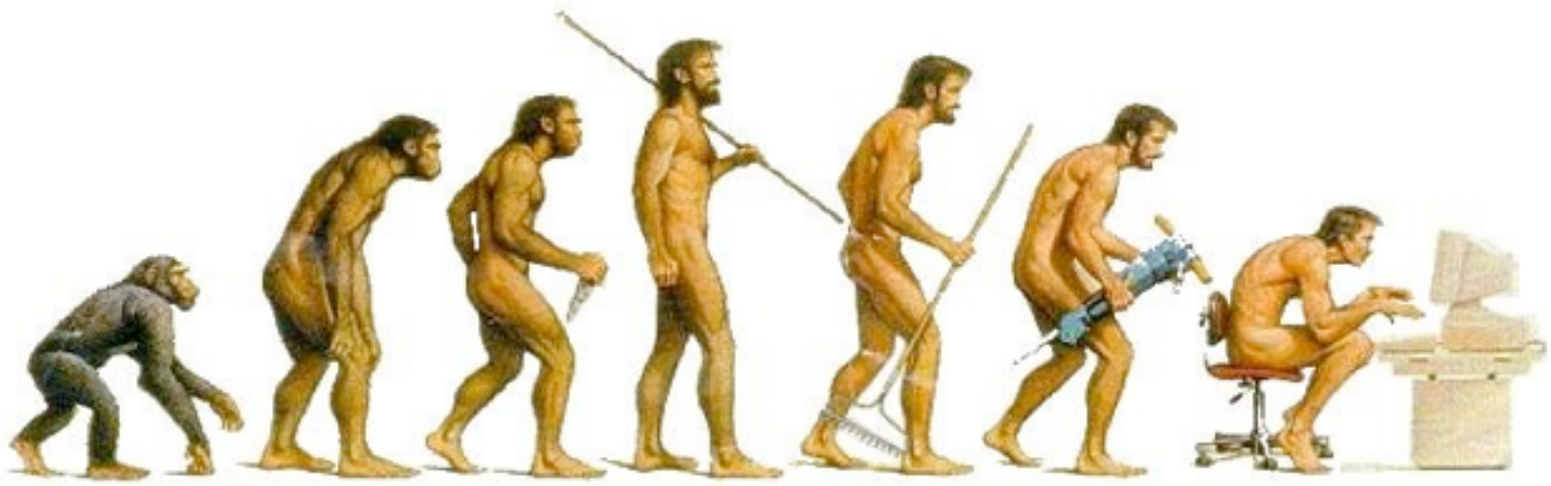
Population Genetic Analyses Using 1000 Genomes Project Data

Ryan Hernandez

UCSF



ryan.hernandez@ucsf.edu twitter: @rdhernand

qb3 ucb·ucsc·ucsf

UCSF University of California San Francisco

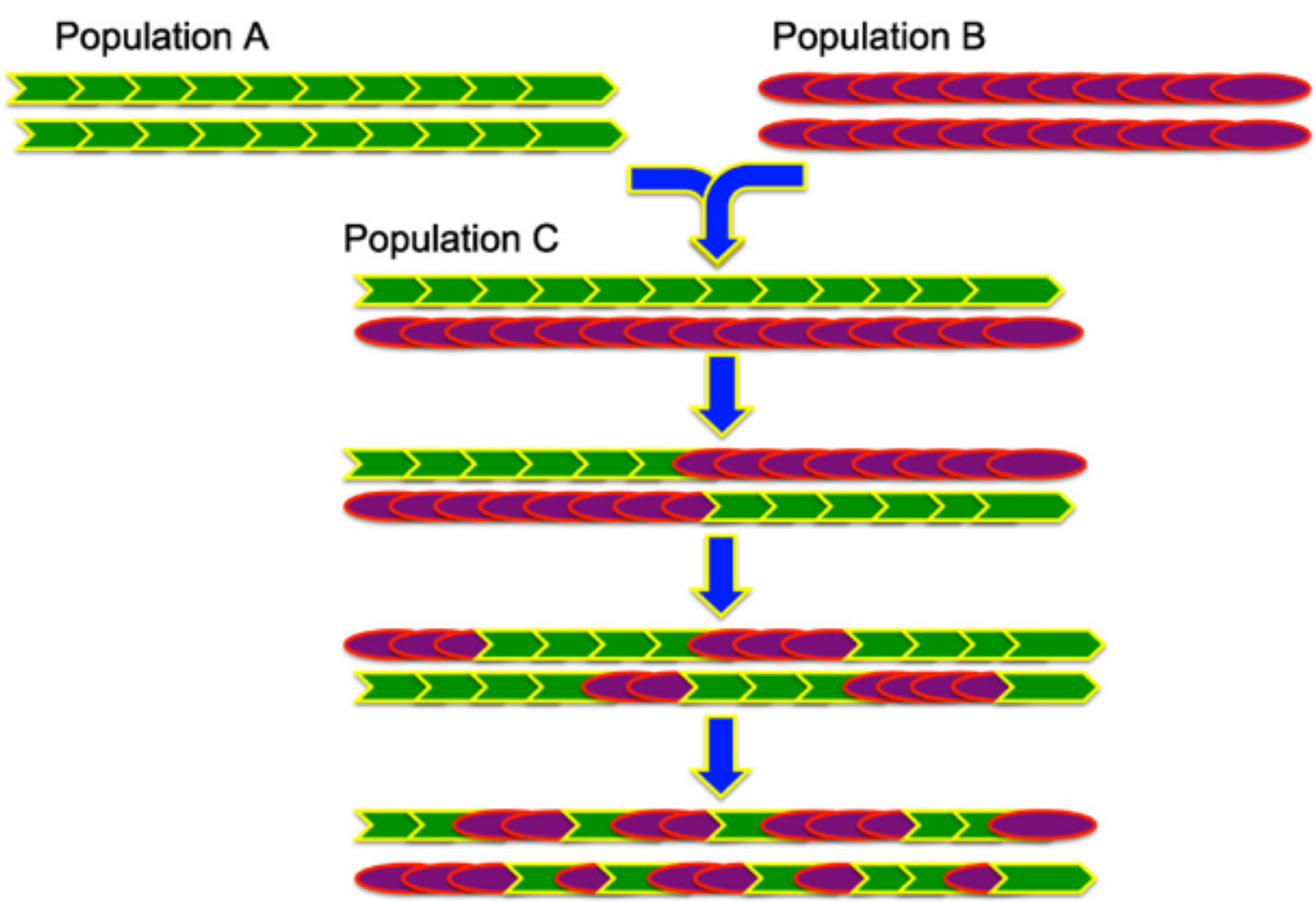
Department of Bioengineering and Therapeutic Sciences

