Population genetic and admixture analyses

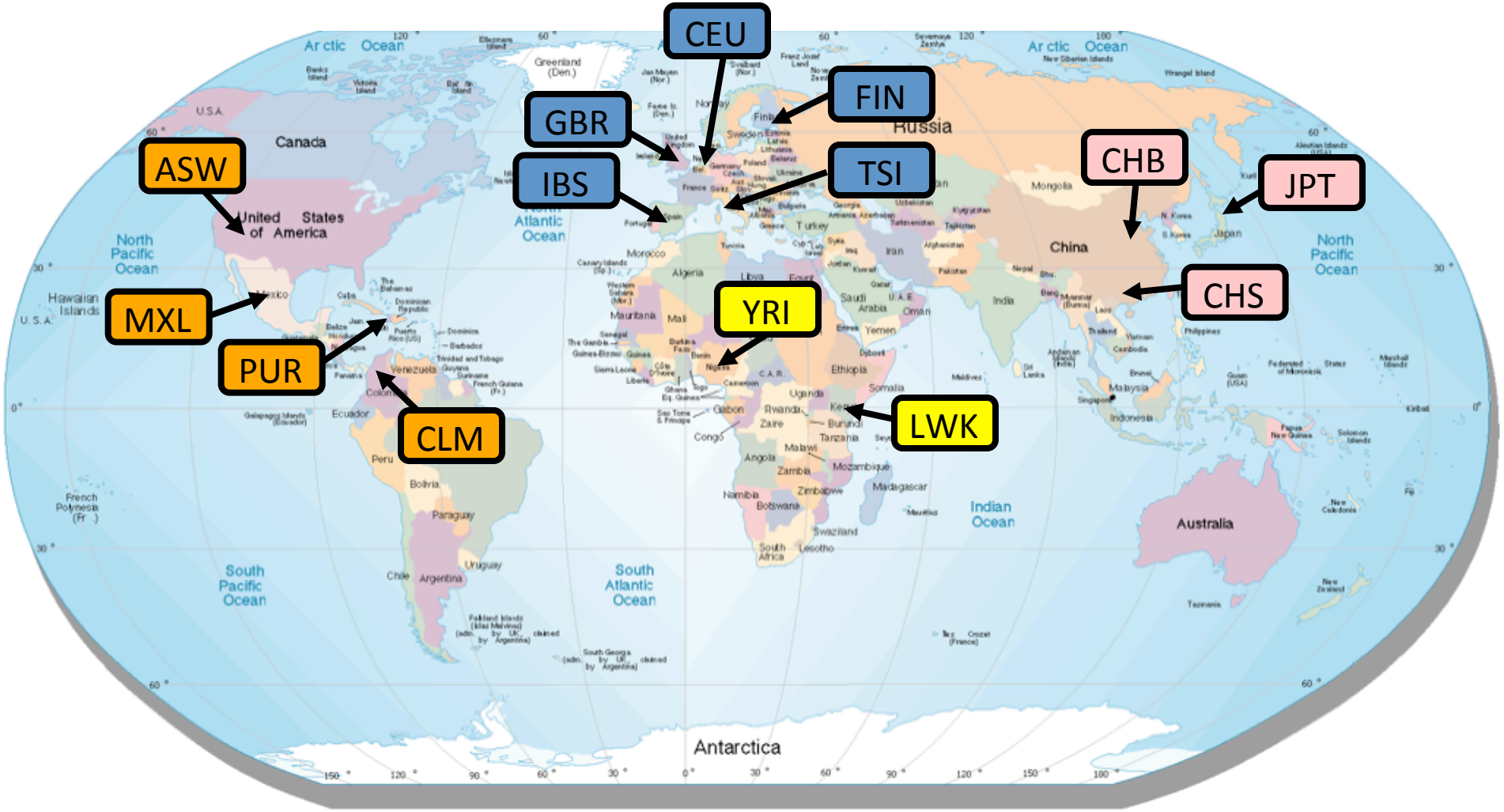
Eimear Kenny

Assistant Professor

Institutes of Personalized Medicine and Multiscale Biology at Mount Sinai School of Medicine

Representing and presenting slides from the 1000 Genomes Project Consortium

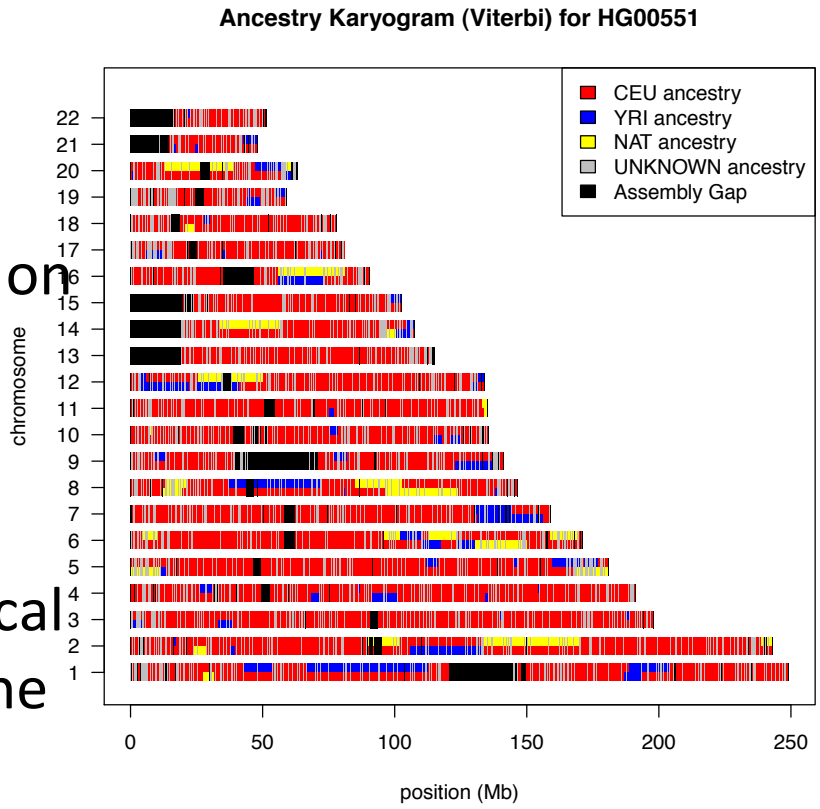
Phase1: 1092 samples from 14 populations



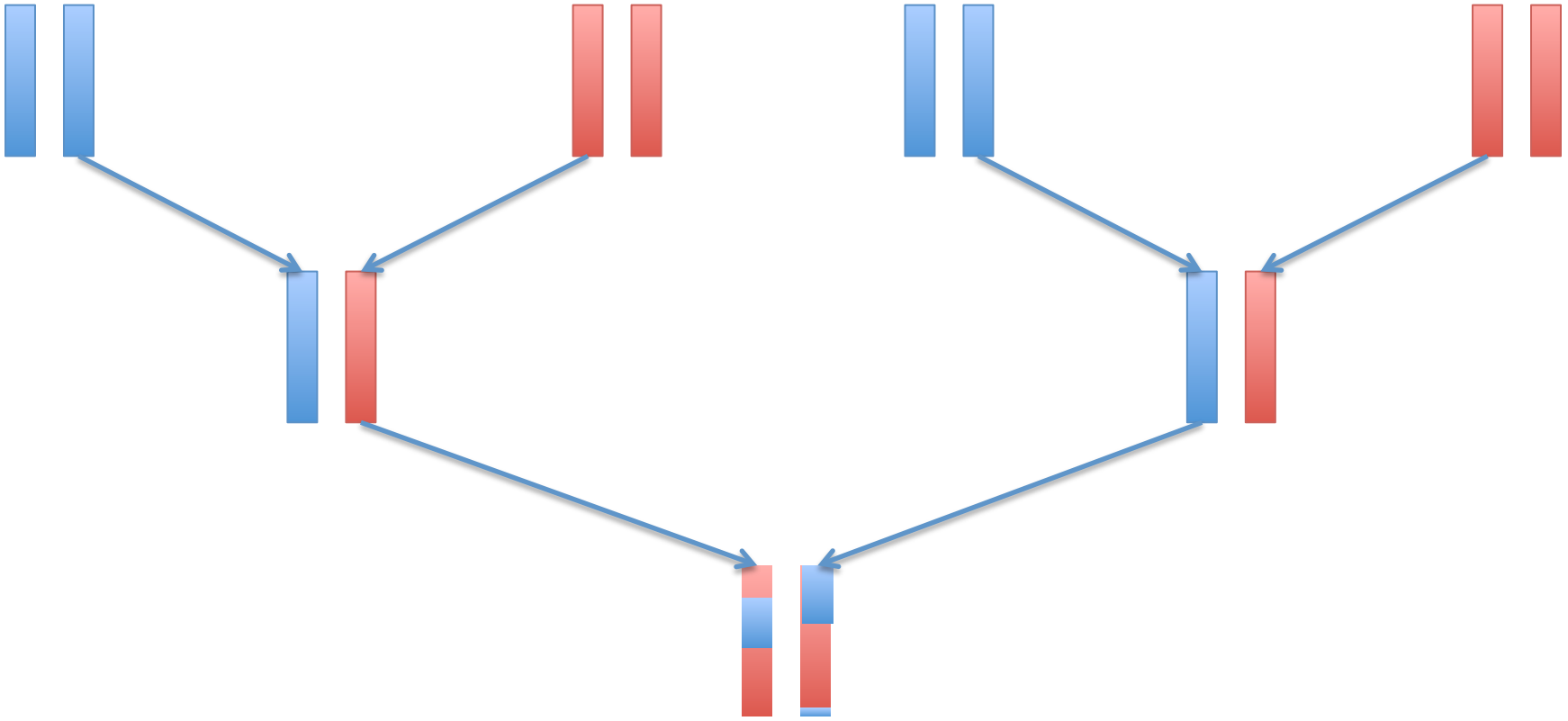
Motivation for Local Ancestry Inference

Using 1000 Genomes admixed populations to:

