Leveraging Local Ancestry with Admixture Mapping

Instructors: Andrey Ziyatdinov and Timothy Thornton

National Autonomous University of Mexico

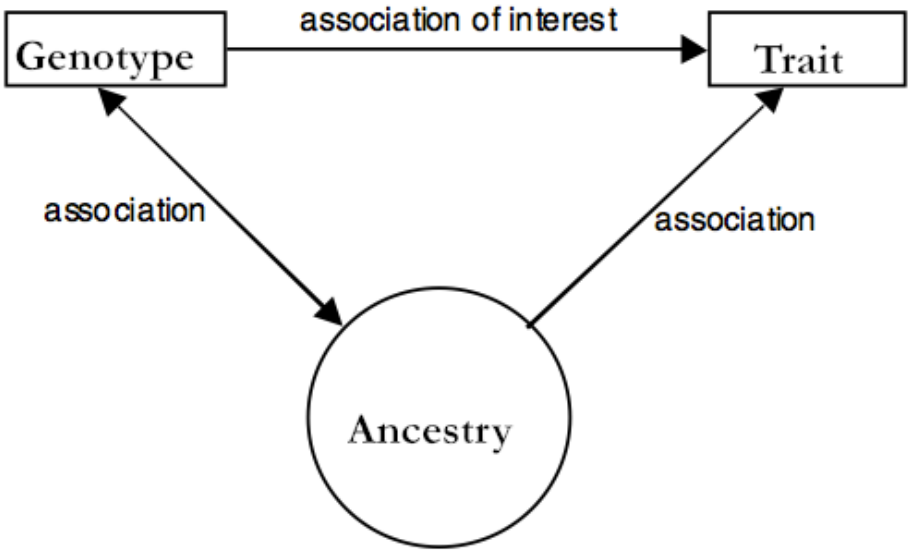
April 27, 2023

Genetic Association Studies in Multi-Ethnic Populations

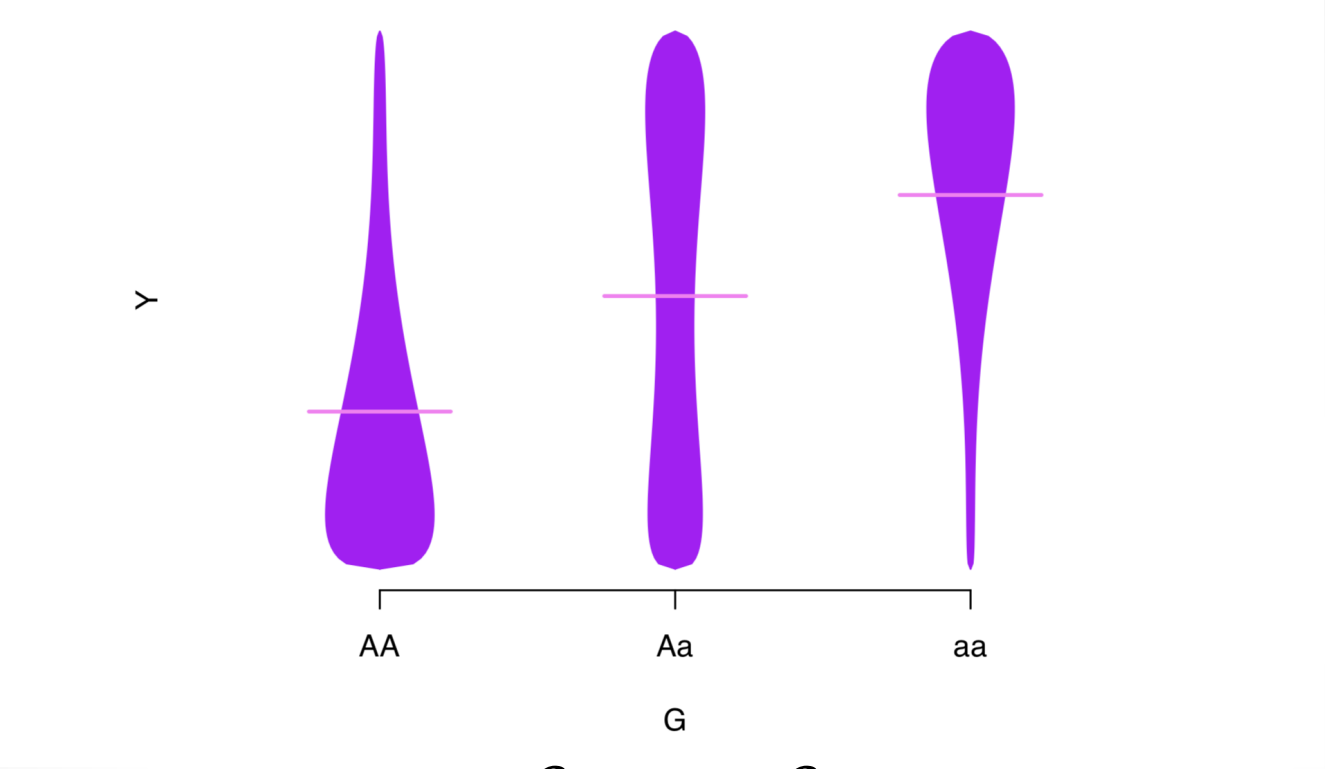
- There remain significant challenges with complex trait mapping in multi-ethnic populations
- Two well know challenges are:
  - **Heterogeneous genetic ancestry and environmental backgrounds among sampled individuals**
  - **Correlated genotype and phenotype data among relatives, known and/or cryptic, in the sample**

Confounding is a serious concerns for genetic studies in multi-ethnic populations

- Ethnic groups (and subgroups) have often share distinct dietary habits and other lifestyle characteristics that result in traits of interest having **different distributions** that are correlated with genetic ancestry and/or ethnicity.



