Ancestry Mapping:

Global Ancestry Mapping with Known Populations

By: Daria Wu

Difficulty Level: Easy

5/23/14

Problem Definition

- Given a set of known populations and an individual, how to accurately determine the individual’s ancestry?
- 2 accuracy factors:
  - Overall percentage for each population
  - Distribution of regions pertaining to each population
- 3 populations:
  - CEU (European)
  - JPT+CHB (Asian : Japanese + Chinese)
  - YRI (Yoruban : African)

Motivation

- In studying disease prevalence in different populations, knowing ancestry improves power of studies
- Conversely not knowing an individual’s ancestry could lead to skewed results
- Better understanding of family history

Baseline Approach

- Compare haplotype structure of test individual to haplotype structures of individuals from each of the known populations

- A haplotype is a considered a contiguous group of SNPs

- Inputs: (1) Sets of genotype SNPs for 30 individuals from each of the known populations and (2) genotype SNP data for a test individual

- Output: String of 0’s (CEU),1’s (JPT+CHB), and 2’s (YRI) representing individual’s ancestry

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	1		
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	1		
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA			
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	**FREQUENCY**	**POS 2 HAPLOTYPE**	**FREQUENCY**
GGCCAACCAA	2		
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2		
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	1
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	1
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	1
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	2
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	2
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	2
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	3
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	2	GCCAACCAAG	3
AGCCAACCAA	1		

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA		GCCAACCAAG	
AGCCAACCAA			

Building Frequency Table

Ind 1: GGCCAACCAAGGAAAAAATTAATTAATT

Ind 2: AGCCAACCAAGGAAAAAATTAATTAATT

Ind 3: GGCCAACCAAGGAAAAAATTAATTAATT

POS 1 HAPLOTYPE	FREQUENCY	POS 2 HAPLOTYPE	FREQUENCY
GGCCAACCAA	0.67	GCCAACCAAG	1.0
AGCCAACCAA	0.33		

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value			
Population 1 Lookup Value			
Population 2 Lookup Value			

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value			
Population 1 Lookup Value			
Population 2 Lookup Value			

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3		
Population 1 Lookup Value	0.1		
Population 2 Lookup Value	0.5		

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3		
Population 1 Lookup Value	0.1		
Population 2 Lookup Value	0.5		

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3		
Population 1 Lookup Value	0.1		
Population 2 Lookup Value	0.5		

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3	0.7	
Population 1 Lookup Value	0.1	0.6	
Population 2 Lookup Value	0.5	0.4	

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3	0.7	
Population 1 Lookup Value	0.1	0.6	
Population 2 Lookup Value	0.5	0.4	

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3	0.7	
Population 1 Lookup Value	0.1	0.6	
Population 2 Lookup Value	0.5	0.4	

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	**Position 2**	**Position 3**
Population 0 Lookup Value	0.3	0.7	0.1
Population 1 Lookup Value	0.1	0.6	0.1
Population 2 Lookup Value	0.5	0.4	0.3

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	**Position 2**	**Position 3**
Population 0 Lookup Value	0.3	0.7	0.1
Population 1 Lookup Value	0.1	0.6	0.1
Population 2 Lookup Value	0.5	0.4	0.3

Using Frequency Table to Determine Ancestry

Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3	[✓ 0.7]	0.1
Population 1 Lookup Value	0.1	0.6	0.1
Population 2 Lookup Value	[✓ 0.5]	0.4	[✓ 0.3]

Using Frequency Table to Determine Ancestry

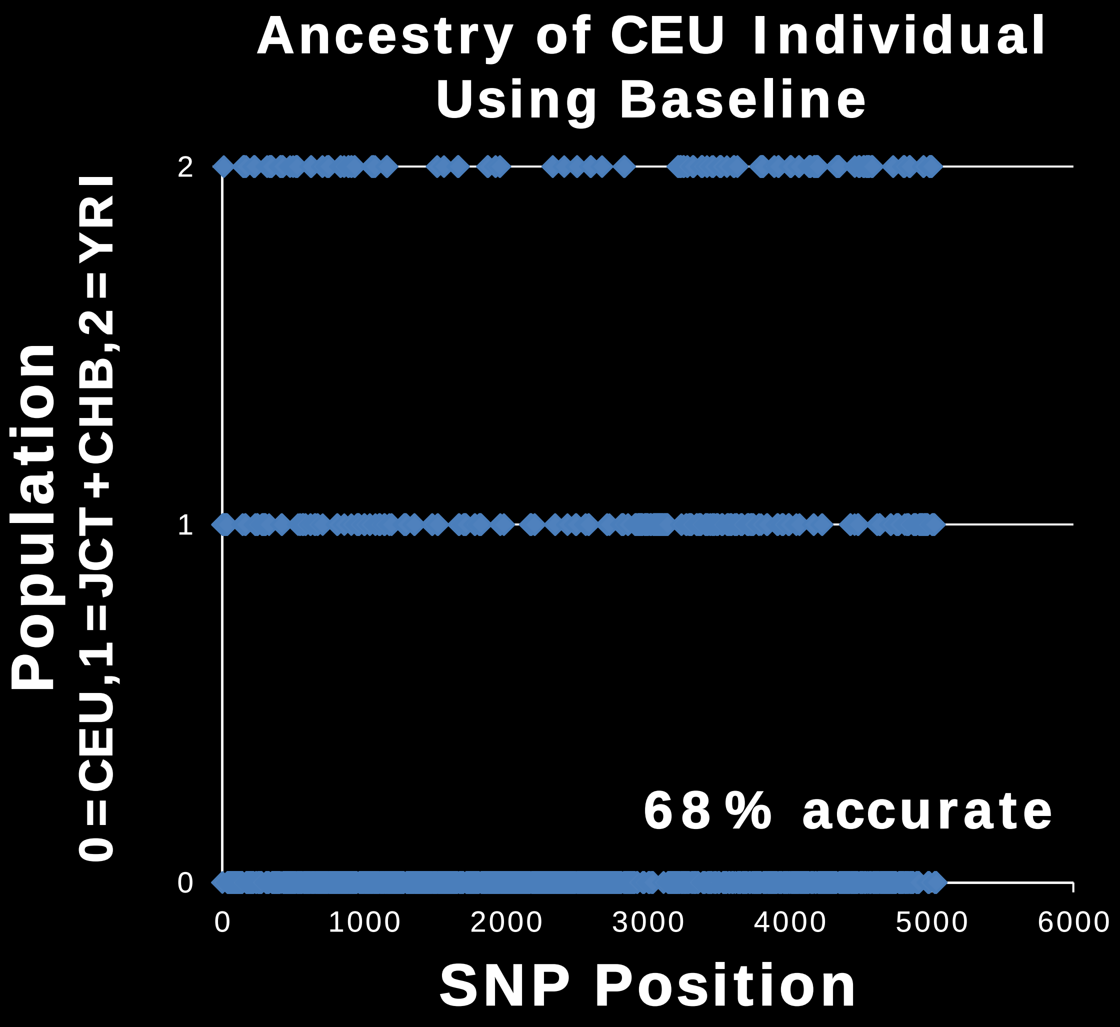
Individual : GGCCAACCAAGGAAAAAATTAATTAATT