a joint department of the UCSF Schools of Pharmacy and Medicine

Types of questions

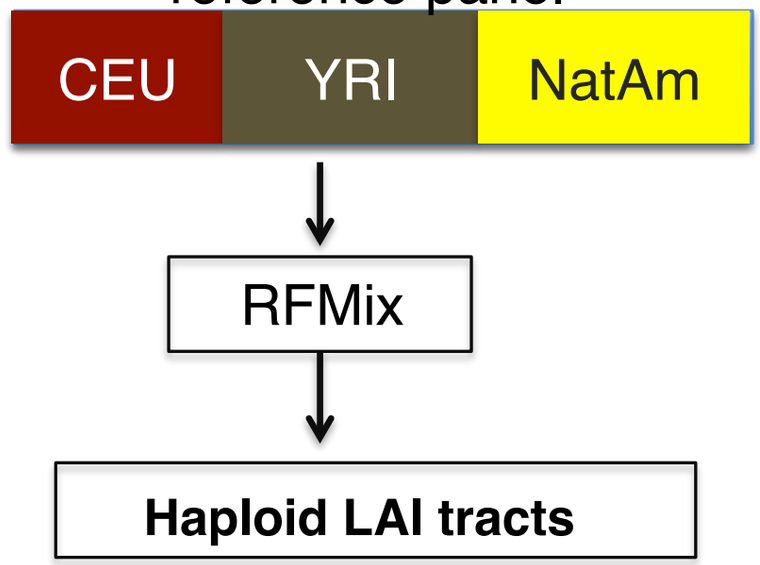
- **Demography**
    - Population expansions/ contractions
    - Population divergence/ mixing
    - Admixture
- **Natural selection**
    - Positive selection
        - Allele frequency distributions
        - Haplotype frequencies
    - Negative selection
        - Allele frequency distributions
        - Allele sharing
        - Haplotype patterns

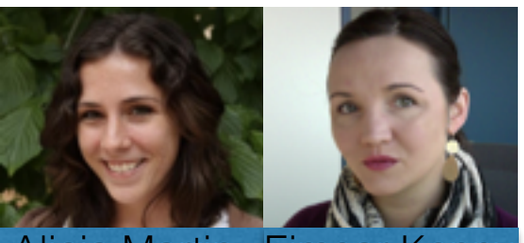
Admixture as a lens into recent human demography



