Genetic data analyses for admixed and multiethnic samples

Seth D. Temple
Sharon R. Browning

BIOST 550
Seattle, WA, USA
May 25-27, 2022

UNIVERSITY of WASHINGTON

Agenda

1. Motivation
   - 1.1 Genetic drift
   - 1.2 Equity in research
2. Global ancestry
   - 2.1 Principal component analysis
   - 2.2 Multidimensional scaling
   - 2.3 Primer: Bayesian data analysis
   - 2.4 *structure* and other methods
3. Local ancestry
   - 3.1 RFMix and other methods
   - 3.2 Admixture mapping

Genetic drift

- Gene pool for populations follows different random trajectory after a population split, e.g. “Out of Africa”
- Some variants drift to 0 (loss) or 1 (fixation) frequency
- Some variants become private to a population
- In general,
  - Small population size $\rightarrow$ more drift
  - Longer split time $\rightarrow$ more drift
  - Migration back and forth $\rightarrow$ less drift

- Consequence: conclusions in one population may not be portable to another population

Other considerations

From Oni-Orisan, et al. (2021), African populations have:

- Greatest genetic diversity
- Smallest LD blocks
- Largest number of population-specific alleles
- Lowest proportion of recent deleterious variants
- Most extensive population structure
- Deepest historical lineage

Consequence: conclusions in one population may not be portable to another population

Contribute to EC discussion to engage w/ this paper more.

Other considerations

Most indigenous populations have some admixture.

- Migration brings populations in contact.
- Exchange of genetic material.
- Allele frequencies depend on haplotype background.
- Ancestry switches occur along chromosomes.

Consequence:

- Many methods assume homogeneous population(s)
- Admixed sample is homogeneous in blocks

Equity in genetics research

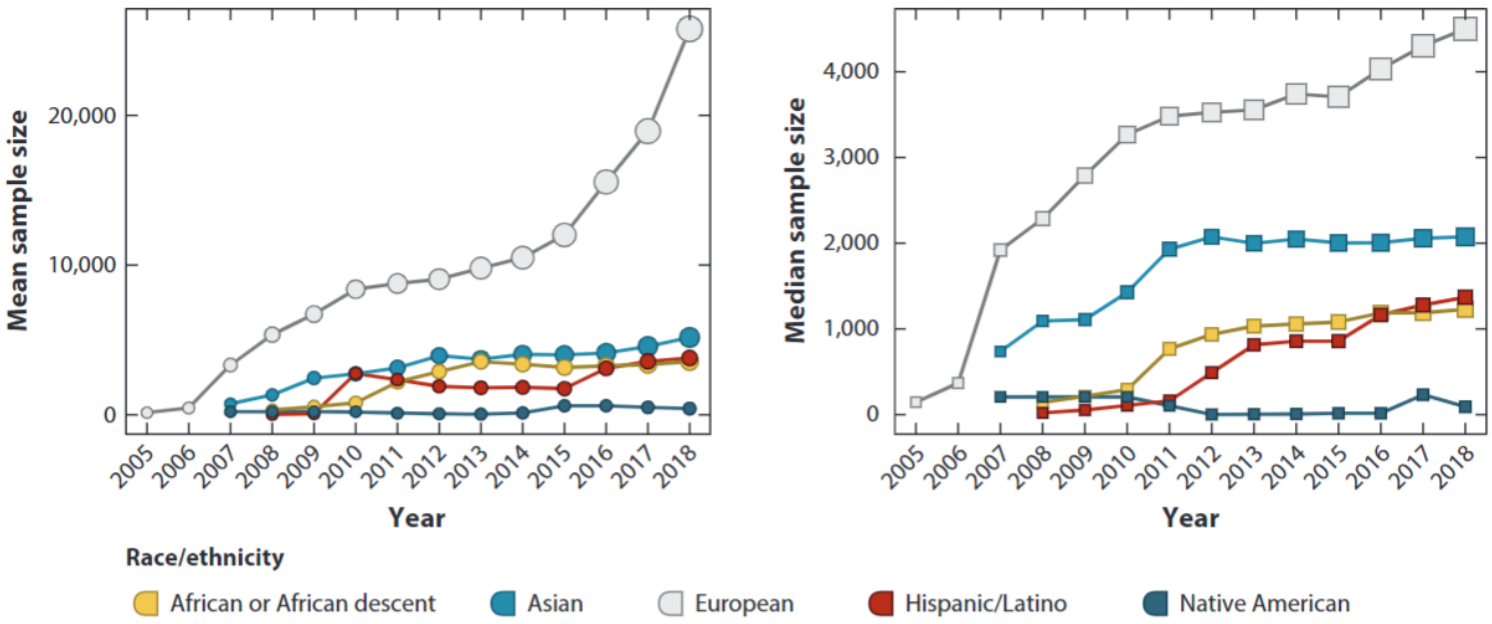

Figure: Mean and median sample sizes for different GWAS in the NHGRI GWAS catalog from 2005-2018.

Equity in genetics research

Problems:

- Some clinical results only benefit one population
- Some clinical tools (PRS, etc.) may be harmful to groups
- Some clinical phenotypes may differ among populations (e.g., dementia in non-white populations)
- Some groups may provide evidence to further disease etiology and our understanding of human genomes

Discussion: please share any additional problems you see with regard to equity in clinical and genetics research.