	Position 1	Position 2	Position 3
Population 0 Lookup Value	0.3	[✓ 0.7]	0.1
Population 1 Lookup Value	0.1	0.6	0.1
Population 2 Lookup Value	[✓ 0.5]	0.4	[✓ 0.3]

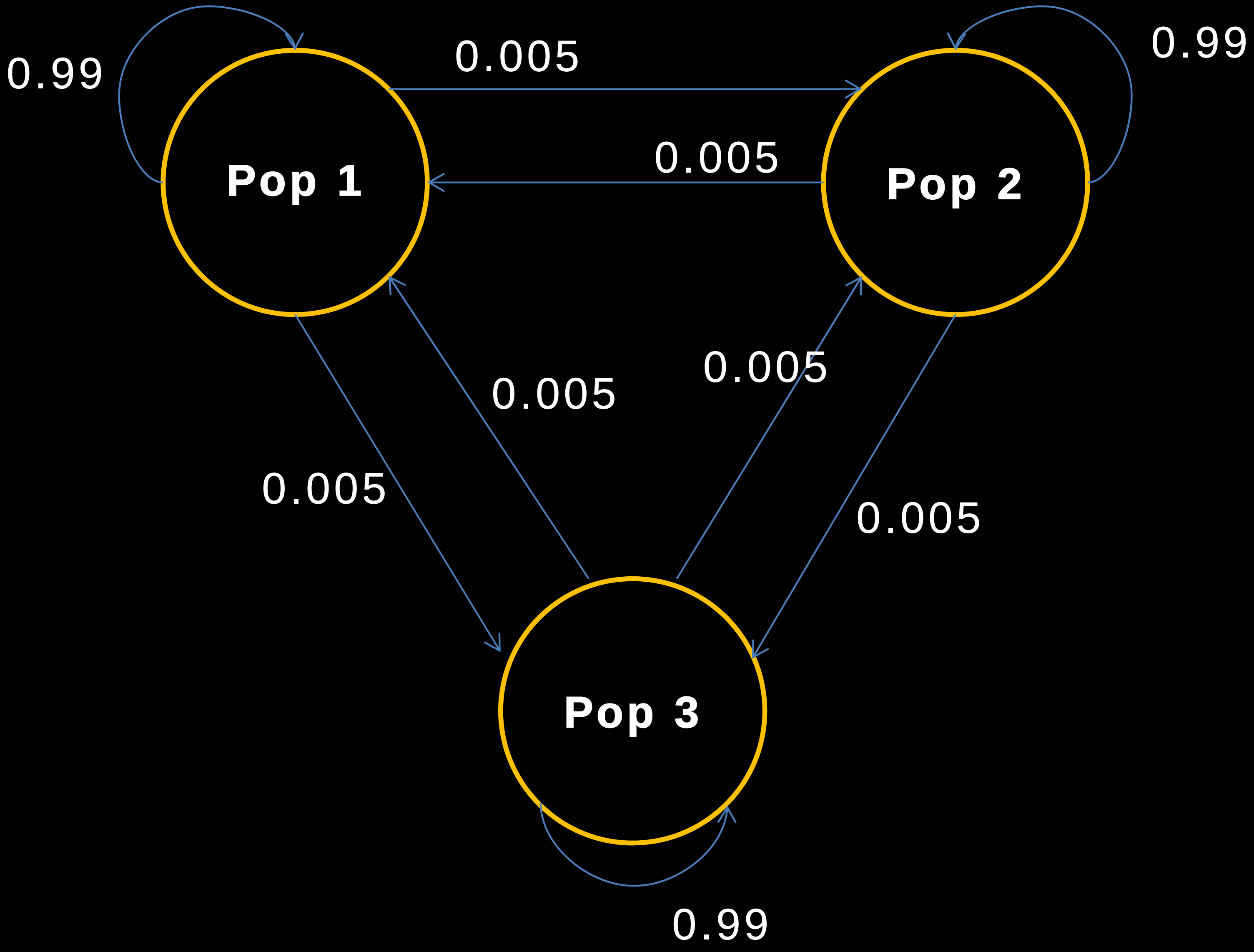
ANCESTRY: 2 0 2

Problems With This Approach

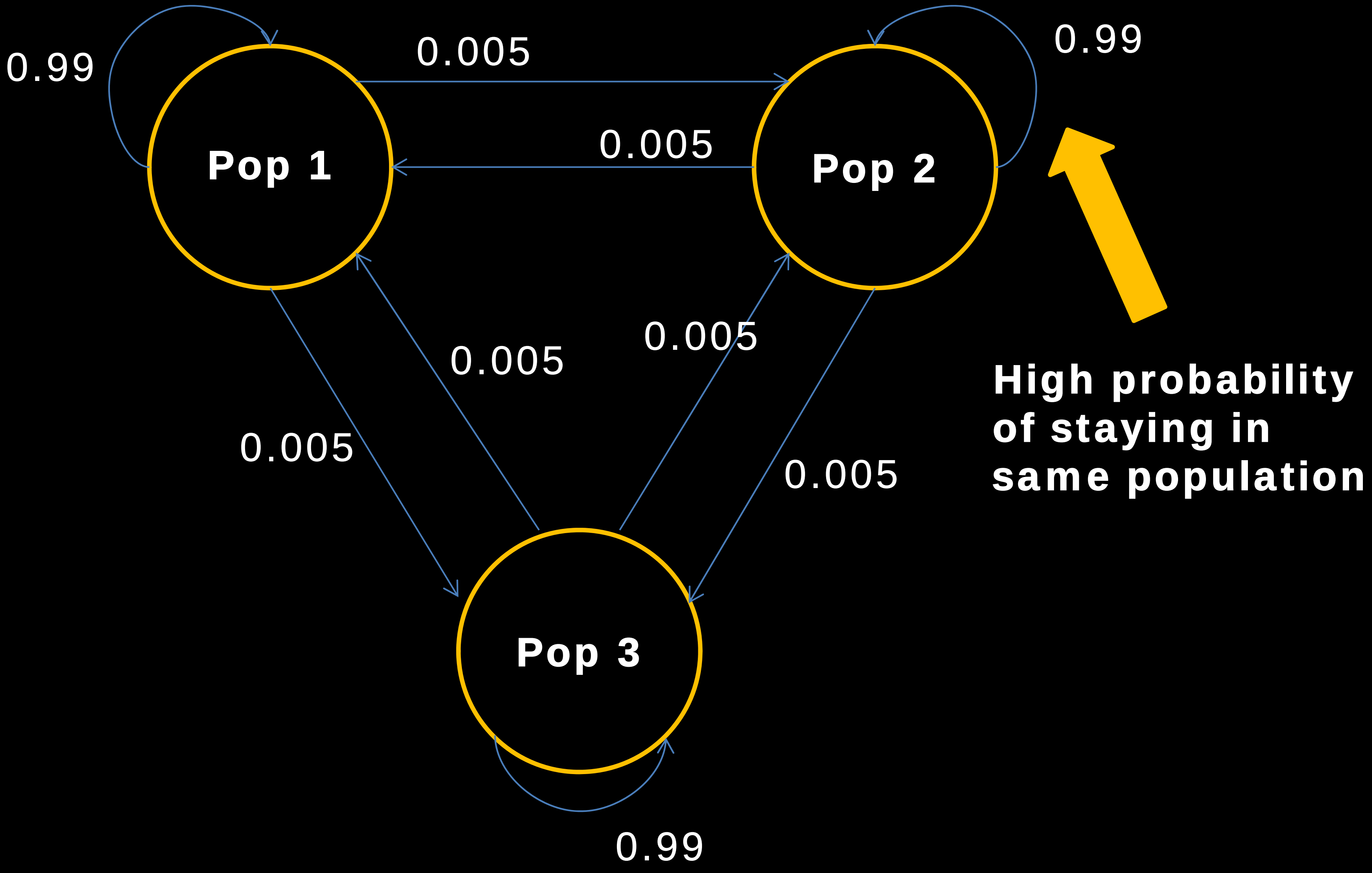
- At every window position a population transition can occur
- This method produces many small, scattered segments from the same population
- In reality individuals are far more likely to have long contiguous segments from the same population