3-way Affy 6.0 pipeline

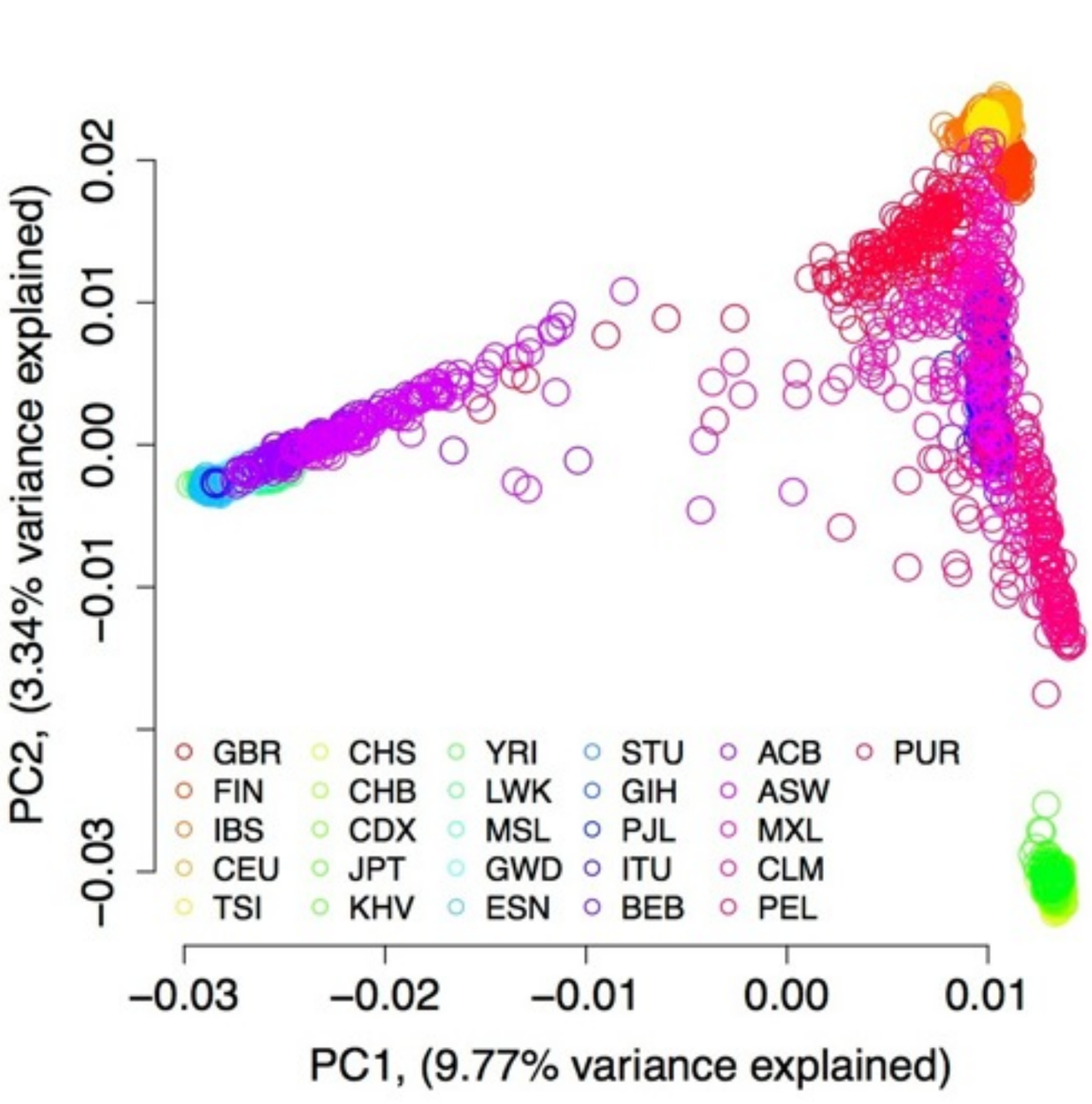
reference panel

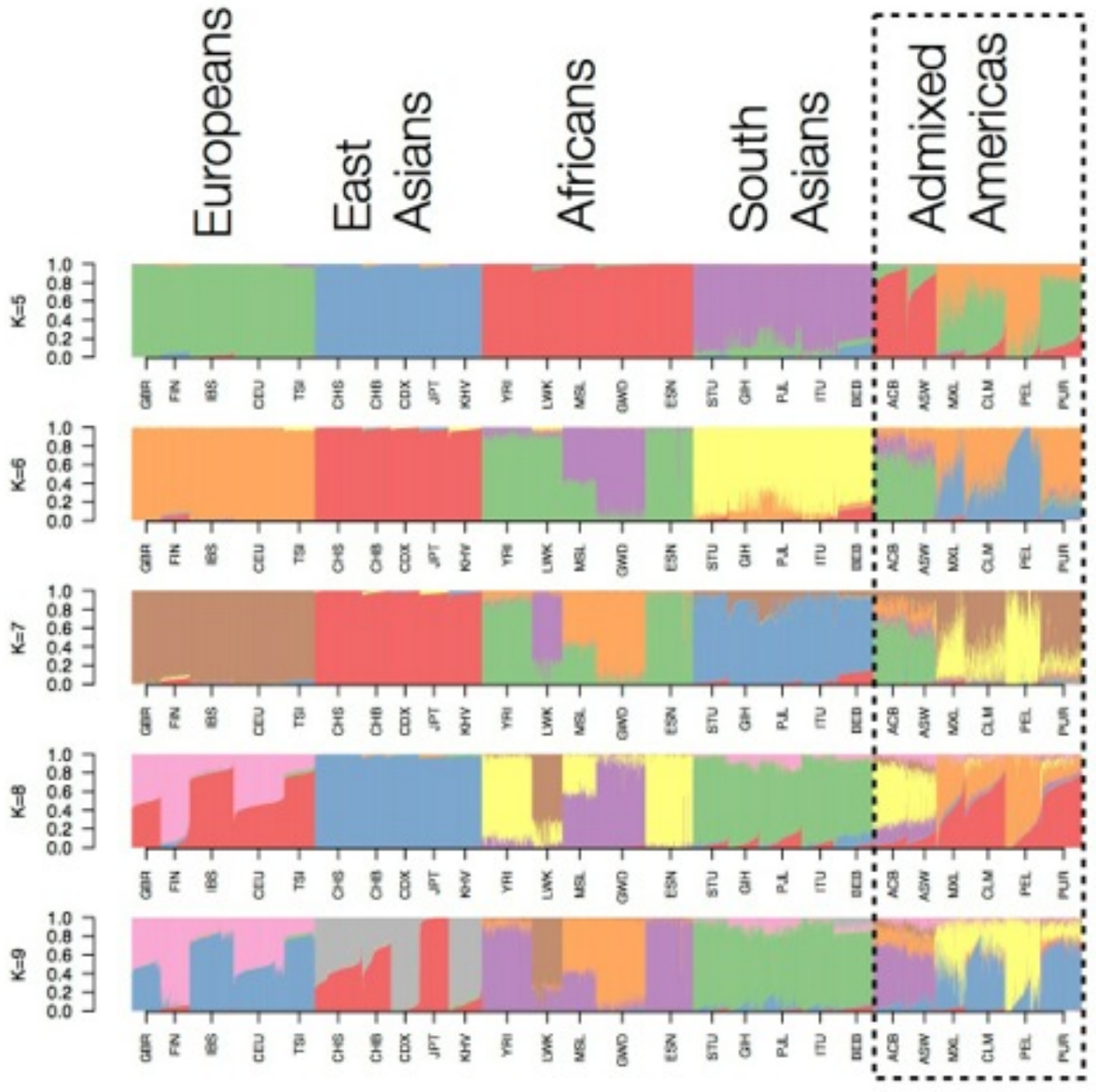




- Recombination breaks haplotypes as well as local ancestry tracts.