- study impact of recent admixture on patterns of genomic variation
- understand pre-historical demography
- inform variant discovery for medical genetics and personalized medicine

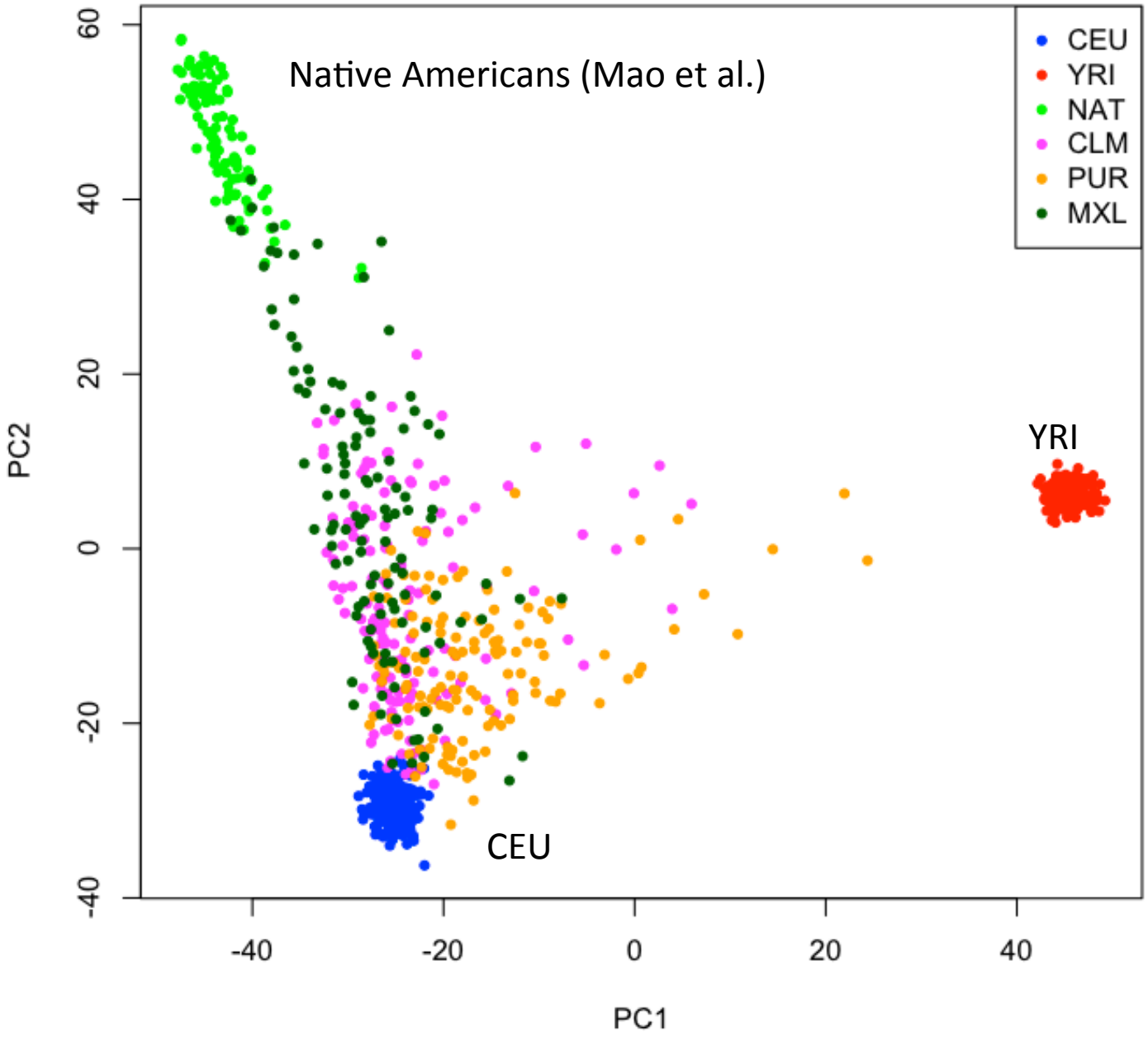


Admixture

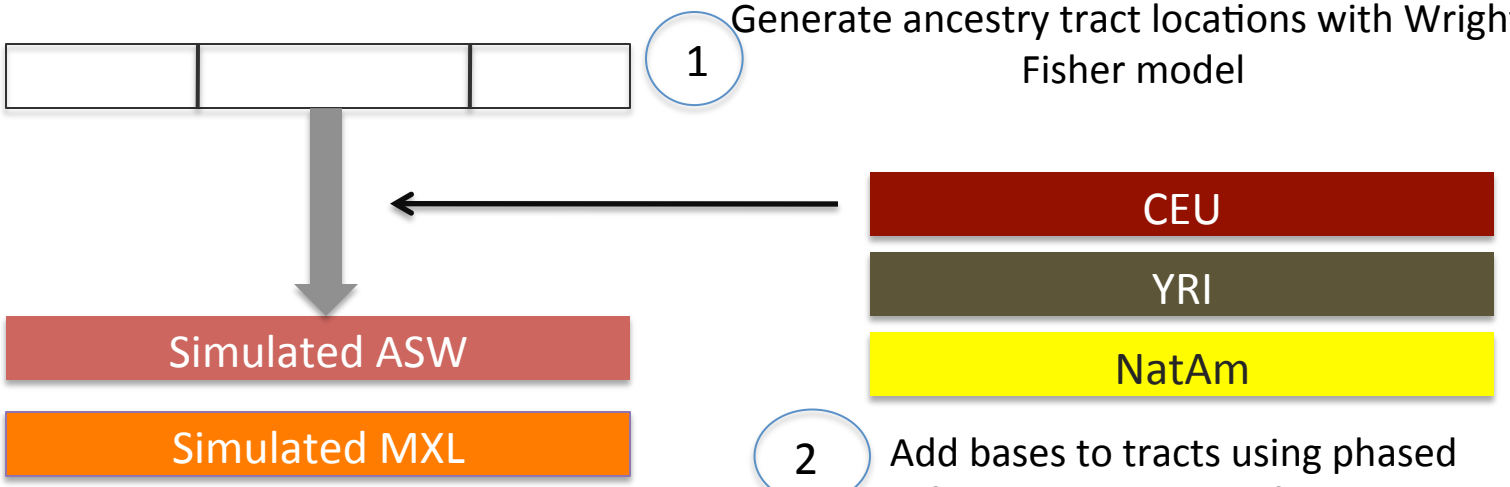


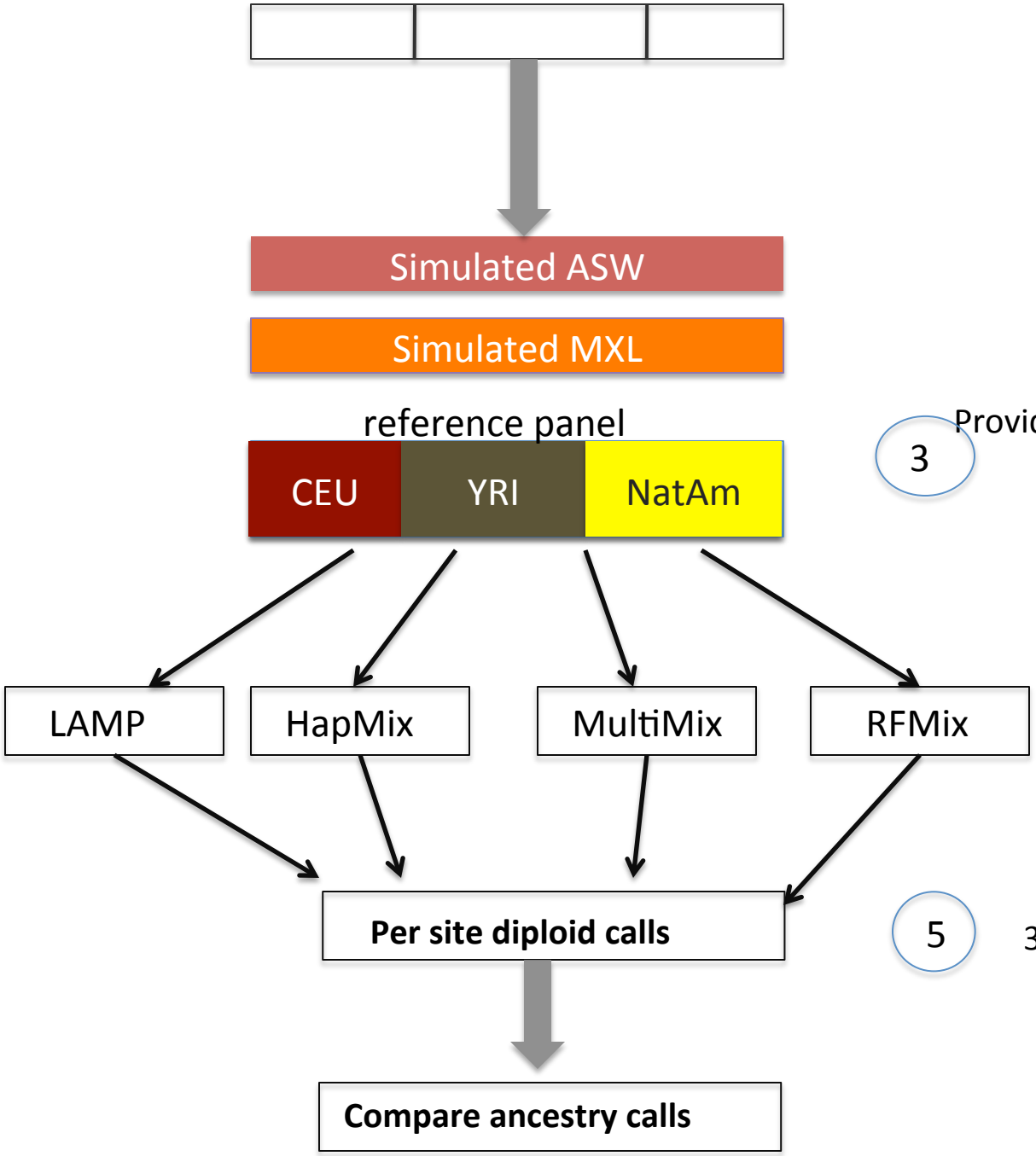
Recombination creates mosaic of ancestry

Phase 1 AMR Samples