Hidden Markov Model

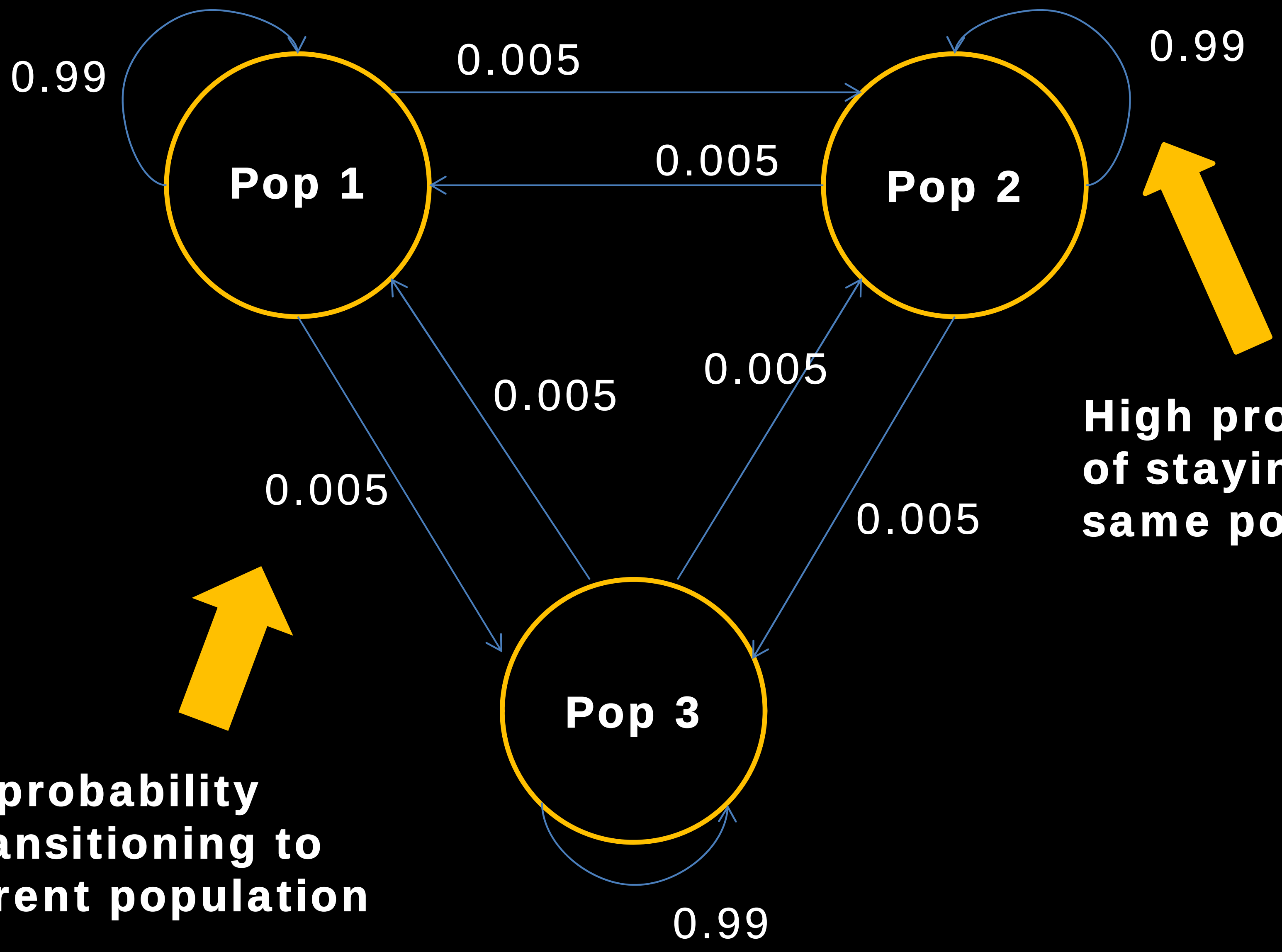


Hidden Markov Model



High probability of staying in same population

Hidden Markov Model



High probability of staying in same population

Low probability of transitioning to different population

The HMM Method

- Determine individual’s ancestry by determining best path through state machine

- At every SNP position the probability of going to the next state depends on the present state

- Same inputs and output as baseline

Building HMM Table

X = max(P(going_to_X|coming_from_state0 and SNP_frequency_in_state0),
P(going_to_X|coming_from_state1 and SNP_frequency_in_state1),
P(going_to_X|coming_from_state2 and SNP_frequency_in_state2)

	Position 1	Position 2
State 0	0.4	X
State 1	0.3	
State 2	0.7	

Building HMM Table

```
X = max( present_state0*transition_prob0*emission_prob0,
         present_state1*transition_prob1*emission_prob1,
         present_state2*transition_prob2*emission_prob2)
```

	Position 1	**Position 2**
State 0	0.4	X
State 1	0.3	
State 2	0.7	

Building HMM Table

```
X = max( 0.4*0.99*0.2,
         0.3*0.005*0.1,
         0.7*0.005*0.6)
```

	Position 1	Position 2
State 0	0.4	X
State 1	0.3	
State 2	0.7	

Building HMM Table

X = max(0.0792, 0.00015, .0021)

	Position 1	**Position 2**
State 0	0.4	X
State 1	0.3	
State 2	0.7	

Building HMM Table

X = max(0.0792, 0.00015, .0021)