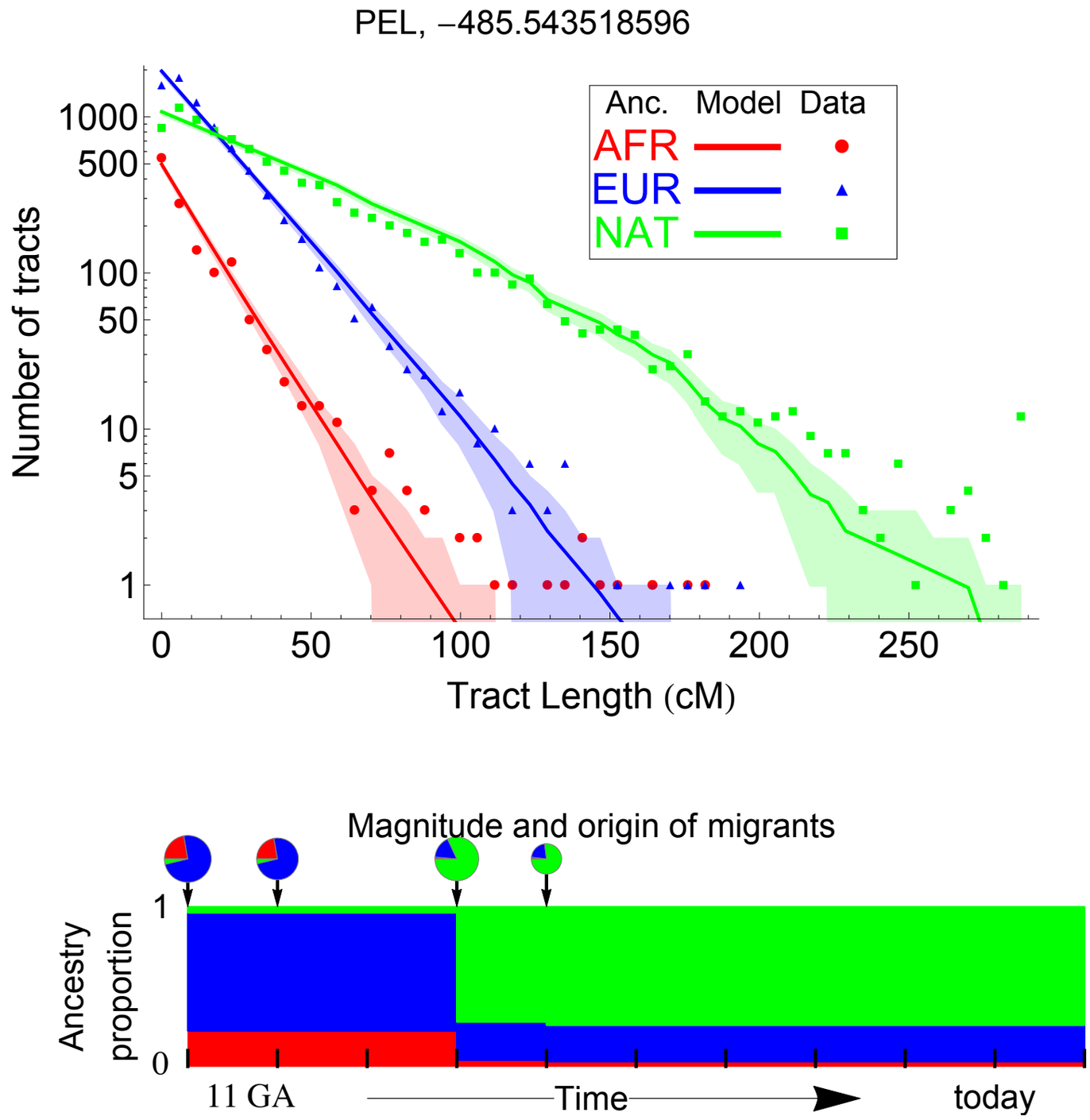
Substantial Global Genetic Diversity in 1000 Genomes





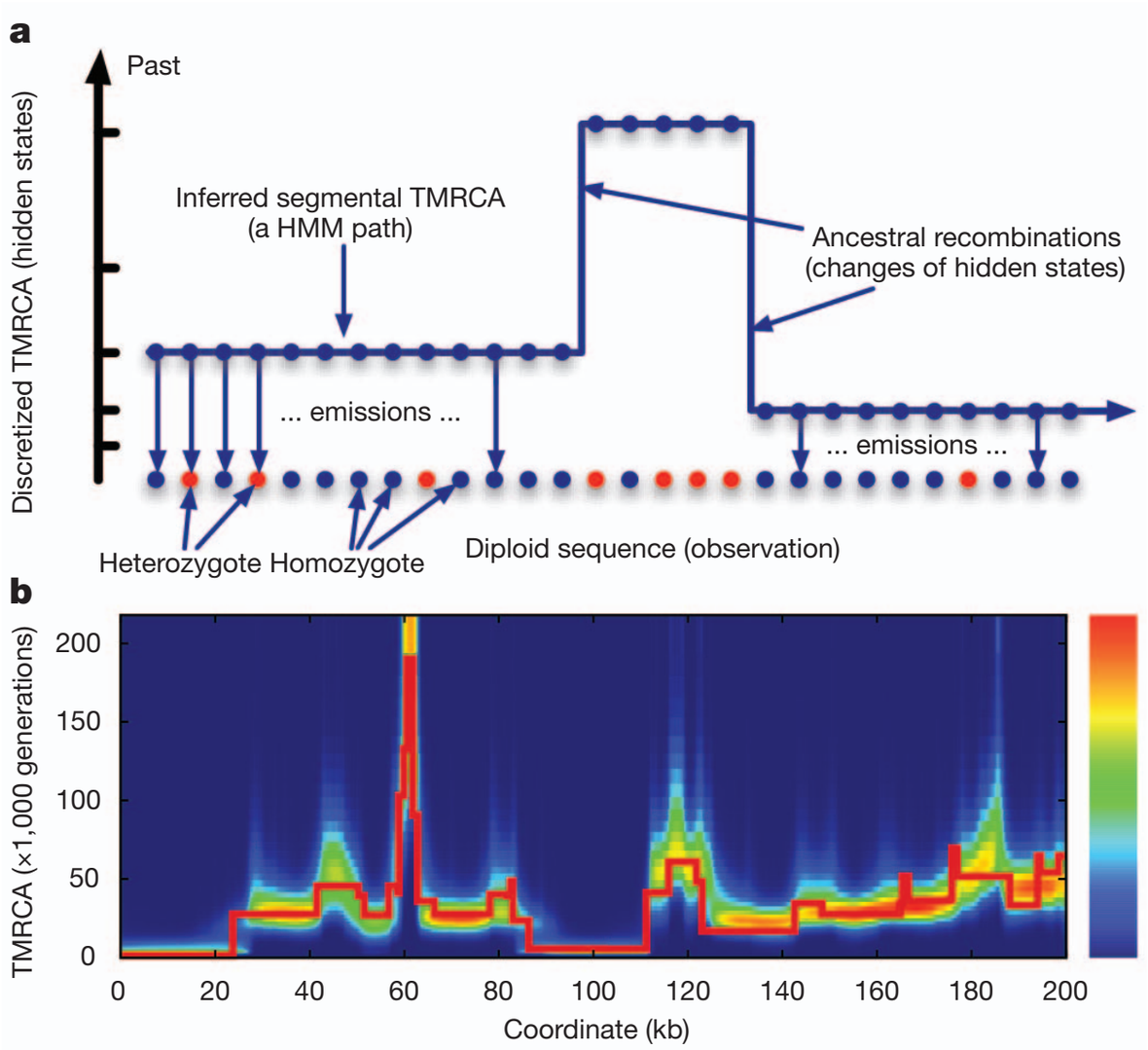
- Ancestry calls available: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140818_ancestry_deconvolution/