Accuracy assessment using simulated haplotypes


1
Generate ancestry tract locations with Wright-Fisher model
CEU
YRI
NatAm
Simulated ASW
Simulated MXL
2
Add bases to tracts using phased reference haplotypes from HapMap + Mao *et al.*

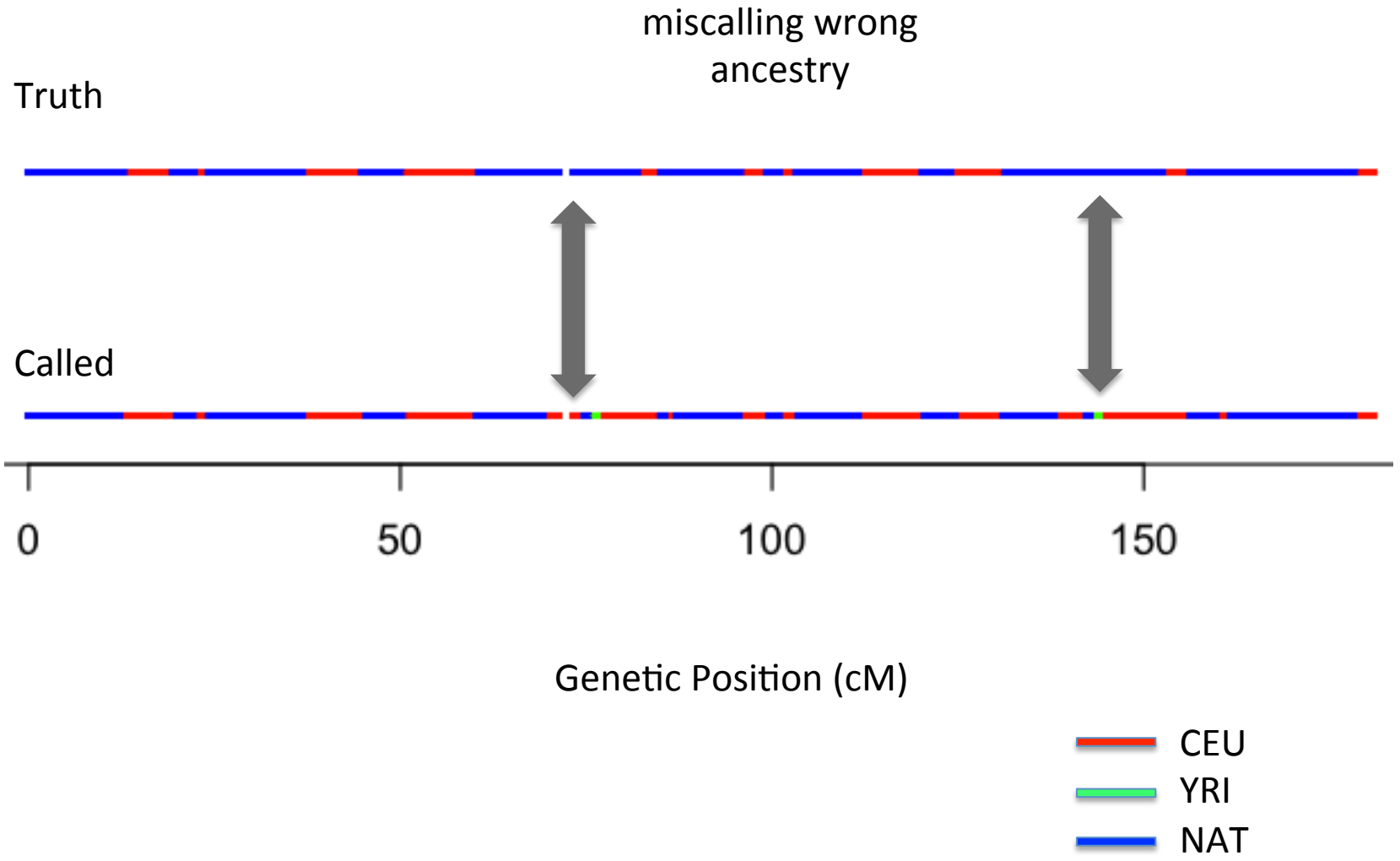


3 Provide reference panel from other HapMap samples + Mao *et al*

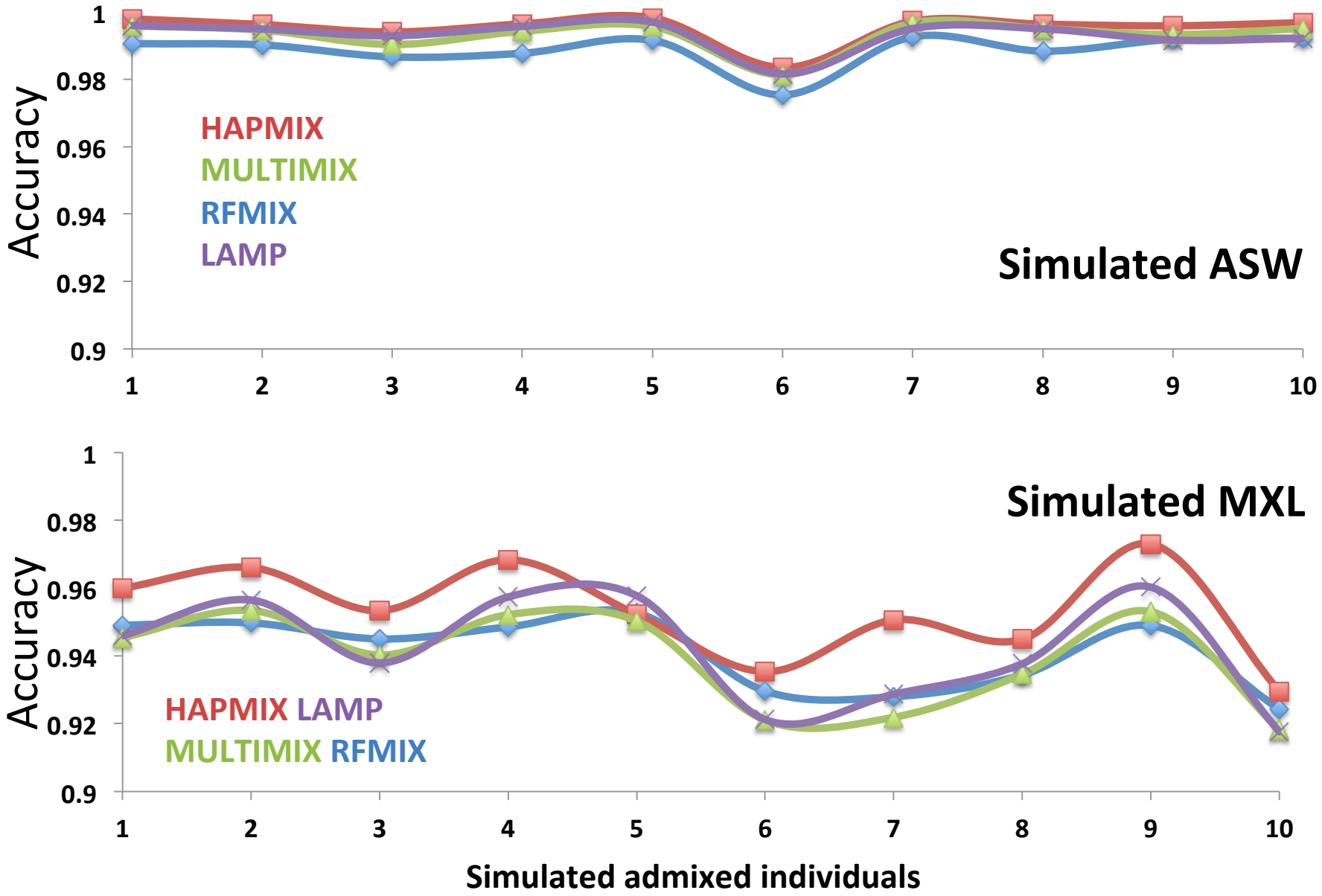
4 Each group runs local ancestry estimation