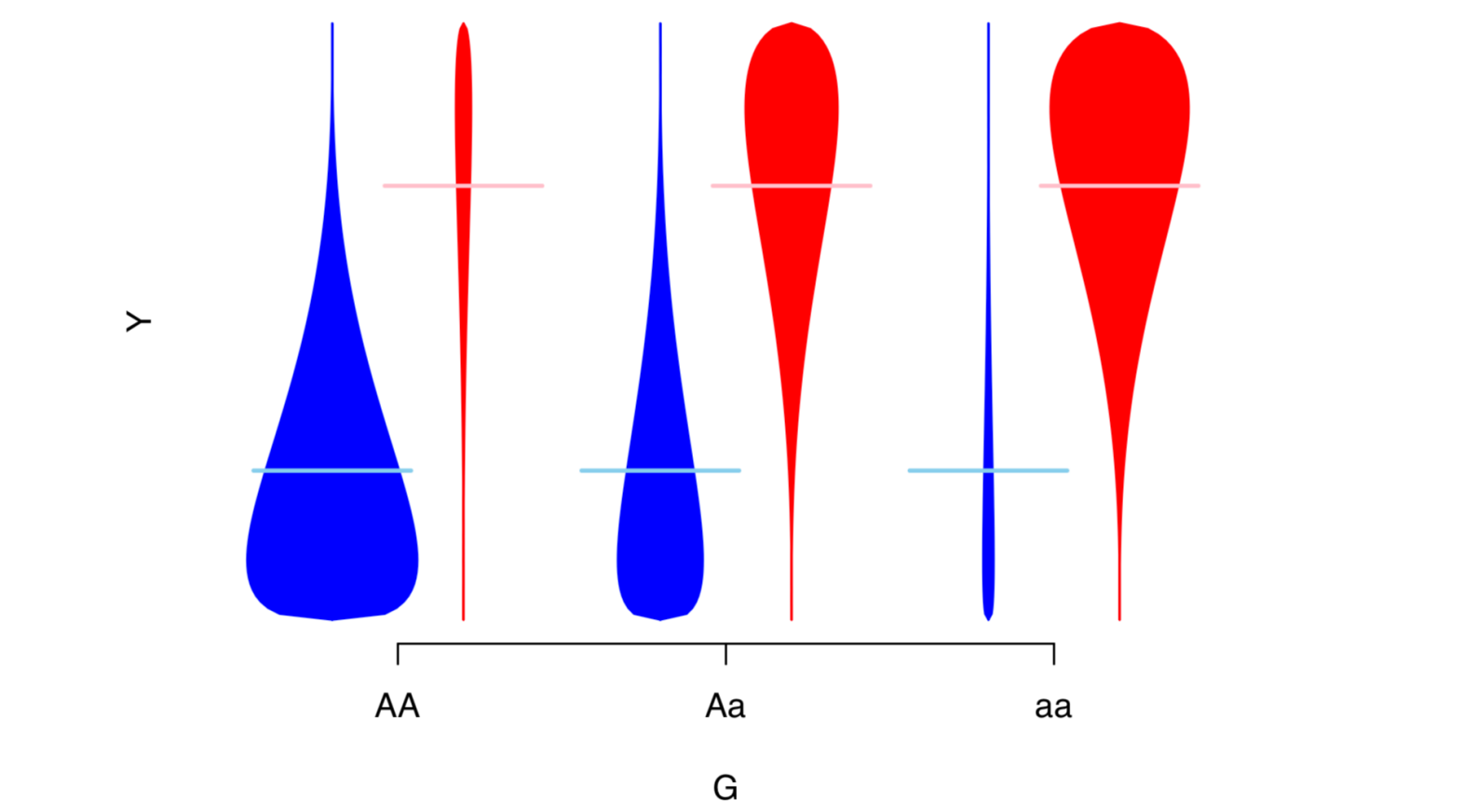
Confounding: multi-ethnic studies



$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{g}_s \beta_1 + \epsilon$$

- $\mathbf{Y}$ is a vector of phenotypes
- $\mathbf{g}_s$ is an additive genotype count vector for a SNP s, where each entry corresponds to the number of reference alleles (A) an individual has, e.g., 0, 1, or 2;
- $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma_e^2 \mathbf{I})$

Confounding: multi-ethnic studies



- The relationship between phenotype vector (**Y**) and genotype vector (**G**) looks much less interesting when broken down and assessed within ancestry groups;
- Well known that the genetic ancestry is a confounder in multi-ethnic studies that can lead to spurious associations

PCs in Regression Models

- In practice, multi-ethnic cohort studies will not have discrete or a fixed number of ancestry groups.
- Eigenvectors (PCs) are often used as surrogates for ancestry (or population structure).
- To protect against spurious association due to genetic ancestry confounding, the top PCs are often included as **fixed effects** in regression models used for assessing genotype/phenotype associations;

$$E(\mathbf{Y}) = \beta_0 \mathbf{1} + \mathbf{g_s}\beta_1 + \gamma_1 PC_1 + \gamma_2 PC_2 + \gamma_3 PC_3 + \cdots$$

i.e., regression model adjusting for PC1, PC2, PC3 etc.
(Logistic, Cox regression can be adjusted similarly)

- Among people with the same ancestry (i.e. the same PCs) β_1 gives us the difference in mean phenotype, per 1-unit difference in $\mathbf{g}_s$
- If the effect of $\mathbf{g}_s$ differs by PCs, β_1 provides a (sensible) average of these genetic effects

Can genetic ancestry be leveraged?

- Genetic association mapping methods generally treat genetic ancestry as a potential confounder
- In multi-ethnic populations that have admixed ancestry derived from multiple continents, such as African Americans and Hispanic/Latino populations, the frequency of causal genetic variants may differ substantially in the underlying ancestral populations
- Admixed individuals with causal variants may be more likely to have inherited the variants from the ancestral population for which the variants are most prevalent

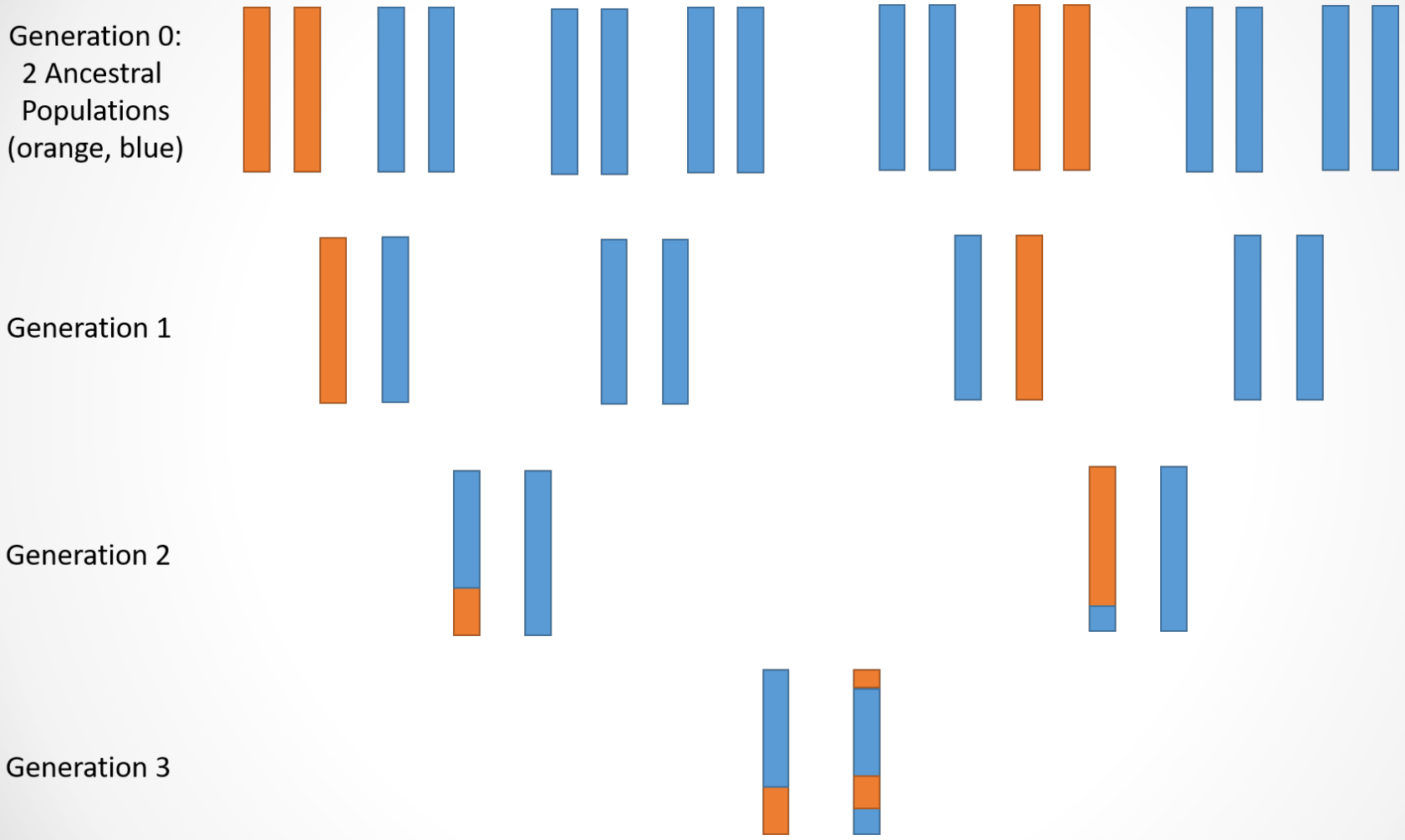
Can genetic ancestry be leveraged?