Migration Timing Events in Peruvians



- Gravel (2012) Population genetics models of local ancestry. Genetics 191, 607-619.
- `tracts`: https://github.com/sgravel/tracts

PSMC: Views into ancient human demography

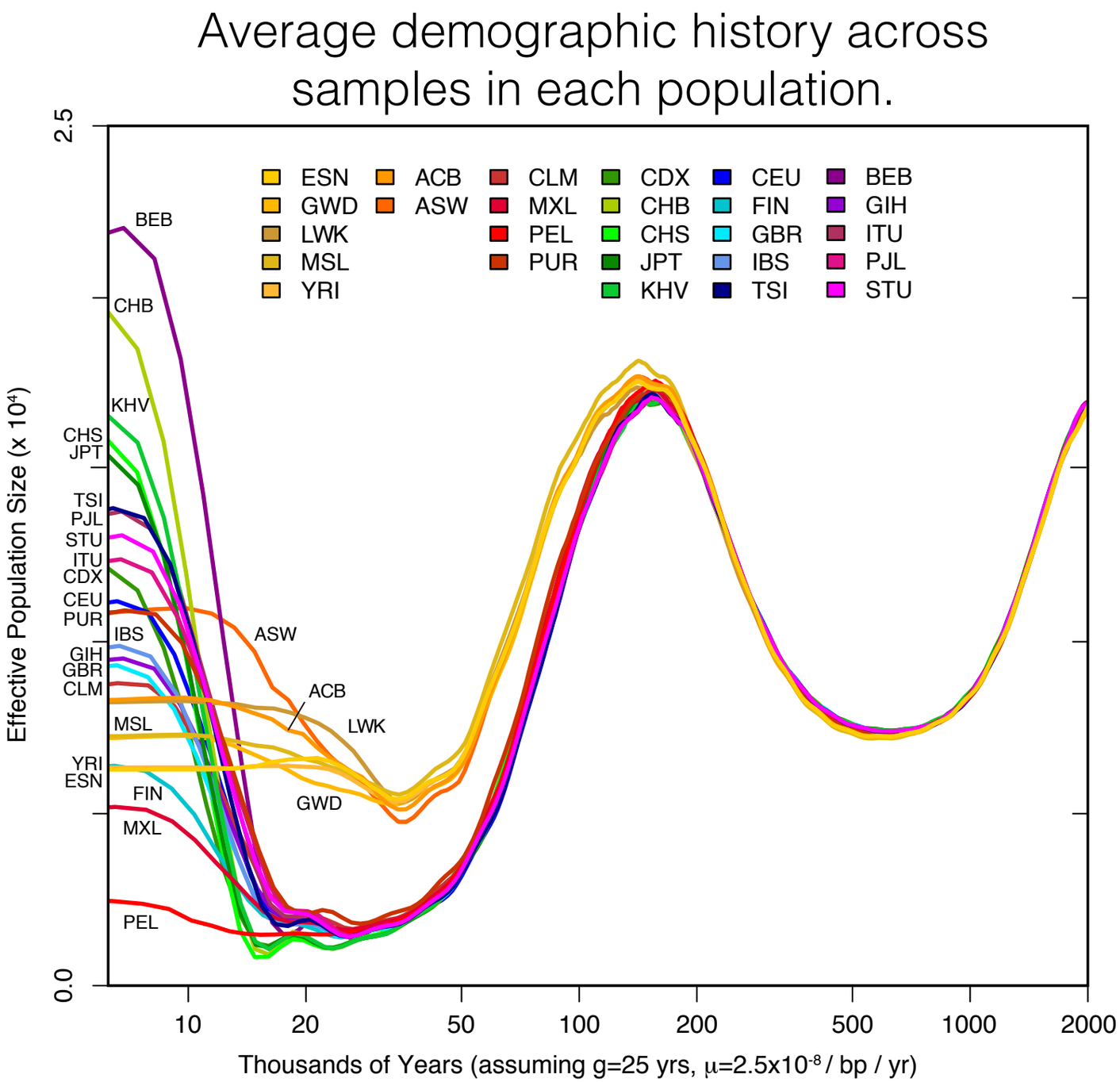


- The number of heterozygous and homozygous positions along an individual's genome is informative about historical population sizes.

- The **P**airwise **S**equential **M**arkov **C**oalescent model is a method for modeling these patterns using an HMM to infer when (and by how much) population sizes have changed throughout time.

- Li and Durbin (2011). Inference of human population history from individual whole-genome sequences. Nature 475(7357):493-6.