5 3- or 6-way diploid call per site

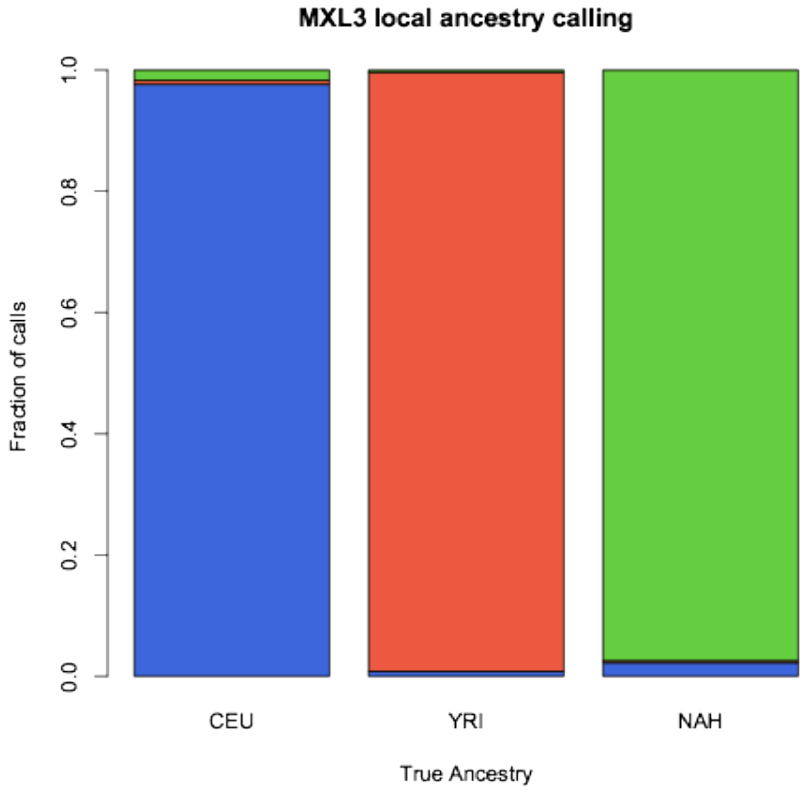
Example miscall

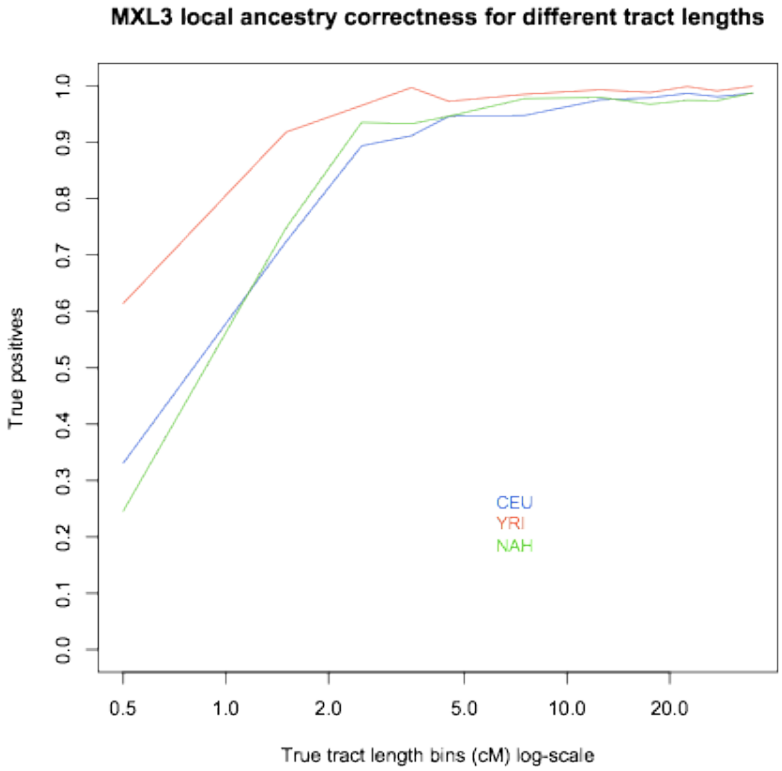


Results from the simulated data: some methods perform better than others

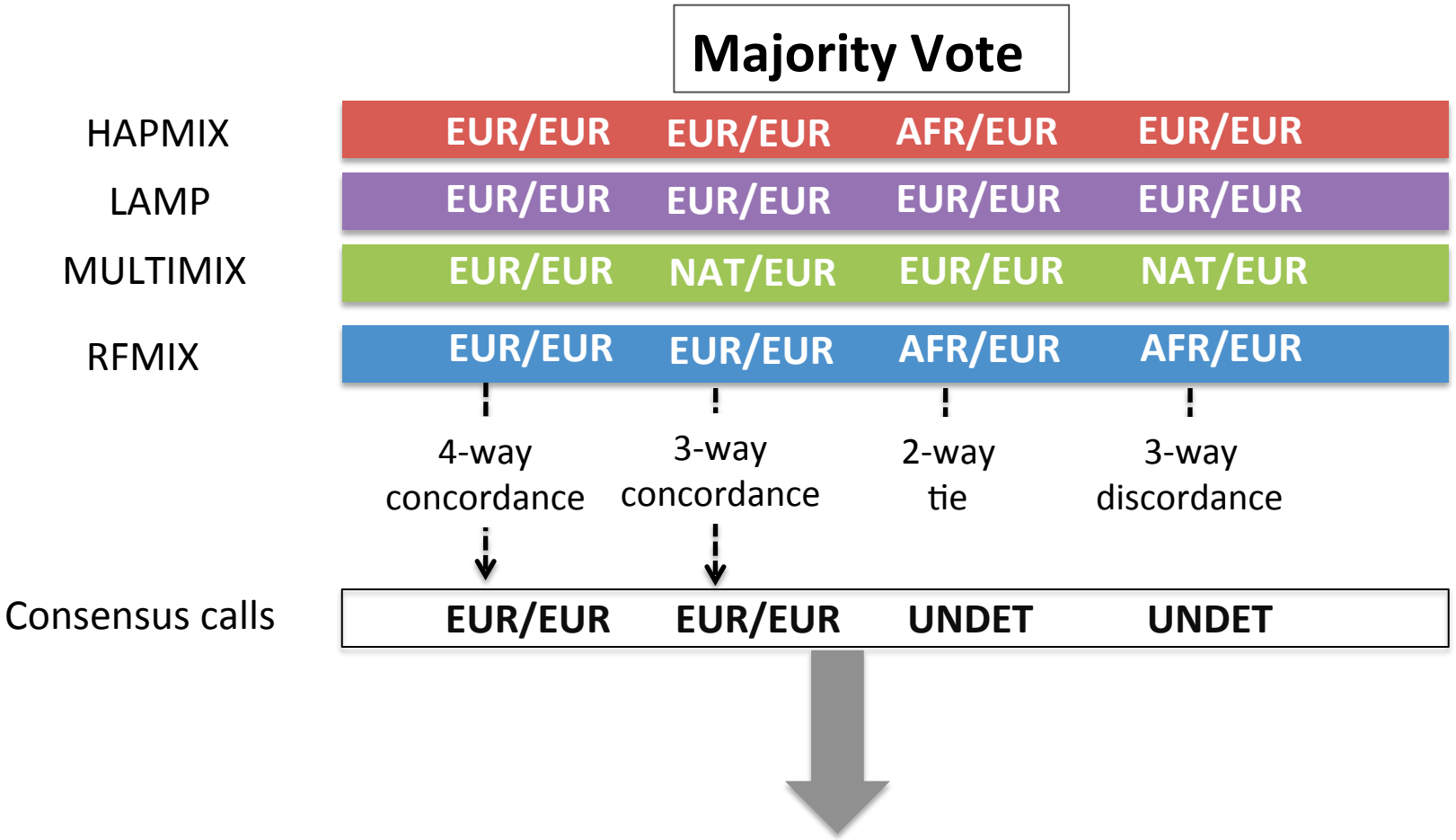


HapMix MXL ($K = 3$) Results