- Mapping by admixture linkage disequilibrium, or **admixture mapping**, leverages heterogeneity in genetic ancestry by identifying loci that
  - (1) have unusual deviations in local ancestry, relative to what would be expected based on genome-wide ancestry

  and

  - (2) and that are significantly associated with a trait
- By leveraging genetic ancestry, regions of the genome harboring genetic variants that differ in frequency across the ancestral populations and that drive the observed phenotypic differences
- can be identified

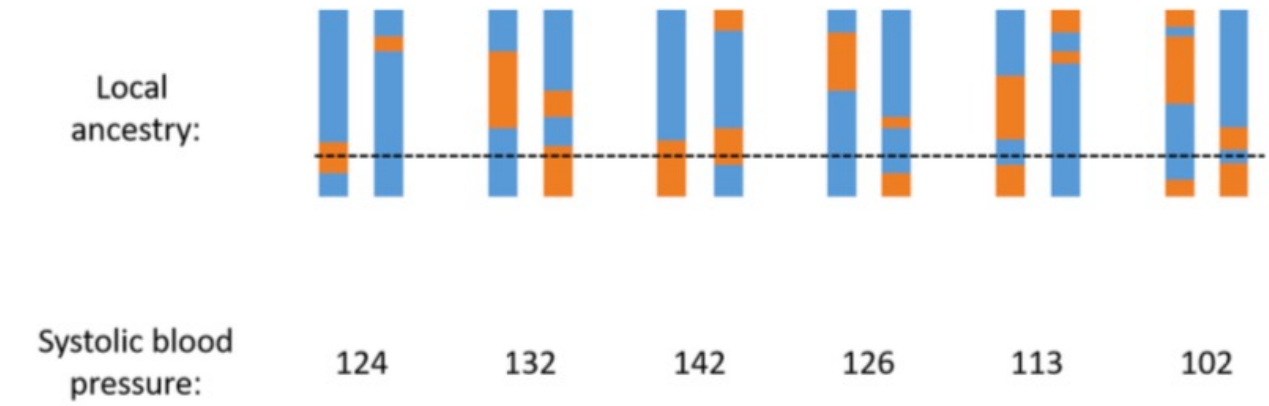
Admixed Populations



Admixture Mapping

Goal: Identify genetic variants associated with trait of interest

How? Look for associations between the trait and local ancestry



- Previously proposed regression approach for admixture with two ancestral populations:

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{a}_{\ell p}\beta_{\ell p} + \mathbf{V}\gamma + \epsilon$$

  - $\mathbf{a}_{\ell p}$ is a local ancestry vector with entries consisting of the number of alleles (0,1,2) inherited from ancestral population *p* at locus ℓ for each individual
  - $\mathbf{V}$ is a vector of covariates for genome-wide average ancestry (e.g., admixture proportions or PCs)

Admixture Mapping with Multiple Ancestral Populations

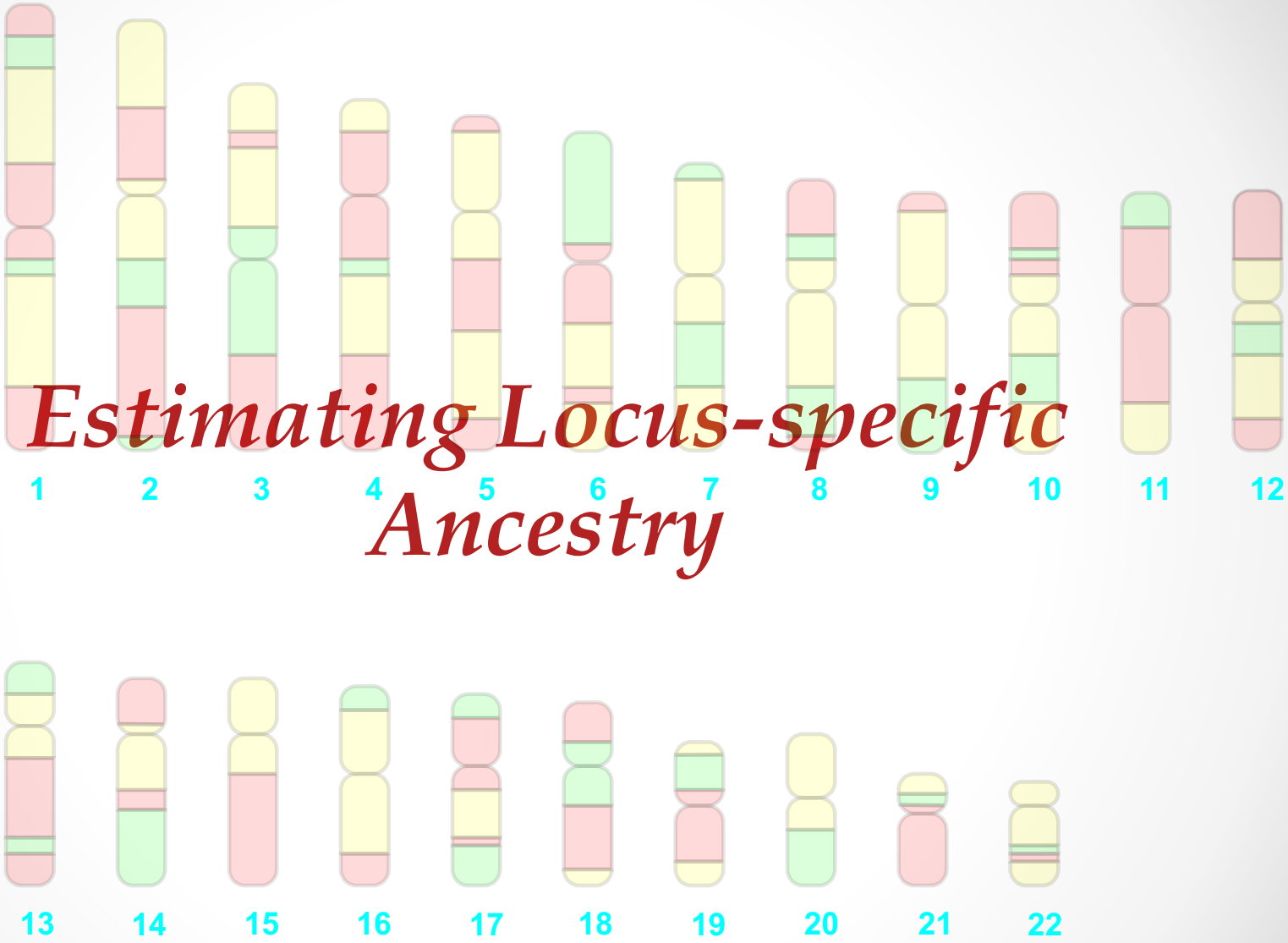
- Proposed admixture mapping model:

