sj} + n_{1,2}^{sj} + n_{1,1}^{sj} + n_{1,0}^{sj}}$$

- $n_{k,u}^{sj} = \text{Number of indivudals with } u \in \{0,1,2\} \text{ minor alleles}$ and $k \in \{1,2\}$ copies of alleles from population s at site j

- Iterate EM to get MLEs of $\overrightarrow{f_1}, \dots, \overrightarrow{f_K}$, where $\overrightarrow{f_j} = (f_{1j}, \dots, f_{nj})$ are minor allele frequencies at n snps for ancestral population j

- Plug MLEs of $\overrightarrow{f_1}, \dots, \overrightarrow{f_K}$ into Eq (1)

Outline

- Overview
- 1) Choose a window length
- 2) MAXVAR: Initial ancestry estimates
- 3) EM within an ICM
- **Issues**

- Initialize EM with f’s = ½ in (4)?
- Simulation: What happens if an allele fixes ancestry in the population? How would you get MLEs of (2)? Toss SNPs that fix ancestry in the admixed population?
- MAXVAR: What if σ_j^2 small?
    - Ad hoc: Avoid divide by 0 error (e.g. R, need to use Python) by replacing it with 0.0001
    - Coding everything in Python, not R, to process SNPs quickly