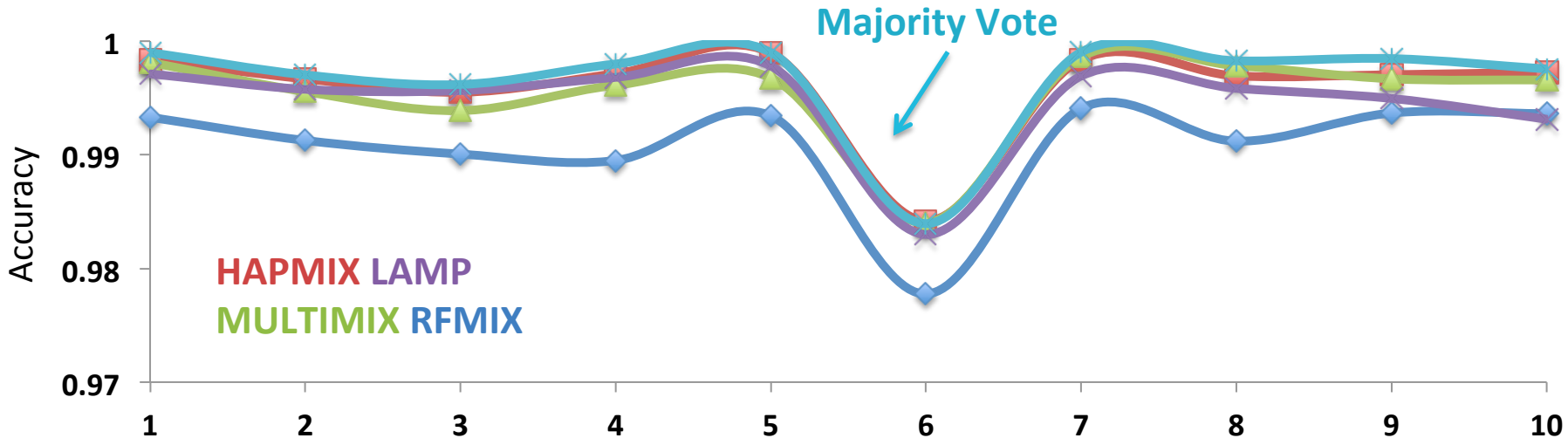


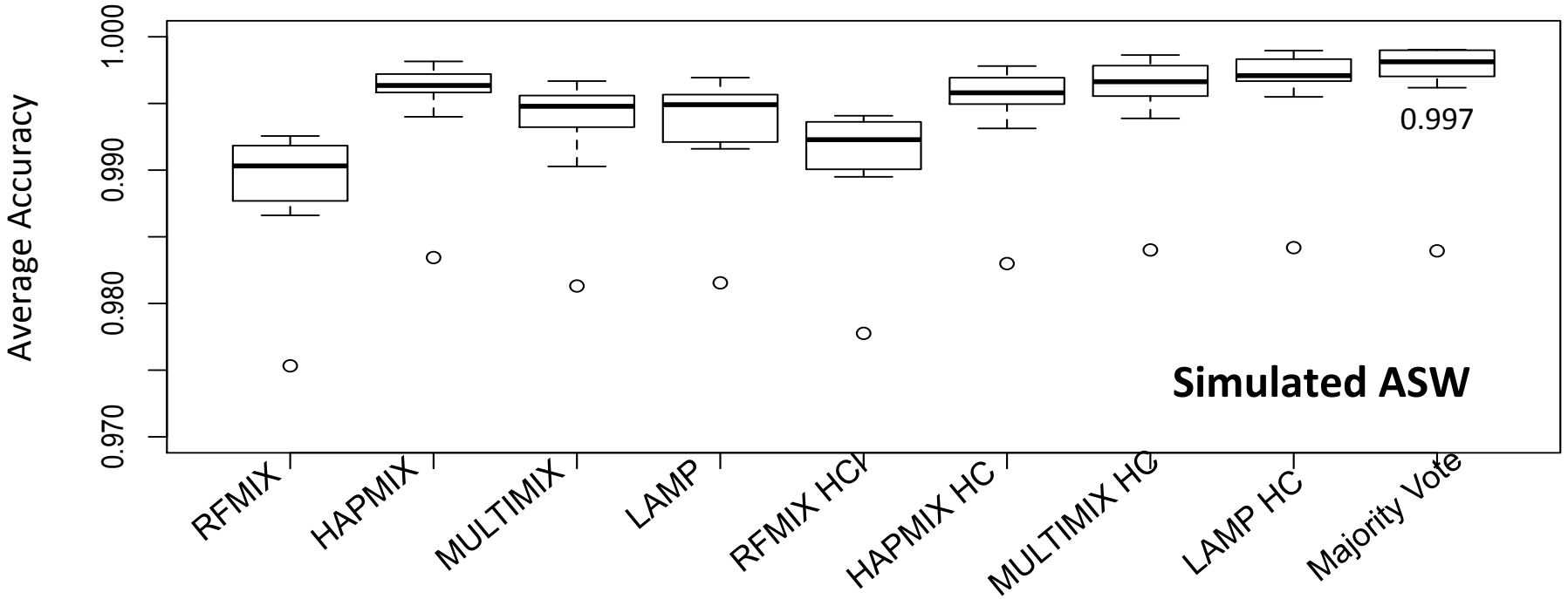
Strategy to obtain highest confidence calls for downstream analysis



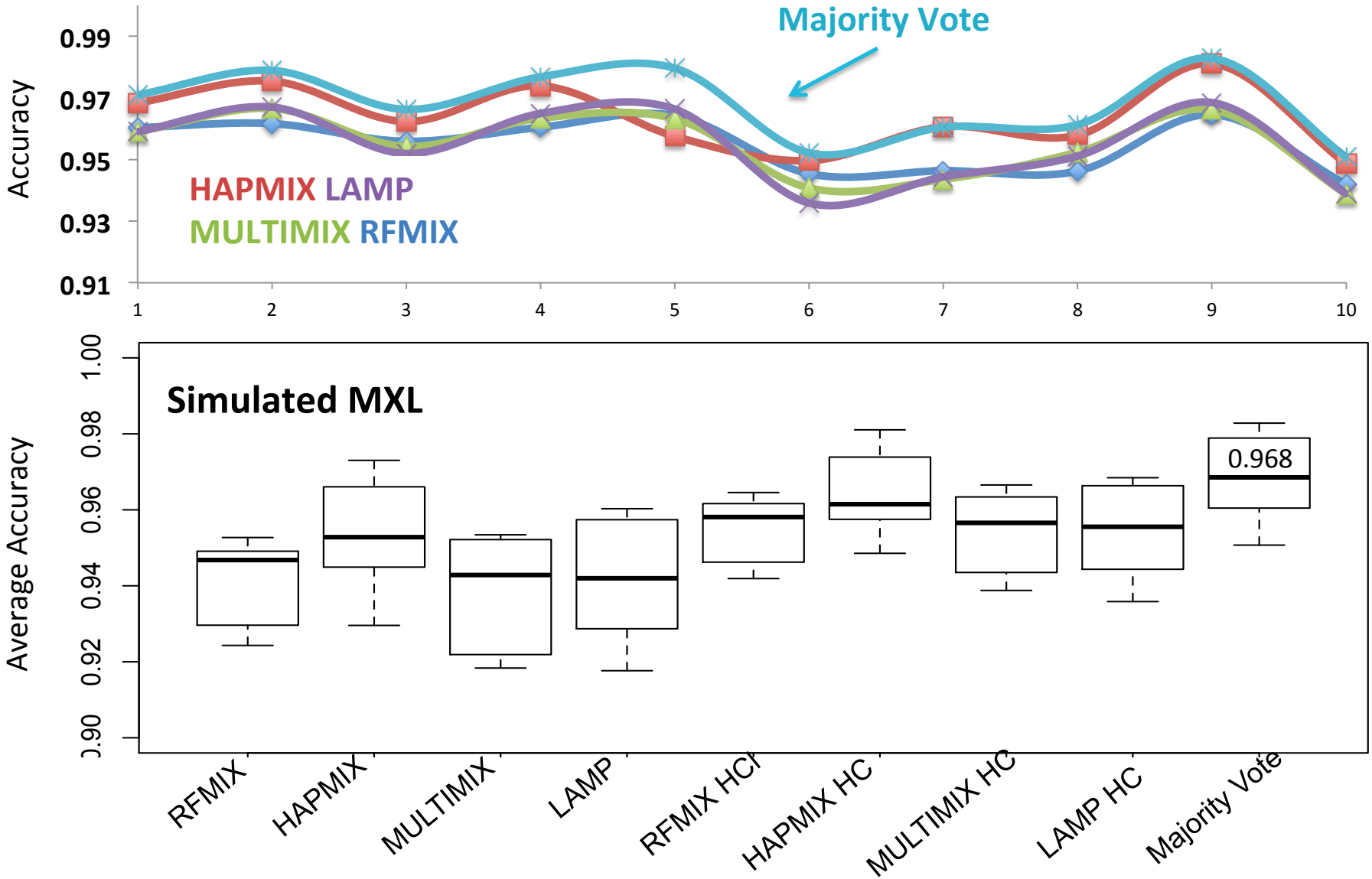
Consensus calls for **97.4%** MXL and **99.3%** ASW genomes