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{a}_{\ell p} \beta_{\ell p} + \mathbf{a}_{\ell p'} \beta_{\ell p'} + \mathbf{V}\gamma + \mathbf{r} + \boldsymbol{\epsilon}$$

- To test for an association with local ancestry at a locus ℓ, we propose using a score test to test the null hypothesis

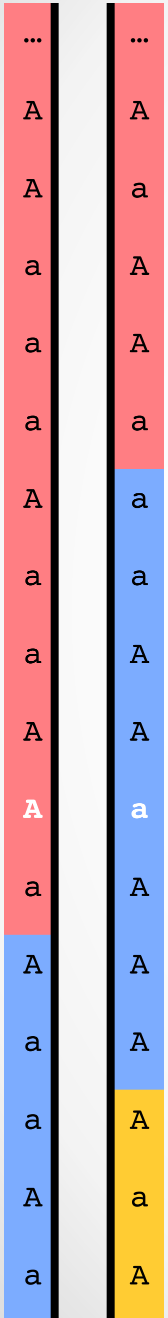
$$H_0 : \beta_{\ell p} = \beta_{\ell p'} = 0$$

Estimating Locus-specific Ancestry



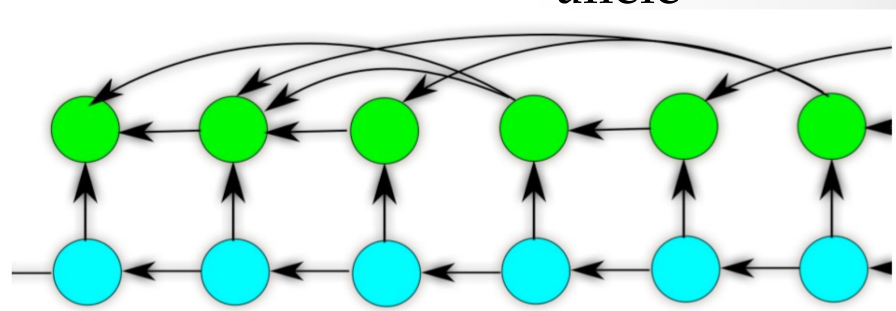
Data structure

Full information



Observed Genotype Data (no phase!)

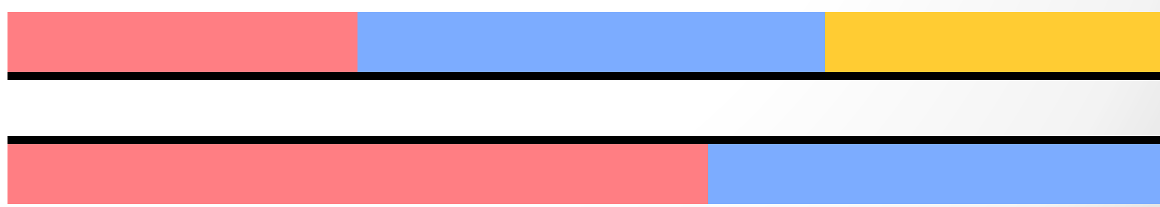
… 2 1 1 1 0 1 0 1 2 1 1 2 1 1 1 1 1 2 …



allele

ancestry

inferred



African
European
Native American

RFMix Method for Local Ancestry Inference

ARTICLE

RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference

Brian K. Maples,[1,2] Simon Gravel,[1,3] Eimear E. Kenny,[1,4,5,6,7,8] and Carlos D. Bustamante[1,8,*]

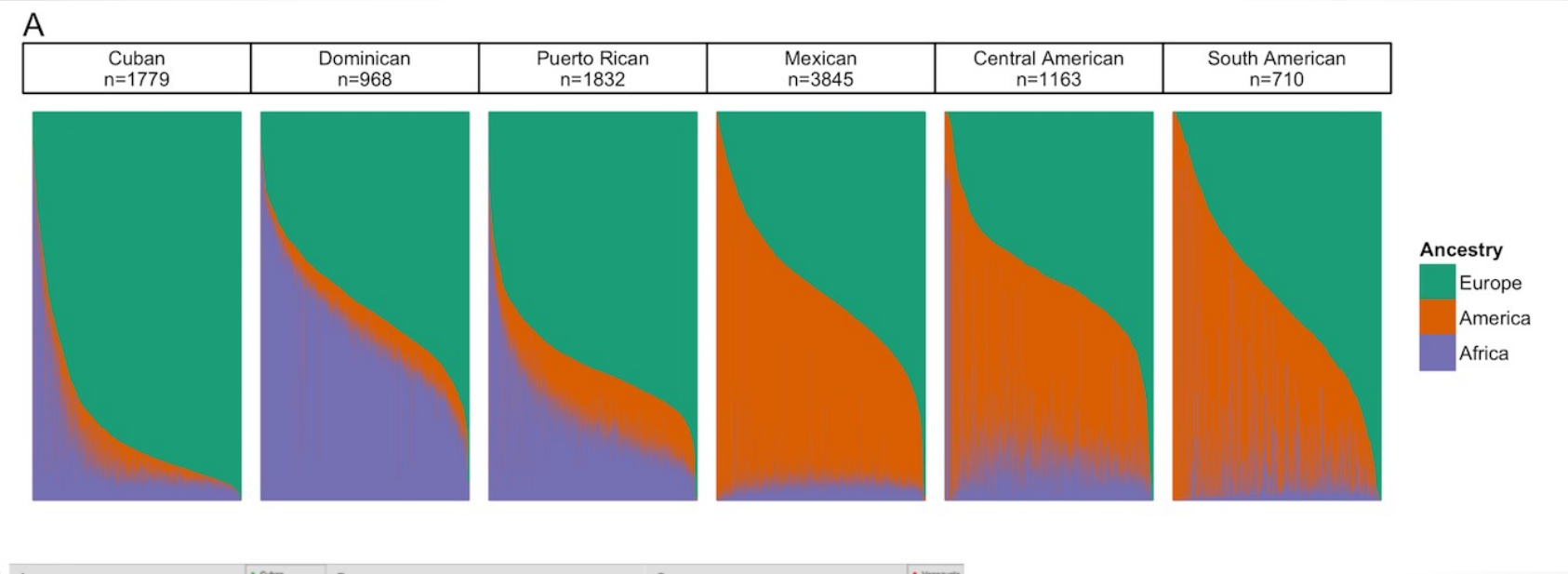
- RFMix (Maples et al.; AJHG 2013) continues to be one of the most widely used software programs as it can accurately infer local ancestry for admixed samples derived from multiple ancestral populations