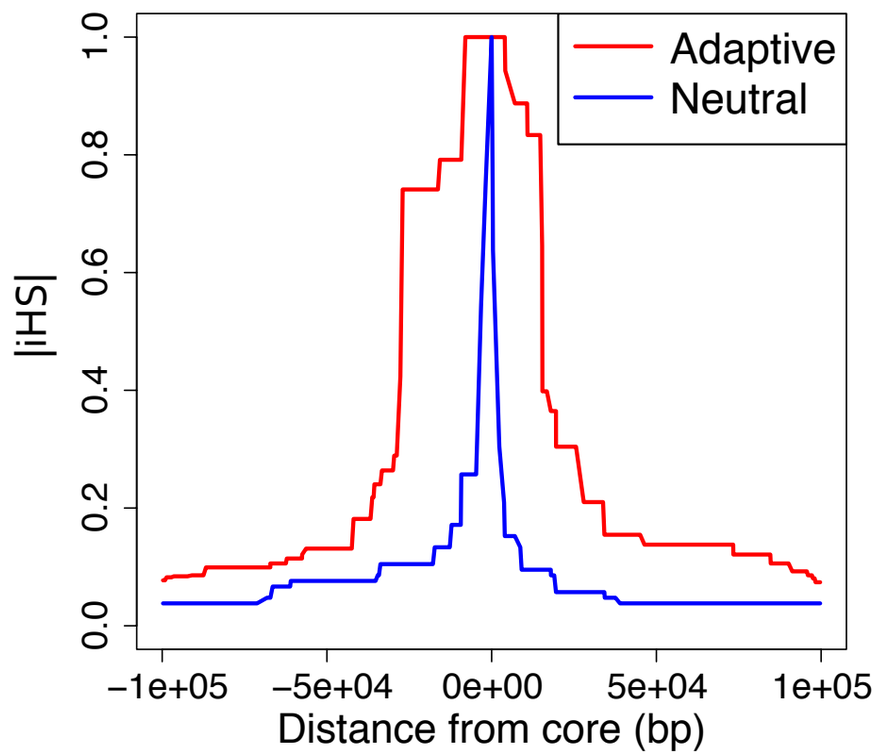
PSMC: Views into human demography



- By analyzing all 2504 samples in TGP, we are gaining deeper insight into human demographic history.

Natural Selection

- Common methods for inferring natural selection model haplotype patterns across individuals.
    - Within populations: iHS
    - Across populations: XP-EHH



MBE Advance Access published July 10, 2014

`selscan`: an efficient multi-threaded program to perform EHH-based scans for positive selection

Zachary A. Szpiech[1,*] and Ryan D. Hernandez[1,2,3]

- We developed an extremely efficient, multithreaded tool that calculates several statistics: https://github.com/szpiech/selscan

- selscan calculates:

    - Extended Haplotype Homozygosity (EHH)

    - Integrated Haplotype Score (iHS)

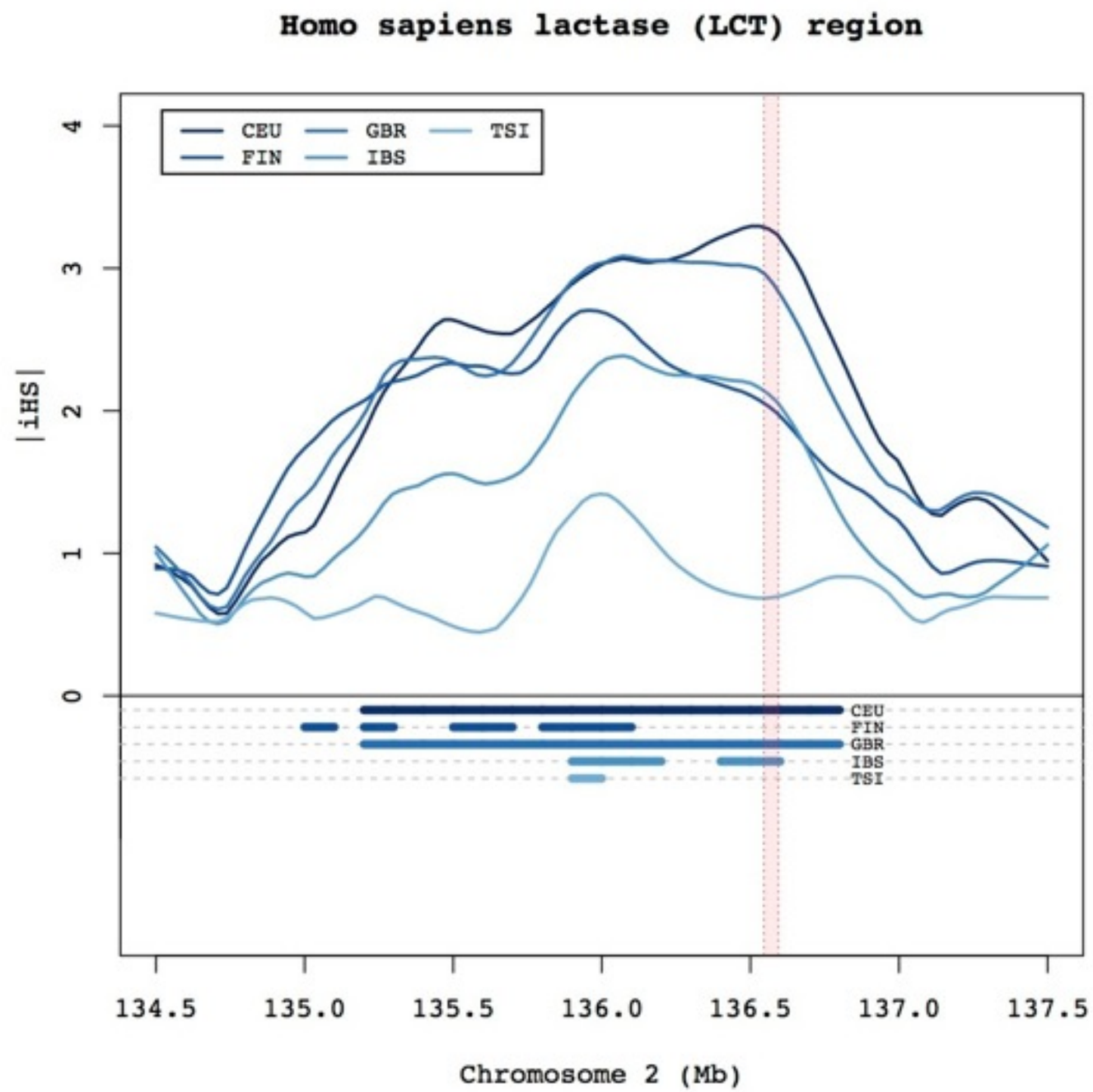
    - Cross-population EHH (XP-EHH)

    - mean pairwise sequence difference (sliding windows)