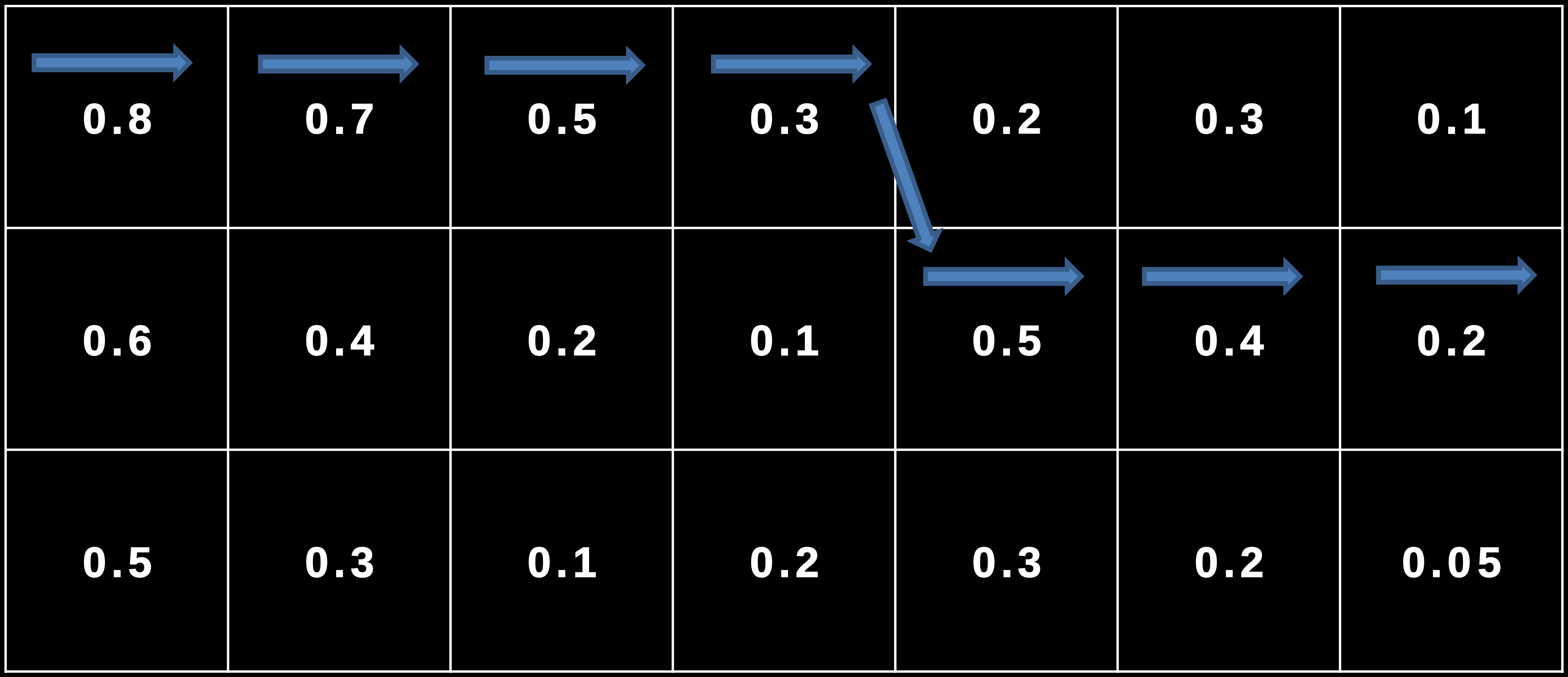
	Position 1	**Position 2**
State 0	0.4	0.0792
State 1	0.3	
State 2	0.7	

Building HMM Table

X = max(0.0792, 0.00015, .0021)