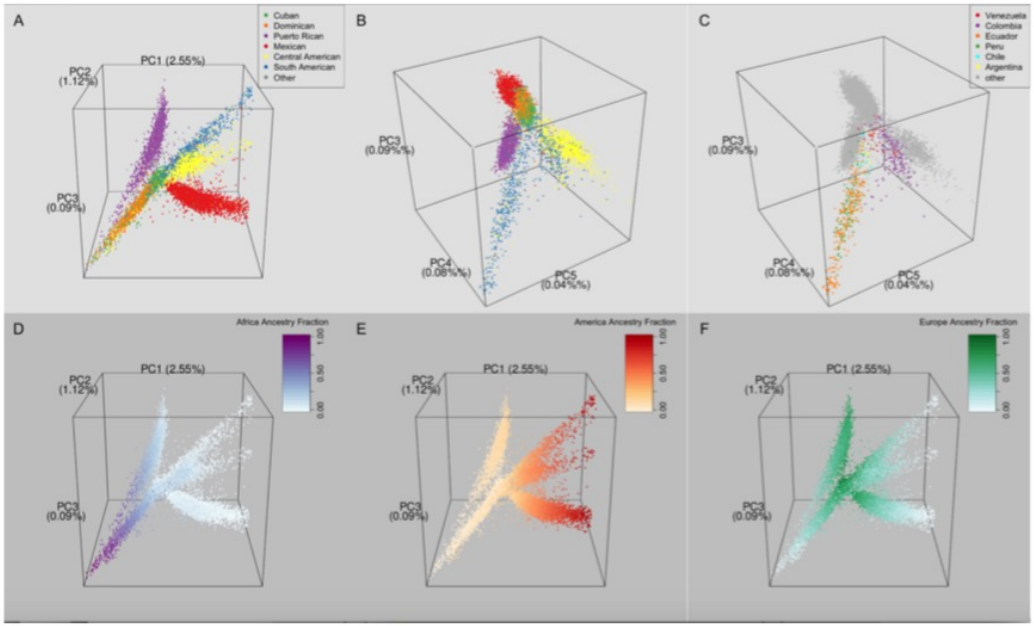
Application to HCHS/SOL

Hispanic Community Health Study/Study of Latinos

- Study Subjects
  - ~13,000 participants who self-identified as Hispanic/Latino
  - randomly selected households from four US cities
- Local ancestry estimation
  - performed at 236,456 SNPs with RFMix

HCHS/SOL Ancestry





ARTICLE

Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

Matthew P. Conomos,[1,14,*] Cecelia A. Laurie,[1,14] Adrienne M. Stilp,[1,14] Stephanie M. Gogarten,[1,14] Caitlin P. McHugh,[1] Sarah C. Nelson,[1] Tamar Sofer,[1] Lindsay Fernández-Rhodes,[2] Anne E. Justice,[2] Mariaelisa Graff,[2] Kristin L. Young,[2] Amanda A. Seyerle,[2] Christy L. Avery,[2] Kent D. Taylor,[3] Jerome I. Rotter,[3] Gregory A. Talavera,[4] Martha L. Daviglus,[5] Sylvia Wassertheil-Smoller,[6] Neil Schneiderman,[7] Gerardo Heiss,[2] Robert C. Kaplan,[6] Nora Franceschini,[2] Alex P. Reiner,[8] John R. Shaffer,[9] R. Graham Barr,[10] Kathleen F. Kerr,[1] Sharon R. Browning,[1] Brian L. Browning,[11] Bruce S. Weir,[1] M. Larissa Avilés-Santa,[12] George J. Papanicolaou,[12] Thomas Lumley,[13] Adam A. Szpiro,[1] Kari E. North,[2] Ken Rice,[1] Timothy A. Thornton,[1] and Cathy C. Laurie[1,*]

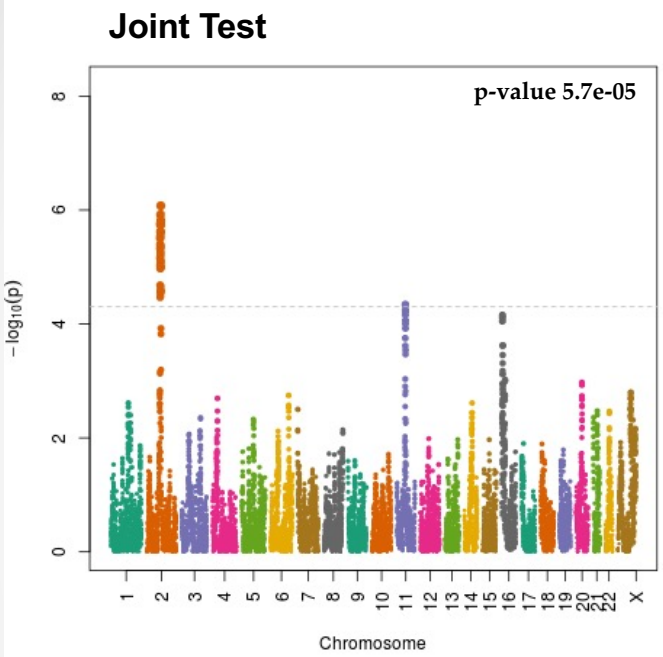
HCHS/SOL: Admixture Mapping of Kidney Traits

Genome-Wide Admixture Mapping of Estimated Glomerular Filtration Rate and Chronic Kidney Disease Identifies European and African Ancestry-of-Origin Loci in Hispanic and Latino Individuals in the United States

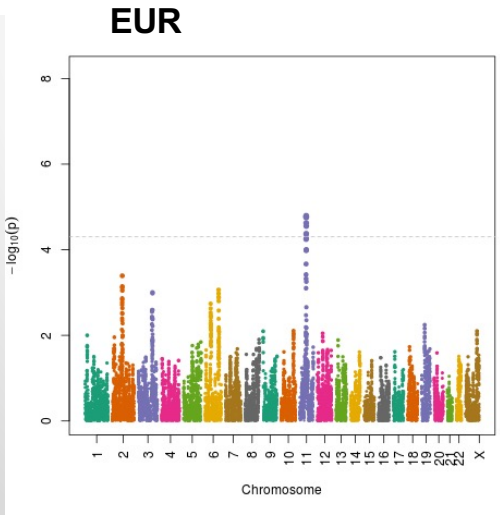
Andrea R.V.R. Horimoto,[1] Diane Xue,[2] Jianwen Cai,[3] James P. Lash,[4] Martha L. Daviglus,[5] Nora Franceschini,[6] and Timothy A. Thornton[1,7]

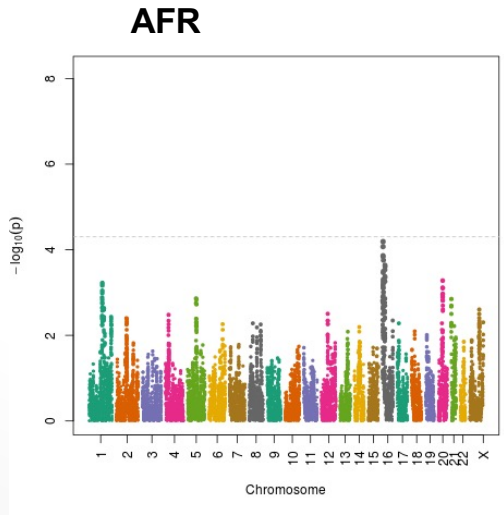
- Horimoto et al. (JASN; 2022)
- Analysis model
  - linear mixed model for both association and admixture mapping
  - adjust for PCs1-5, genetic background group, sex, age, center, smoking status
- admixture mapping significance threshold: 6e-5