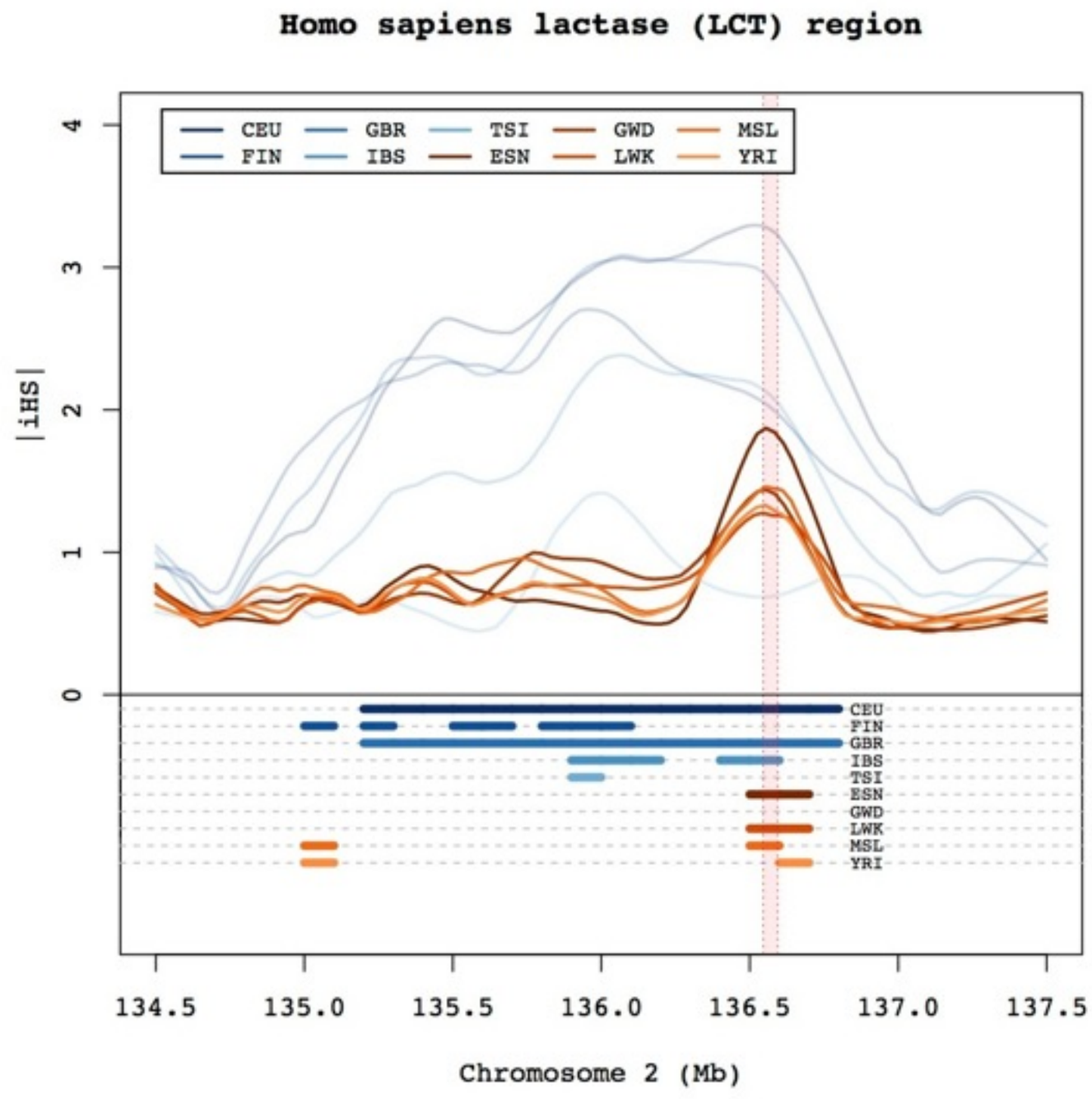
    - Also a novel method for inferring soft sweeps (coming soon).

Data Set	`ihs`	`rehh`*	`selscan` threads = 1	2	4	8	16
IHS250	19,275	563	618	306	162	84	58
IHS500	45,547	1,652	1,554	782	399	220	150
IHS1000	> 100,000	4,834	4,018	2,019	1,040	566	380
IHS2000	> 100,000	12,652	7,054	3,633	1,869	1,046	752
CEU22	19,434	588	353	182	93	50	33