	Position 1	**Position 2**
State 0	0.4	0.0792
State 1	0.3	
State 2	0.7	

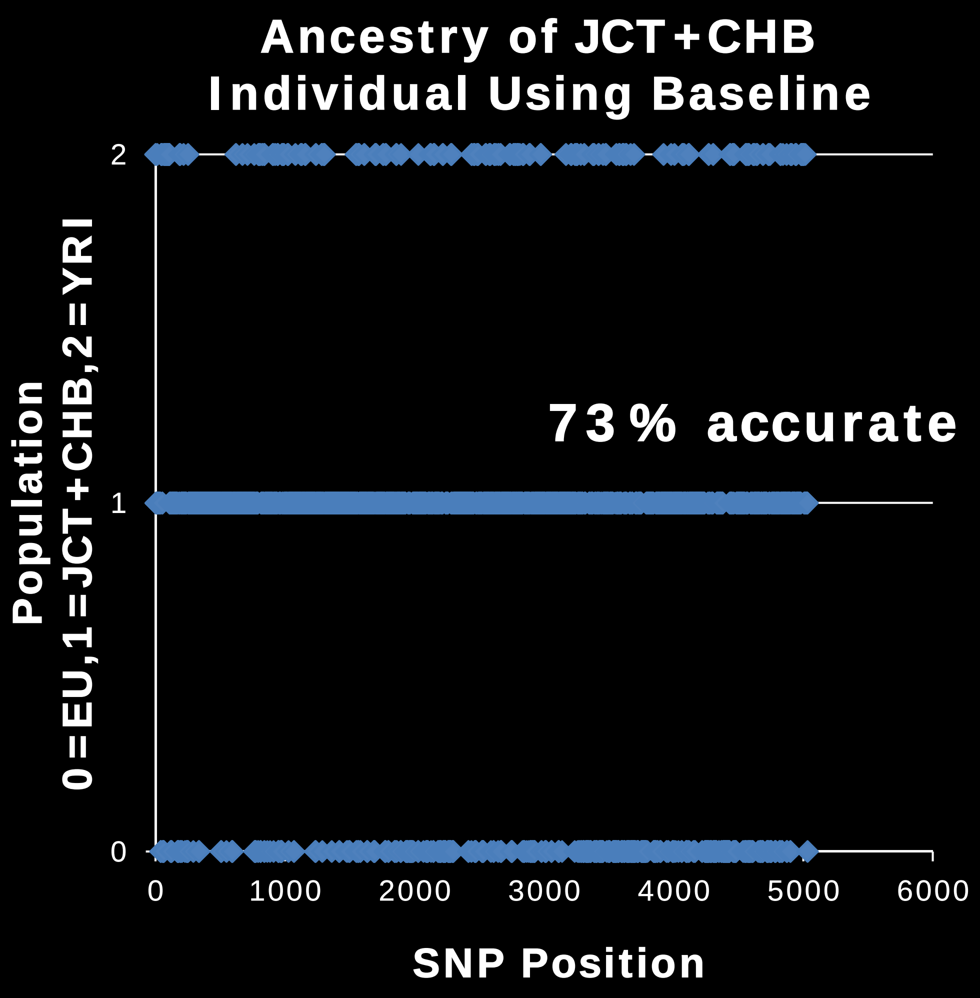
The Best Path



0.8	0.7	0.5	0.3	0.2	0.3	0.1
0.6	0.4	0.2	0.1	0.5	0.4	0.2
0.5	0.3	0.1	0.2	0.3	0.2	0.05