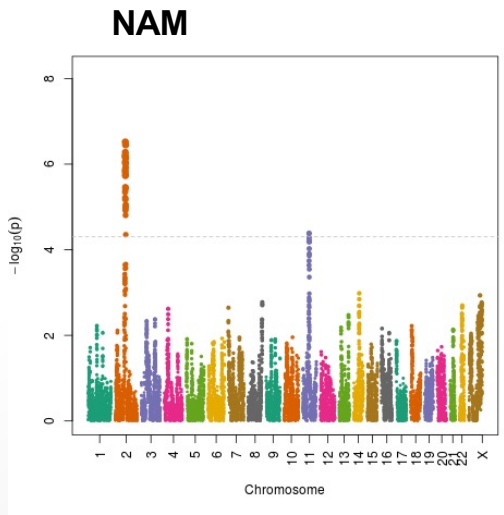
HCHS/SOL: Albumin-to-creatinine ratio Admixture Mapping



chr	# SNPs	pos	p-value	gene
2	1 (rs2139376)	112143413	8.4e-07	MIR4435-2HG (2q13)
11	13	69533129 - 69579070	4.5e-05	intergenic region
16	8	4964076 - 4999392	7.0e-05	PPL (16p13.3)



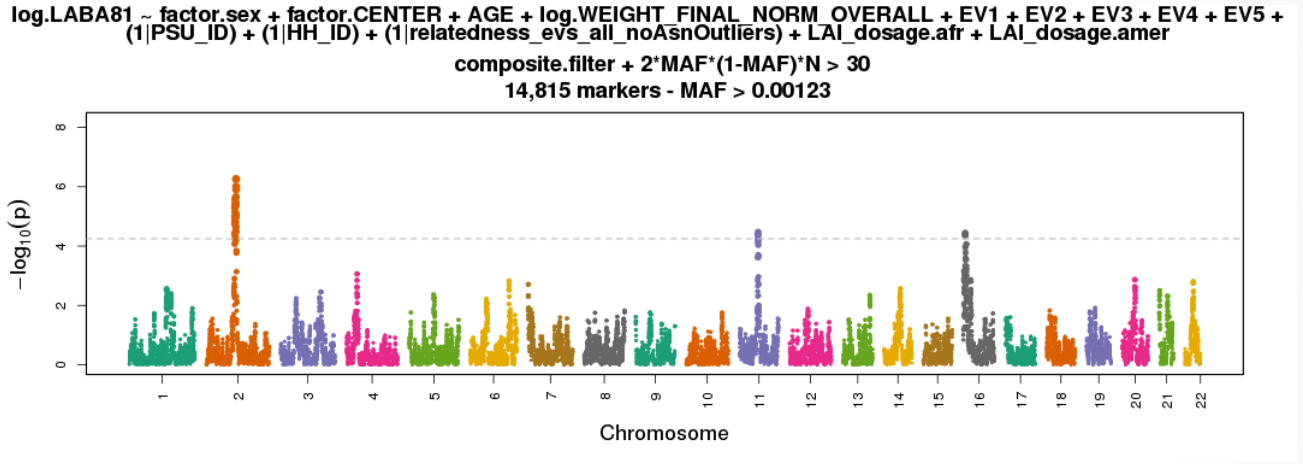




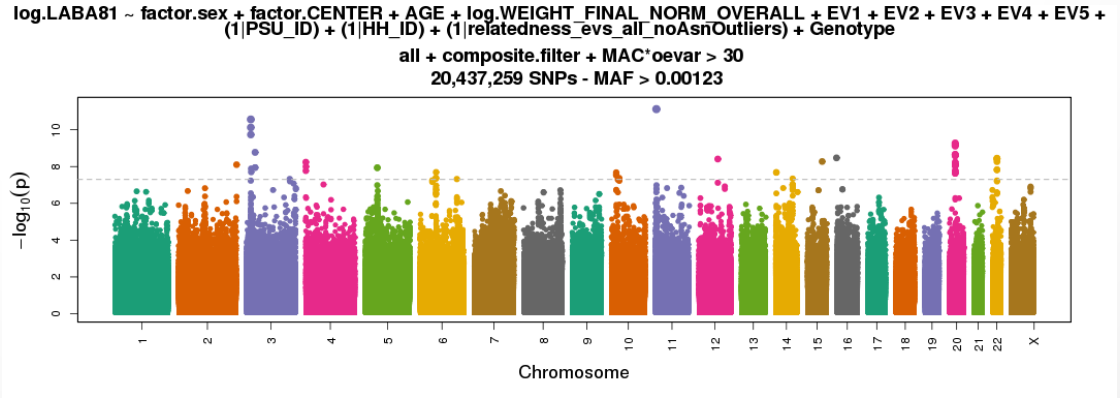
HCHS/SOL: Admixture and Association Mapping