Marginal improvement for ASW

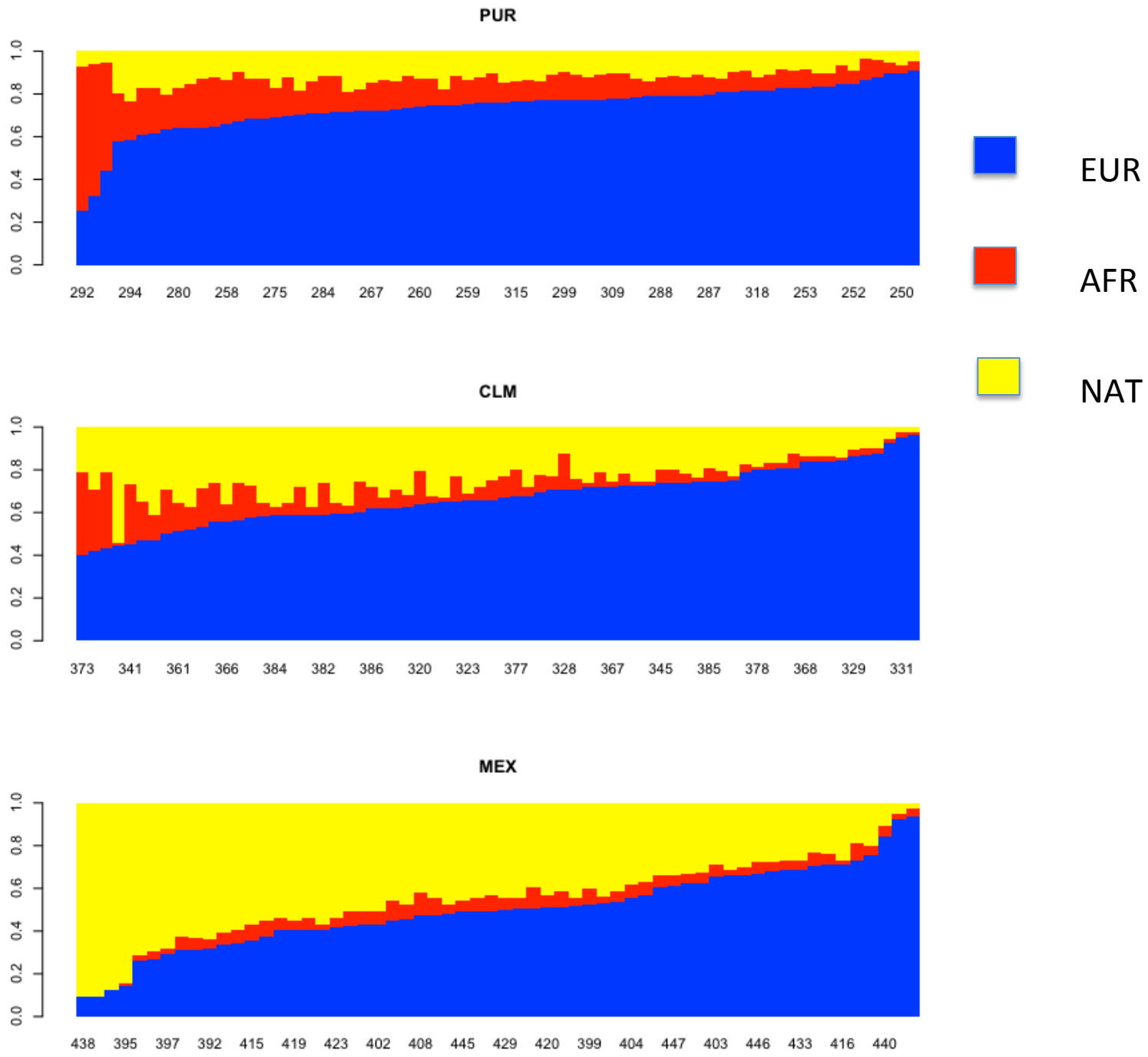




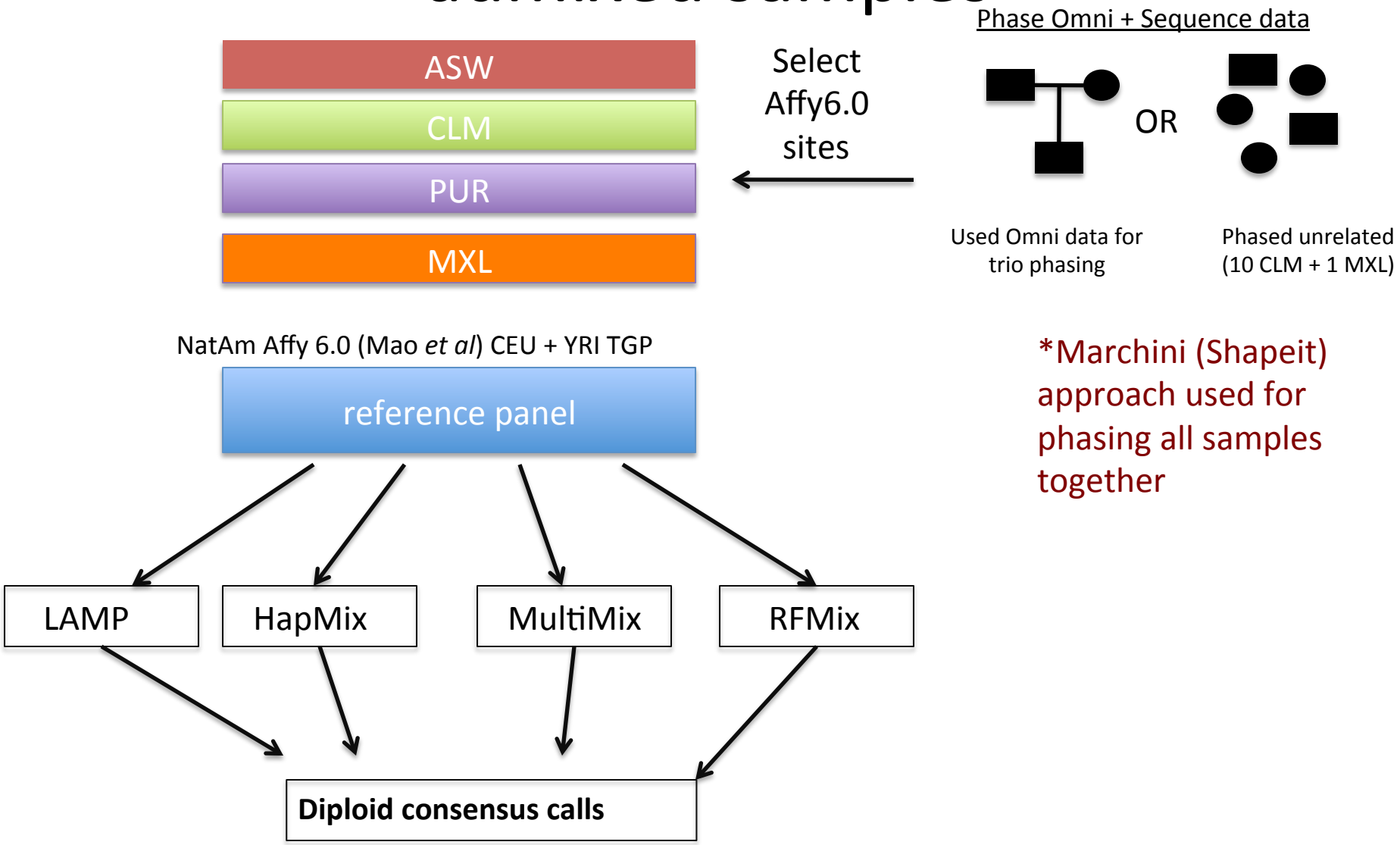
For downstream analysis select high confidence regions