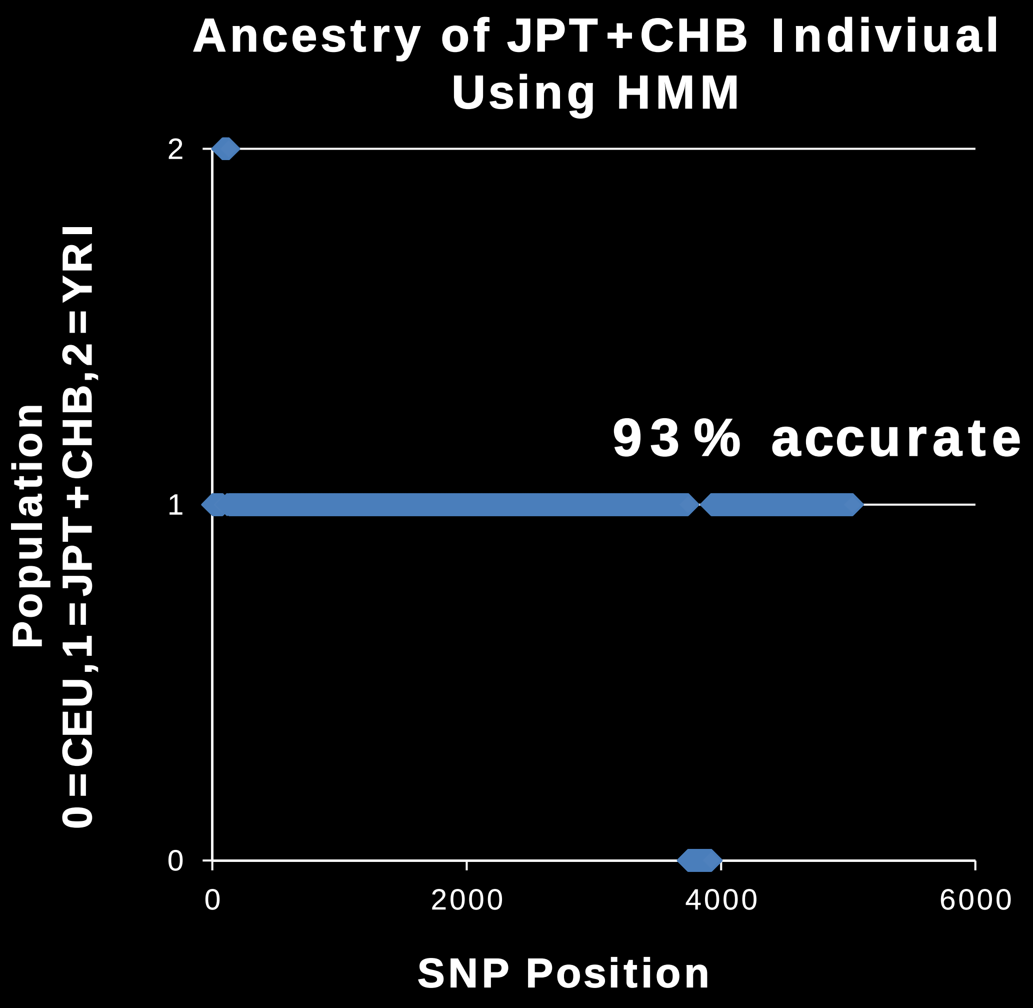
ANCESTRY: 0000111

Comparison: Accuracy on Homogenous Individual


Ancestry of JCT+CHB Individual Using Baseline

73% accurate

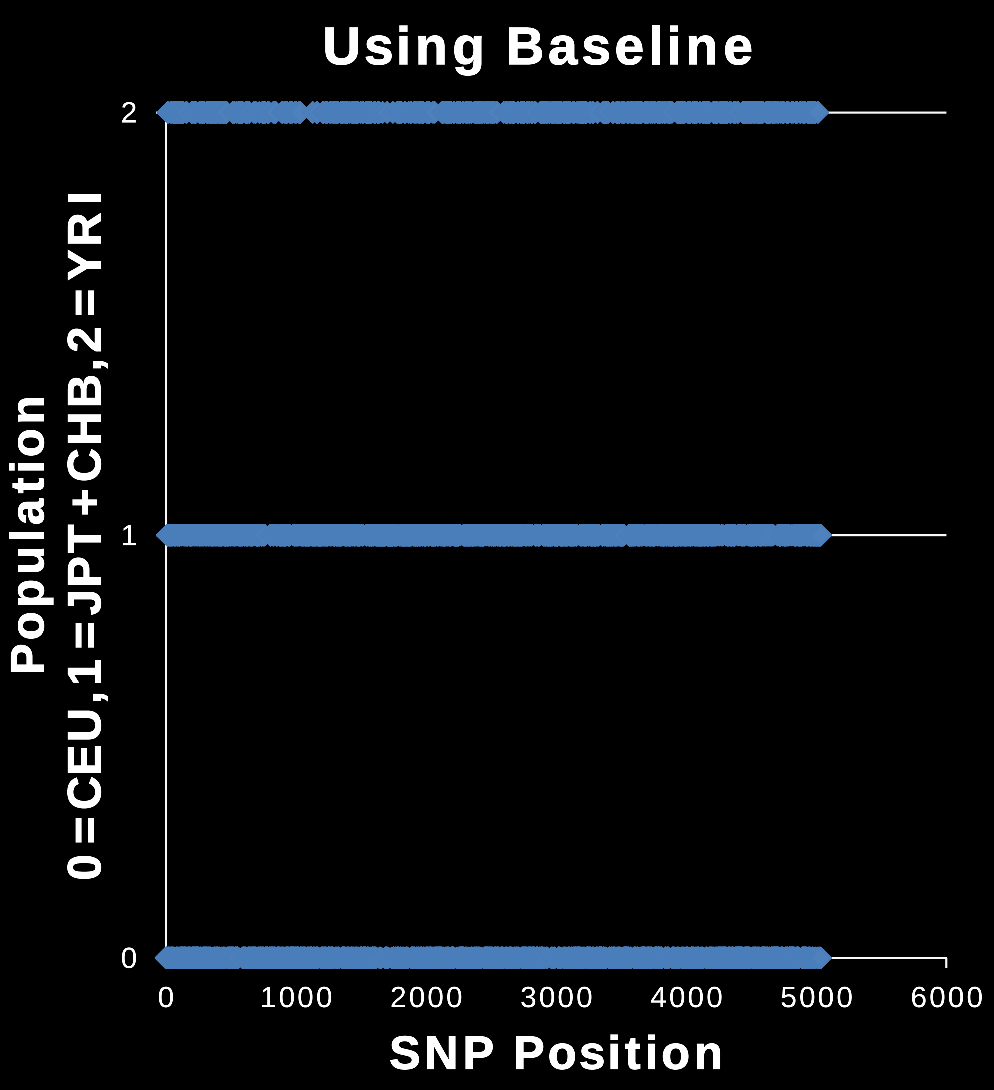
Many small, scattered segments

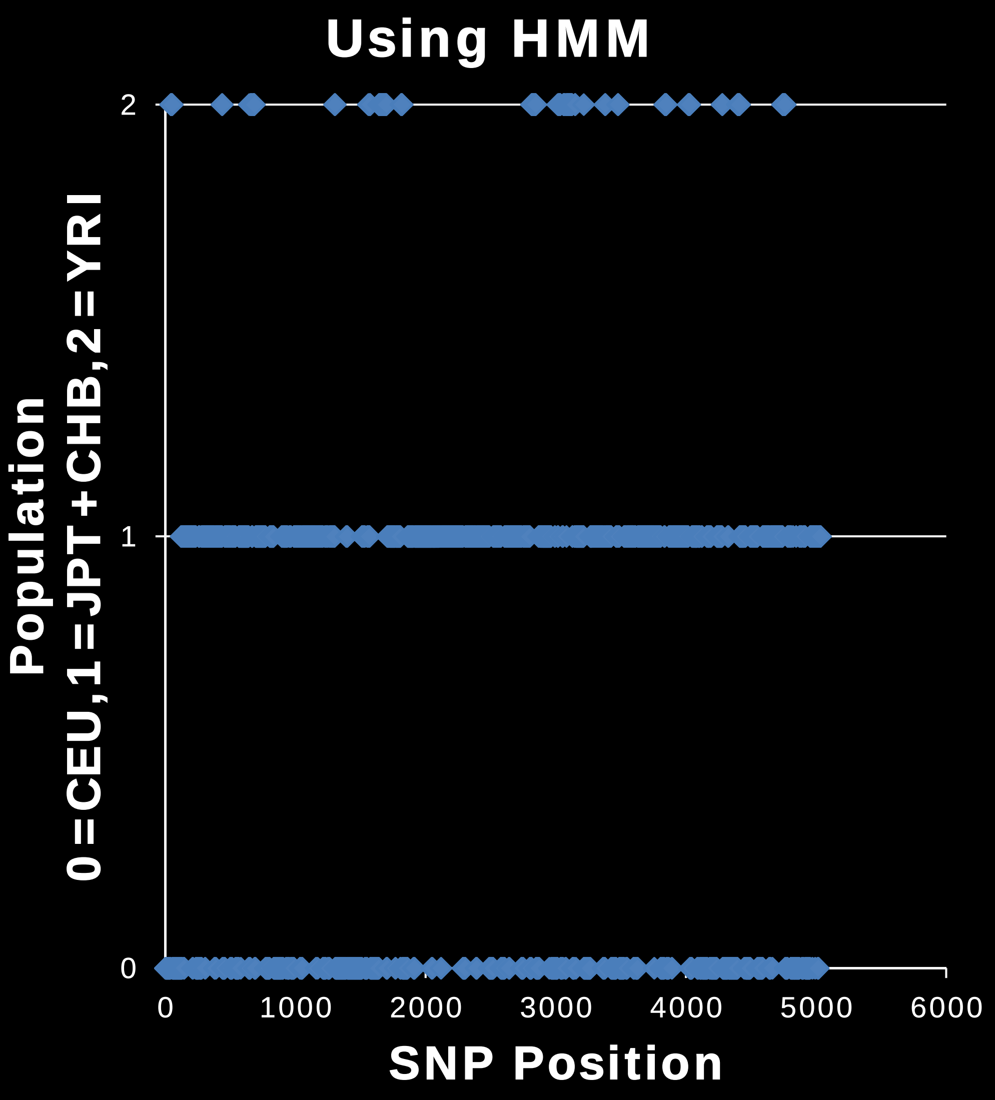