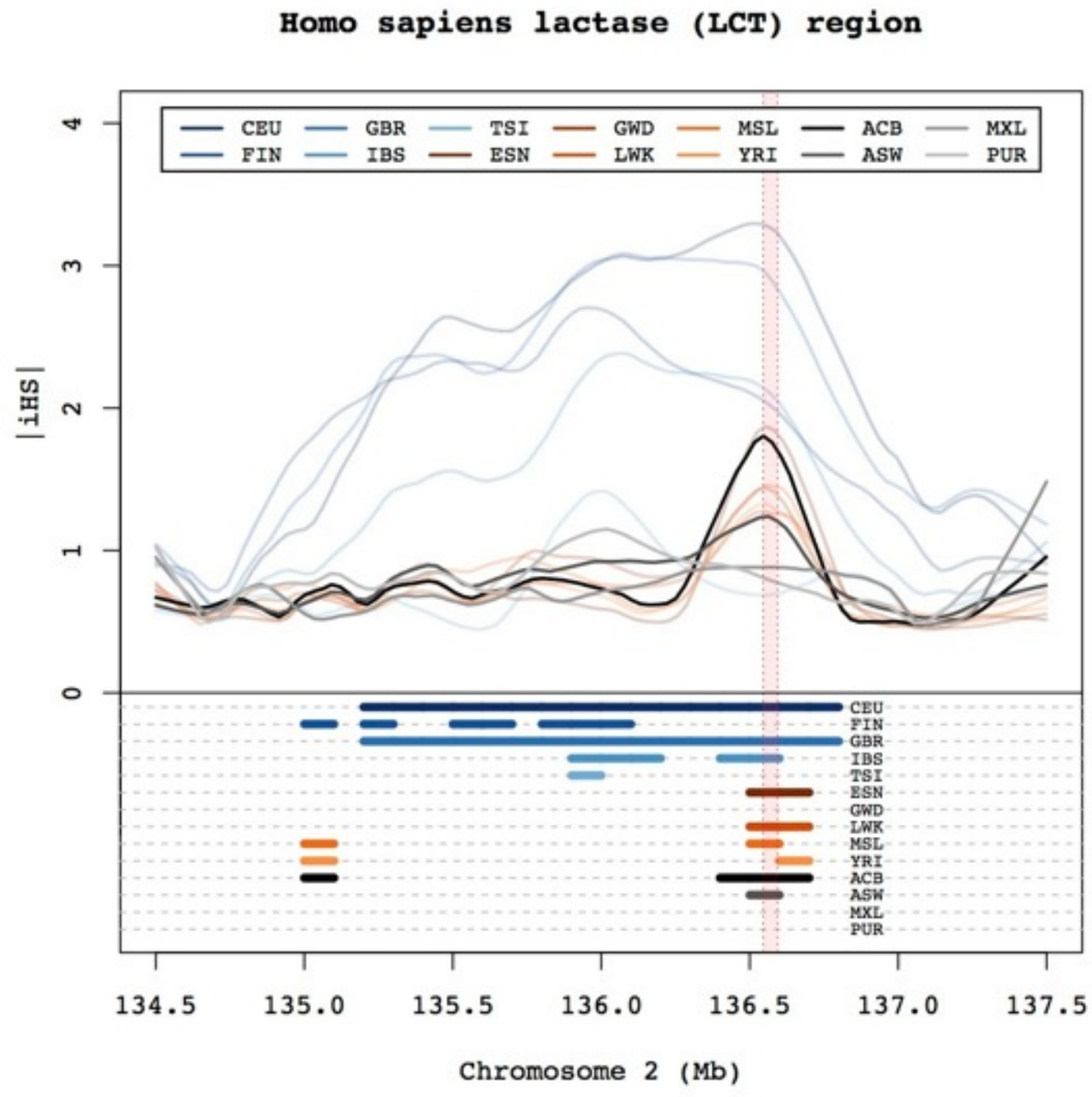
iHS around LCT



iHS around LCT

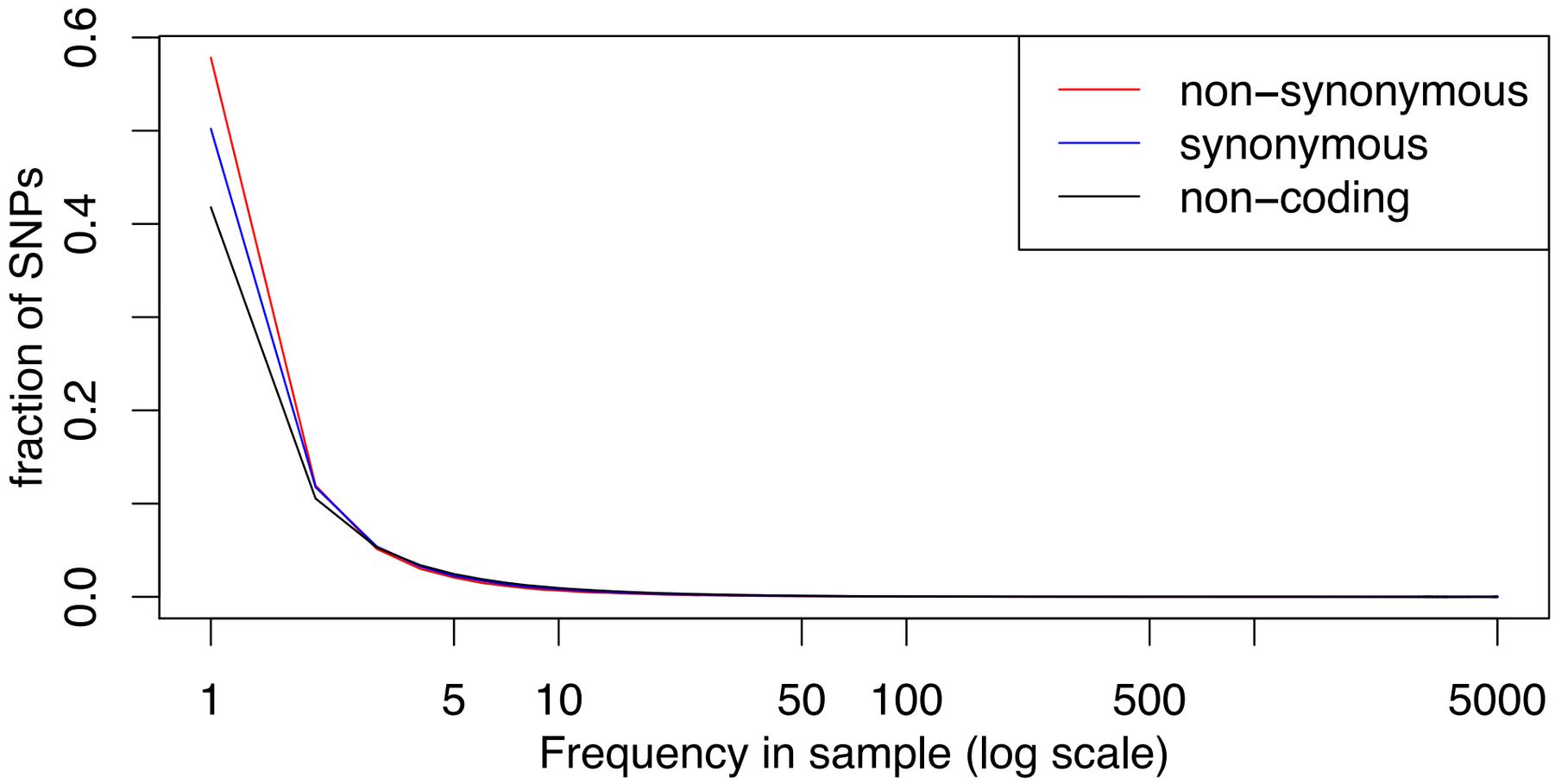


iHS around LCT



What you probably should not do

- Compare statistics between coding and non-coding regions.
- High coverage exomes vs low coverage WGS means that the patterns of diversity observed in the two regions are generally not comparable without correction factors.



Acknowledgements

- Zachary Szpiech (UCSF)
- Alicia Martin (Stanford)
- Eimear Kenny (Mt. Sanai)
- Adam Auton (Einstein)