Sequence diversity in admixed TGP samples

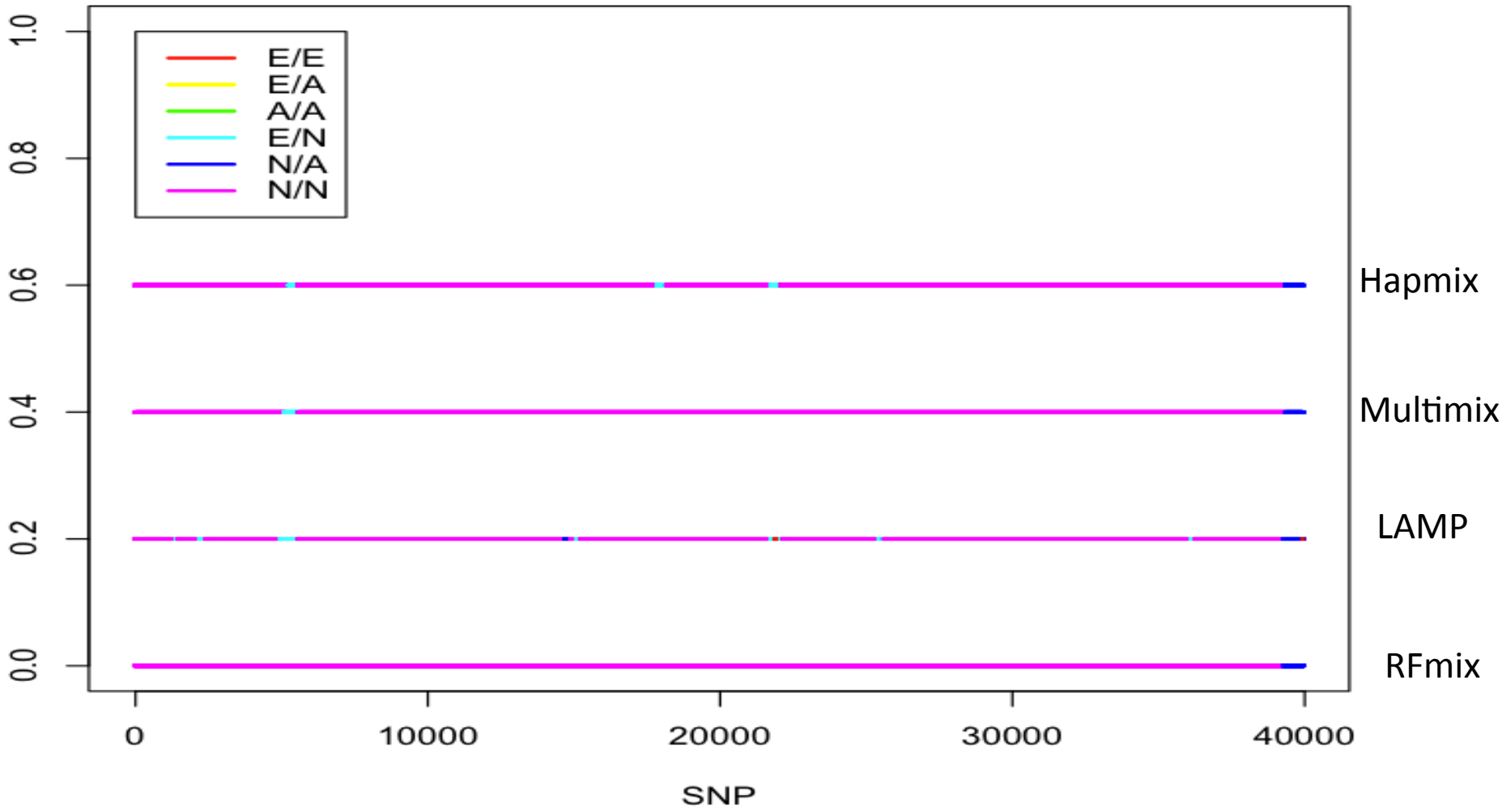

PUR
0.0
0.2
0.4
0.6
0.8
1.0
292 294 280 258 275 284 267 260 259 315 299 309 288 287 318 253 252 250
CLM
373 341 361 366 384 382 386 320 323 377 328 367 345 385 378 368 329 331
MEX
438 395 397 392 415 419 423 402 408 445 429 420 399 404 447 403 446 433 416 440
EUR
AFR
NAT

Inferred local ancestry calling in admixed samples



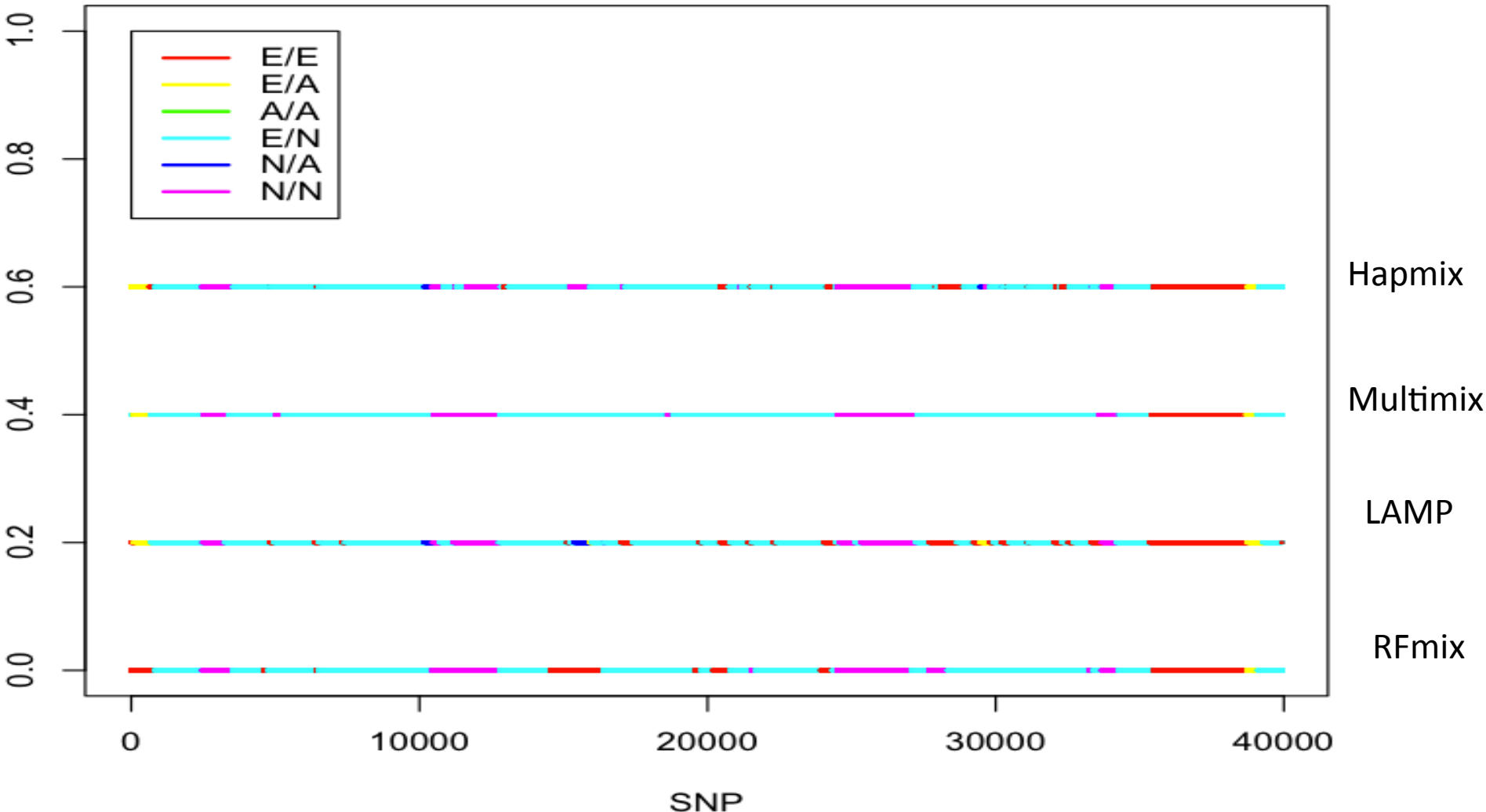
*Marchini (Shapeit) approach used for phasing all samples together

Good agreement



Not so good agreement

(Probably best to “mask” and not make call…)