Phenotype: Albumin-to-creatinine ratio

Admixture Mapping



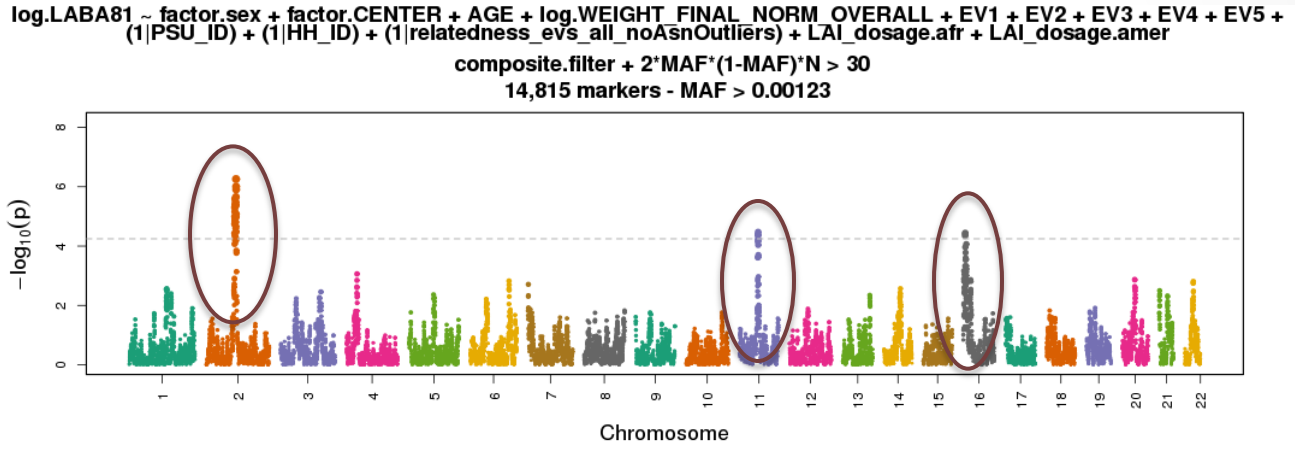
Association Mapping



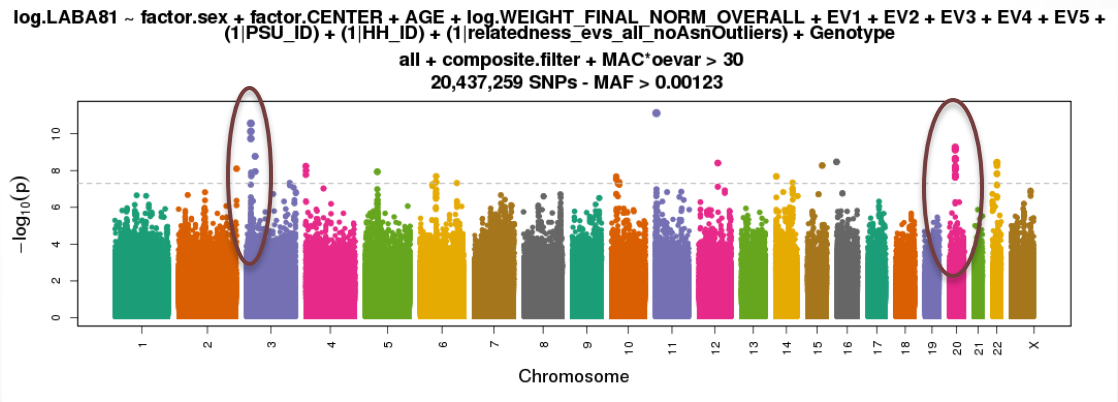
HCHS/SOL: Admixture and Association Mapping

Phenotype: Albumin-to-creatinine ratio

Admixture Mapping



Association Mapping

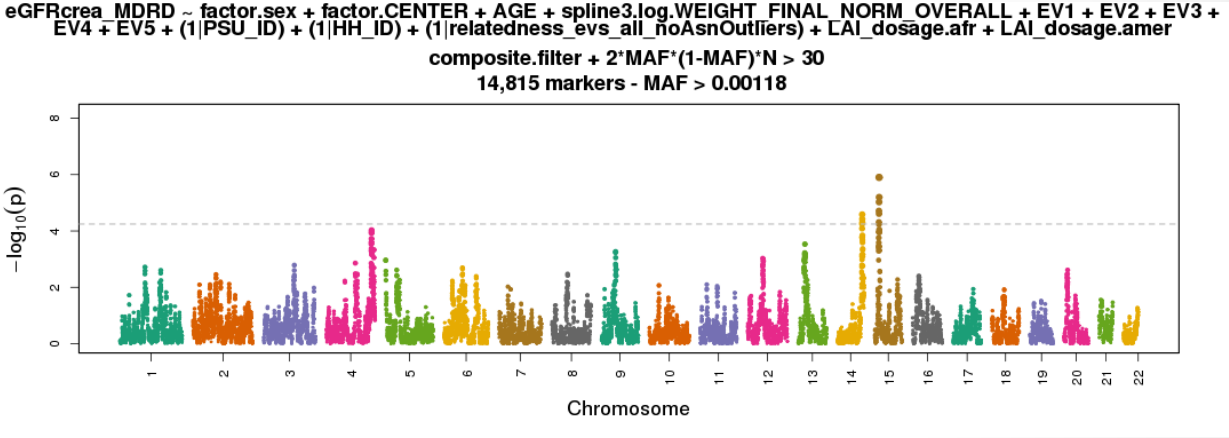


Different signals for the two methods

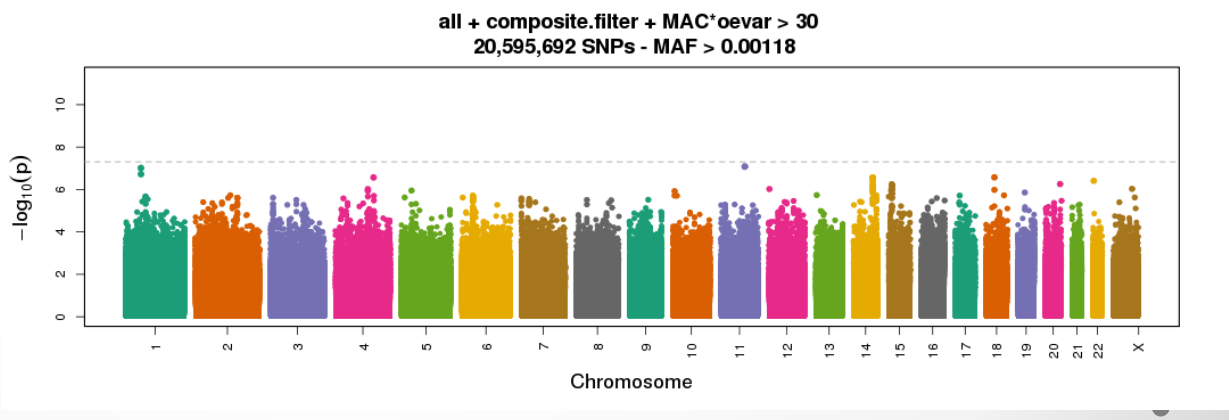
HCHS/SOL: Admixture and Association Mapping