Ancestry of JPT+CHB Indiviual Using HMM

93% accurate

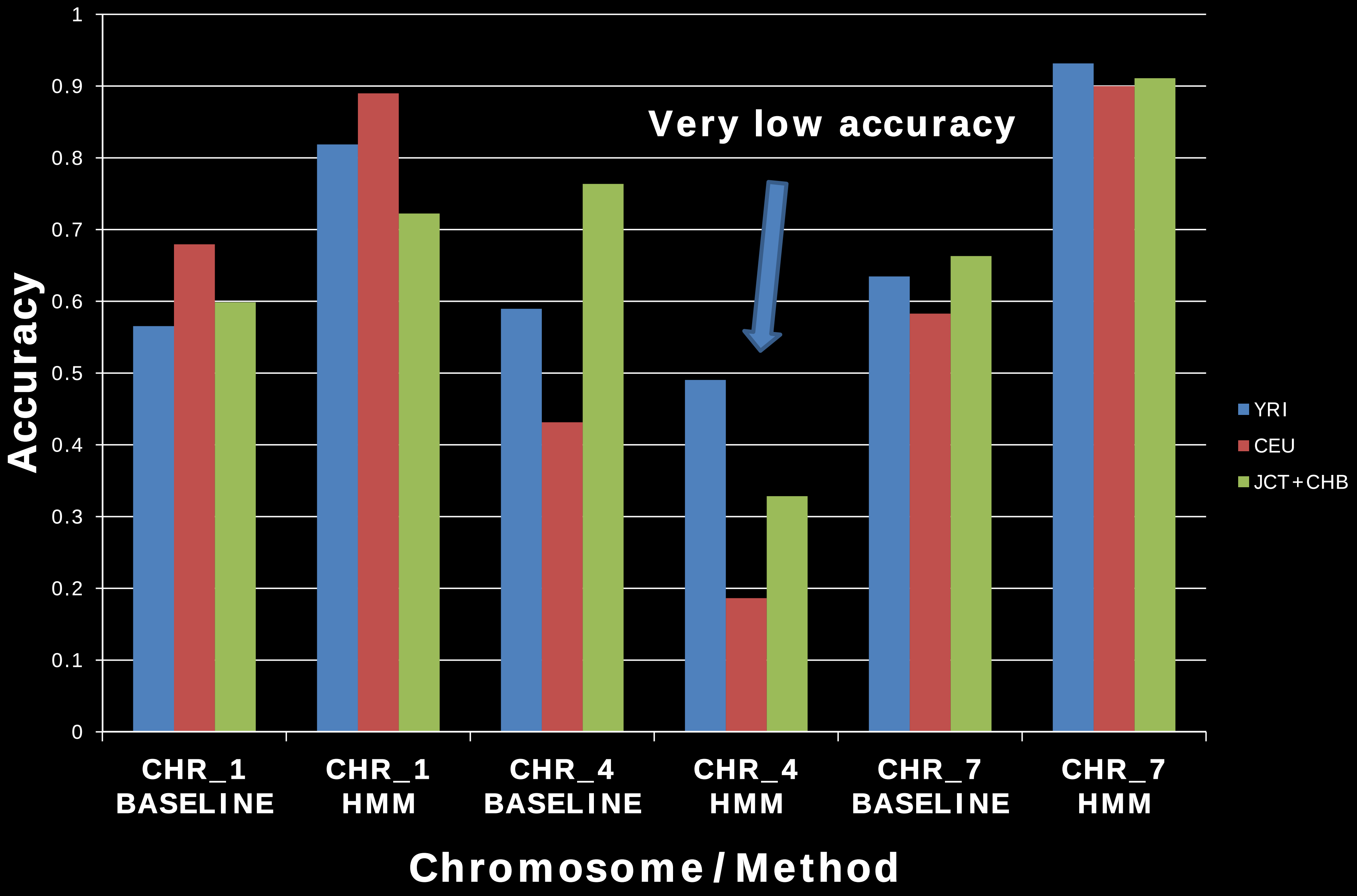
Large contiguous segments

Comparison: Accuracy for ½ CEU ½ JPT+CHB Individual





Comparison: Method Accuracy Across Different Regions



Very low accuracy

Closing Remarks

- The HMM is not perfect
  - Small population segments can be cut out by the low transition rates
  - A person who deviates from his/her own population can be misrepresented

- Problems/Extensions
  - Different chromosome regions produce varying results
  - Results from mixed individual were not accurate with either method