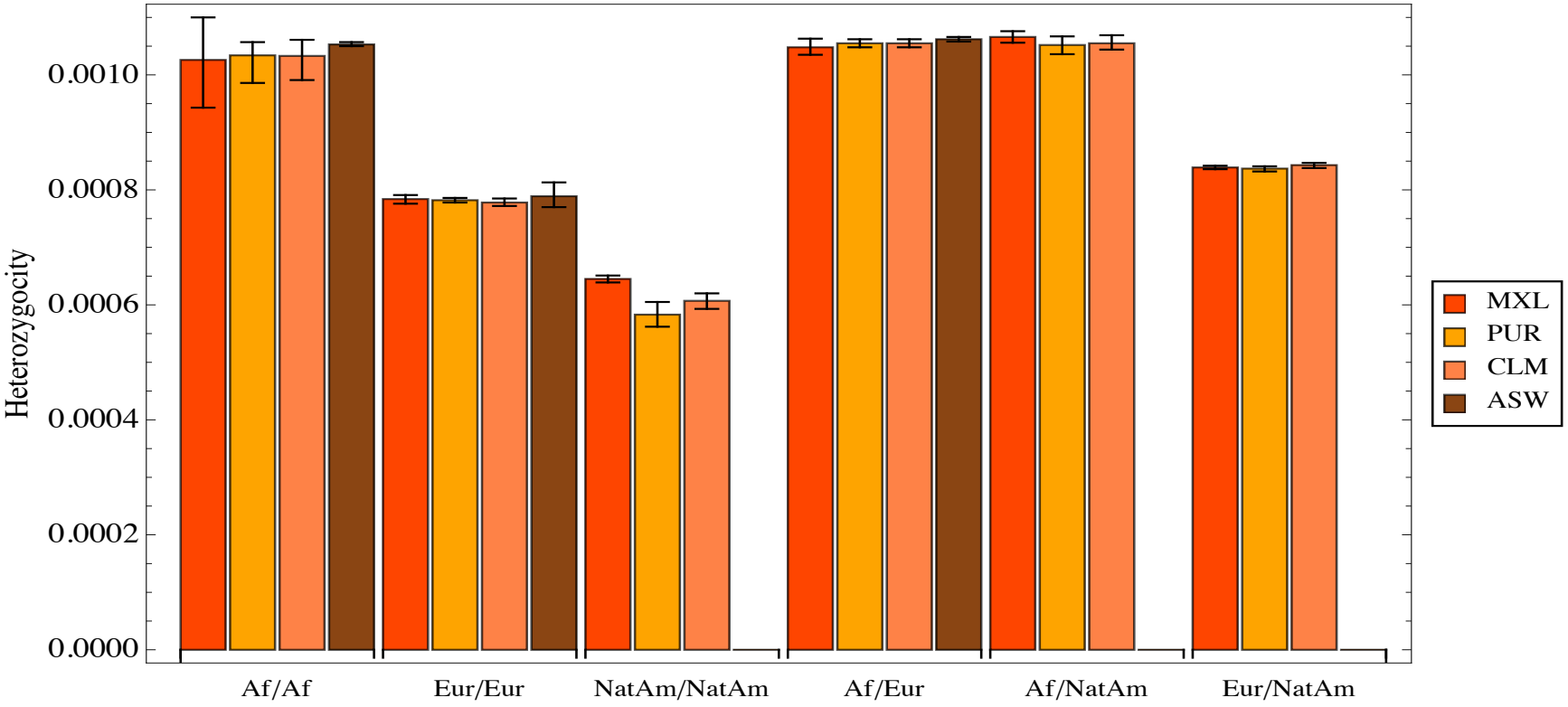


Summary

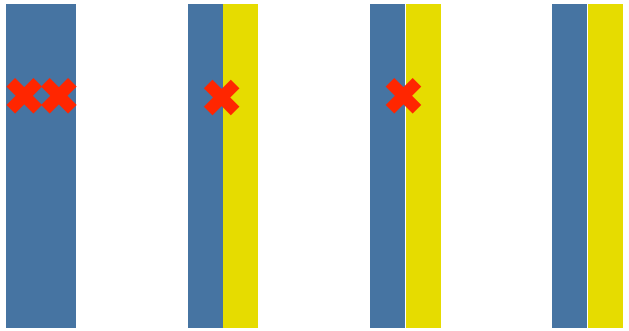
- Admixture deconvolution produced for all Phase 1 samples and pushed to DCC
  - Merged calling improved accuracy
  - Estimated 99%+ accuracy for ASW and 96-97% for MXL, CLM, PUR samples
- Local ancestry calls are available as bed files per individual on the DCC

Some cool applications

Heterozygosity per population



Inferring “ancestral” allele frequencies

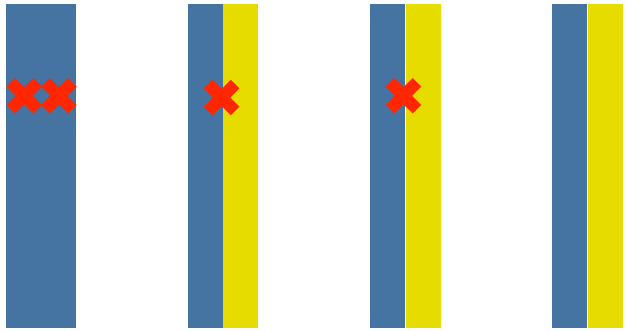


What are the allele frequencies in each population?

-Bayesian answer

$$P(\mathbf{f}|D) = \frac{P(D|\mathbf{f})P(\mathbf{f})}{\int d\mathbf{f}' P(D|\mathbf{f}')P(\mathbf{f}')}$$

Inferring “ancestral” allele frequencies



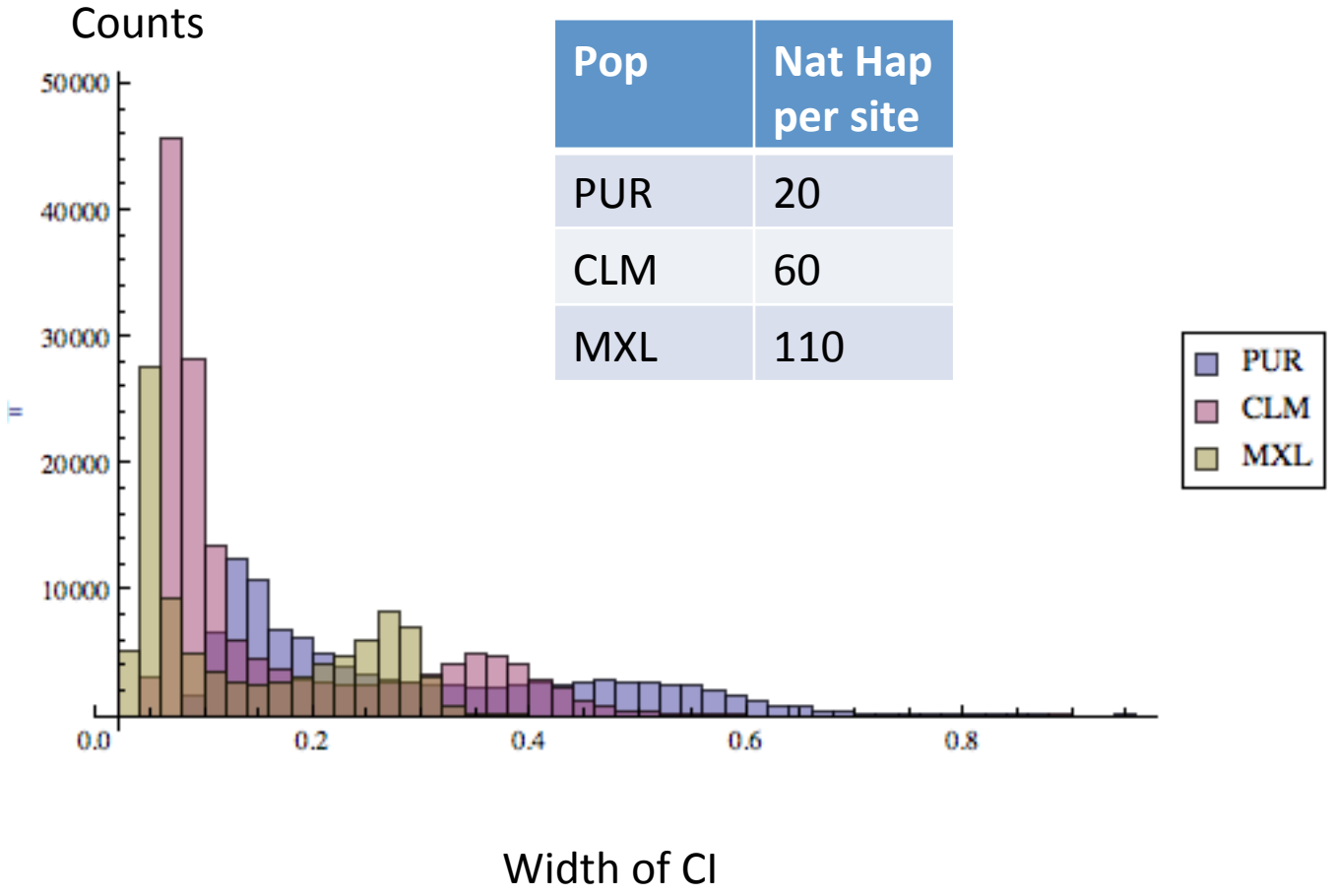
What are the allele frequencies in each population?

-Bayesian answer

$$P(\mathbf{f}|D) = \frac{P(D|\mathbf{f})P(\mathbf{f})}{\int d\mathbf{f}' P(D|\mathbf{f}')P(\mathbf{f}')}$$

Calculate across all sites

Confidence interval width varies by population



Pop	Nat Hap per site
PUR	20
CLM	60
MXL	110

Summary

- SNP diversity, novelty rate, and Non-Syn/Syn ratios mirror demography
- Recover allele frequencies in ancestral populations
- Samples and pipelines are useful beyond TGP

Admixture Working Group

- Bustamante lab (Eimear Kenny, Simon Gravel, Fouad Zakharia, Brian Maples)
- Marchini lab (Claire Churchouse)
- Halperin lab (Yael Baran)
- Myers lab (Anjali GuptaHinch)
- Burchard (Chris Gignoux)
- Abigail Bigham and Mark Shriver (Mao et al. Affy 6.0 data)

Credits

More information at www.
1000genomes.org