Phenotype: Estimated Glomerular Filtration rate

Admixture Mapping



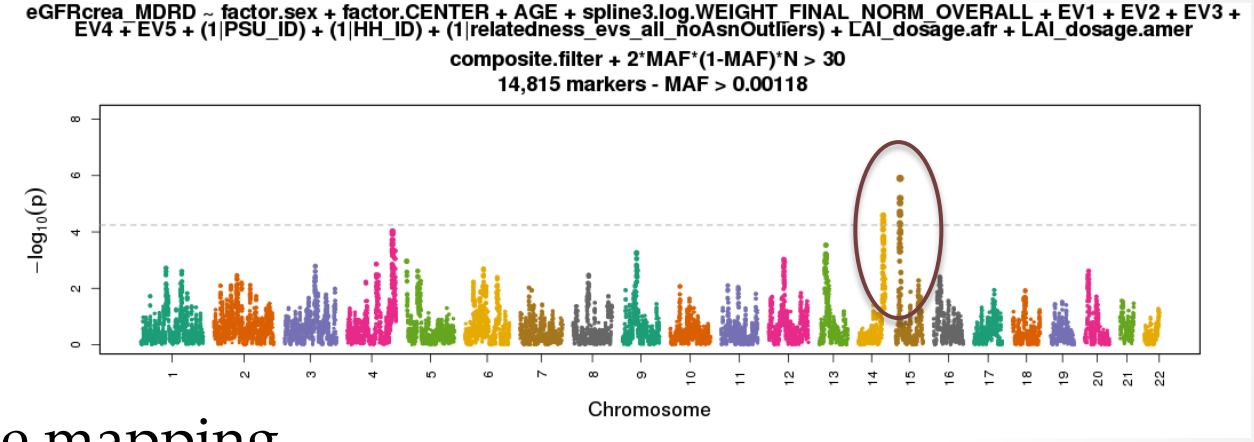
Association Mapping



HCHS/SOL: Admixture and Association Mapping

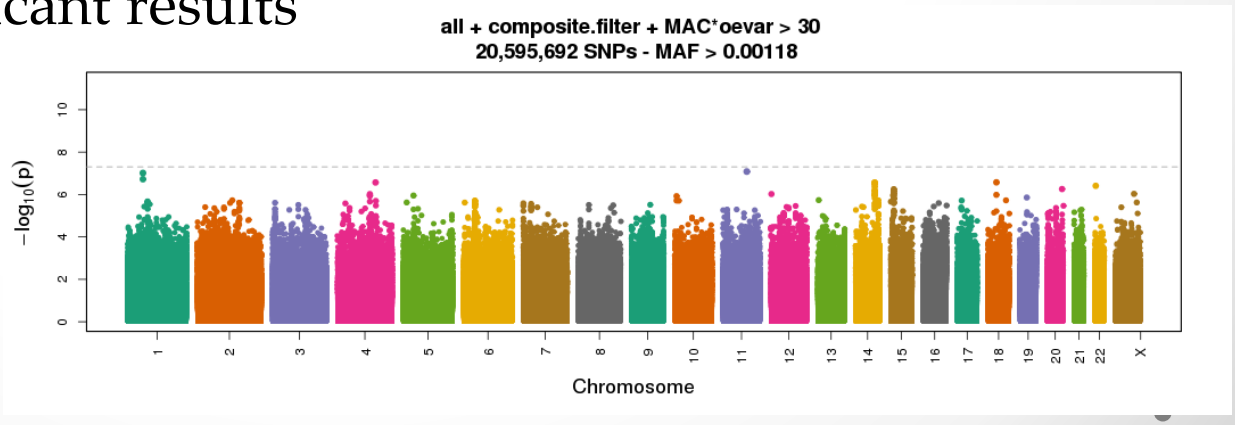
Phenotype: Estimated Glomerular Filtration rate

Admixture Mapping



Only admixture mapping provides significant results

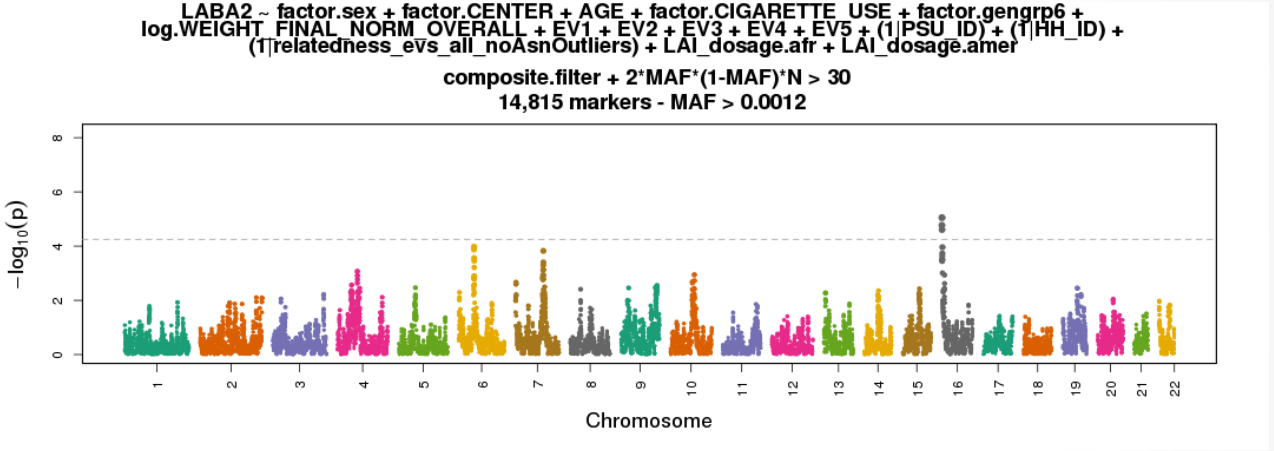
Association Mapping



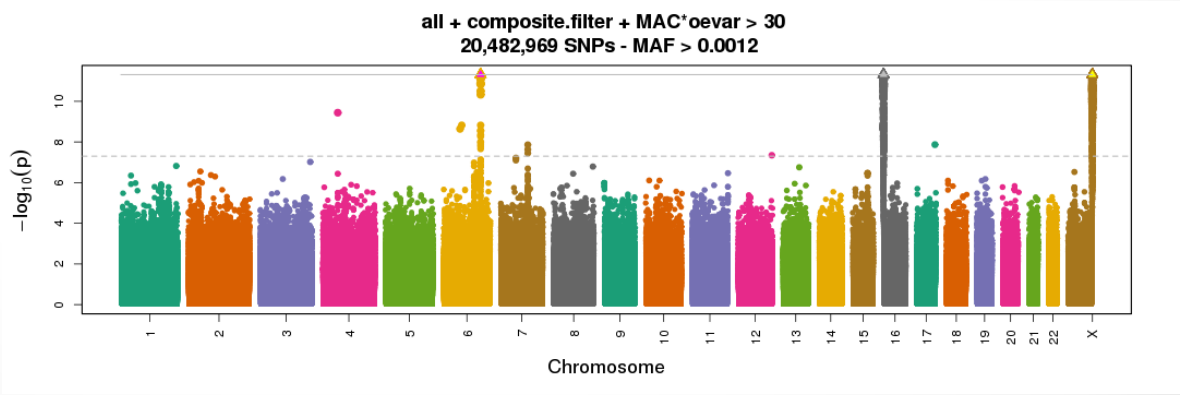
HCHS/SOL: Admixture and Association Mapping

Phenotype: Red blood cell count

Admixture Mapping



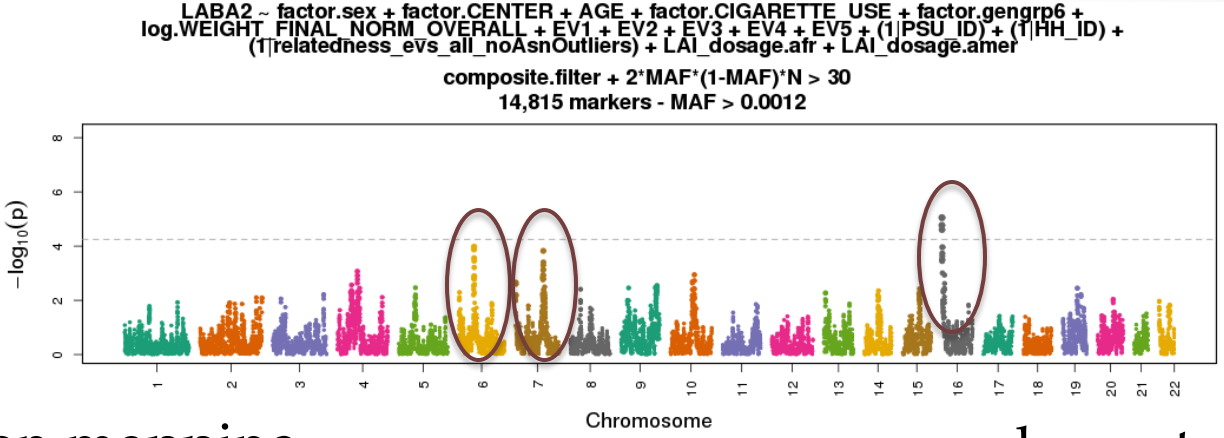
Association Mapping



HCHS/SOL: Admixture and Association Mapping

Phenotype: Red blood cell count

Admixture Mapping



association mapping gives more significant results

complementary results

Association Mapping