Equity in genetics research

Resources:

- Journal club on diversity in genetics
- An array for multiethnic samples
- Embracing genetic diversity to improve Black health
- Malaspinas, Anna Sapfo, Michael C. Westaway, Craig Muller, Vitor C. Sousa, Oscar Lao, Isabel Alves, Anders Bergström, et al. 2016. “A Genomic History of Aboriginal Australia.” Nature 538 (7624): 207–14.

Discussion: please share any additional resources you know.

Agenda

1. Motivation
   - 1.1 Genetic drift
   - 1.2 Equity in genetics research
2. Global ancestry
   - 2.1 Principal component analysis
   - 2.2 Multidimensional scaling
3. Local ancestry
   - 3.1 Primer: Bayesian data analysis
   - 3.2 *structure* and other methods
   - 3.3 RFMix and other methods
4. Admixture mapping

Global ancestry

Global refers to ancestry composition of entire genome

- E.g., an African-American individual may have 70% West African ancestry and 30% European ancestry
- Here "ancestry" means similarity with ref. pop.

Motivation: linear modeling

- Model says $\mathbb{E}[g(y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$
  - Link $g(\cdot)$ is usually identity $g(y) = y$ or log-odds
- Interpretation: β_1 is average effect of increment +1 in x_1, holding $x_2, \ldots$ fixed
  - Include data $x_2, \ldots$ so as to give stronger evidence to association signal between y and genetic marker x_1

Genetic data on diverse samples

- Hard to identify population groups with a few loci, even if population-specific allele frequencies are distinct
- Arrays contain thousands of genetic markers
- Sequences contain millions of genetic markers
- Use **dimension reduction** methods to cluster samples, aggregating over small signals of population-specific allele frequencies
  - Largest signal is African versus non-African
  - Second signal is Asian and American versus Europe
  - Finer signals could be large inversion regions, relatives
    - Caution: do not over-interpret your results

Principal component analysis

Goal: find an orthogonal (perpendicular) matrix U that determines a change of variables $Z = XU$ such that

- Columns of Z are ordered by increasing variability
- Columns of Z are uncorrelated
- (Normalized) $\sum_{i=1}^{p} u_{ji}^2 = 1$ for each column of U

Keep q much less than p of the U, Z columns

- Columns of U are loadings
- Columns of Z are PCs

Interpretation: loadings define **direction of most variation** in a high dimensional (vector) space

Principal components analysis

Why does PCA make sense for genetic data analysis?

- Genetic markers are correlated
- Many genetic markers
- Efficient matrix calculations
- PC is linear combination of loadings,
  so interpretation is possible
- Promotes visual/graphical diagnostics

Interpretation:

- Direction of most variation is African versus non-African
- Direction of second most variation is Asian and American versus European

PCA: Bryc, et al. (2010)

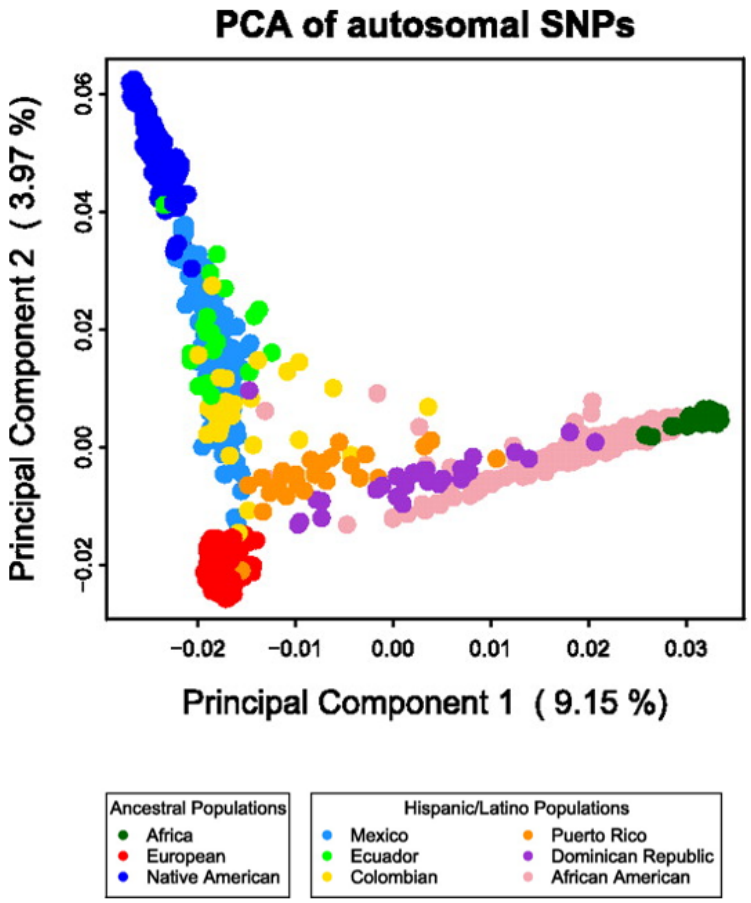

Figure: PC 1 by PC 2 for labeled Hispanic/Latino populations

PCA: Novembre, et al. (2008)

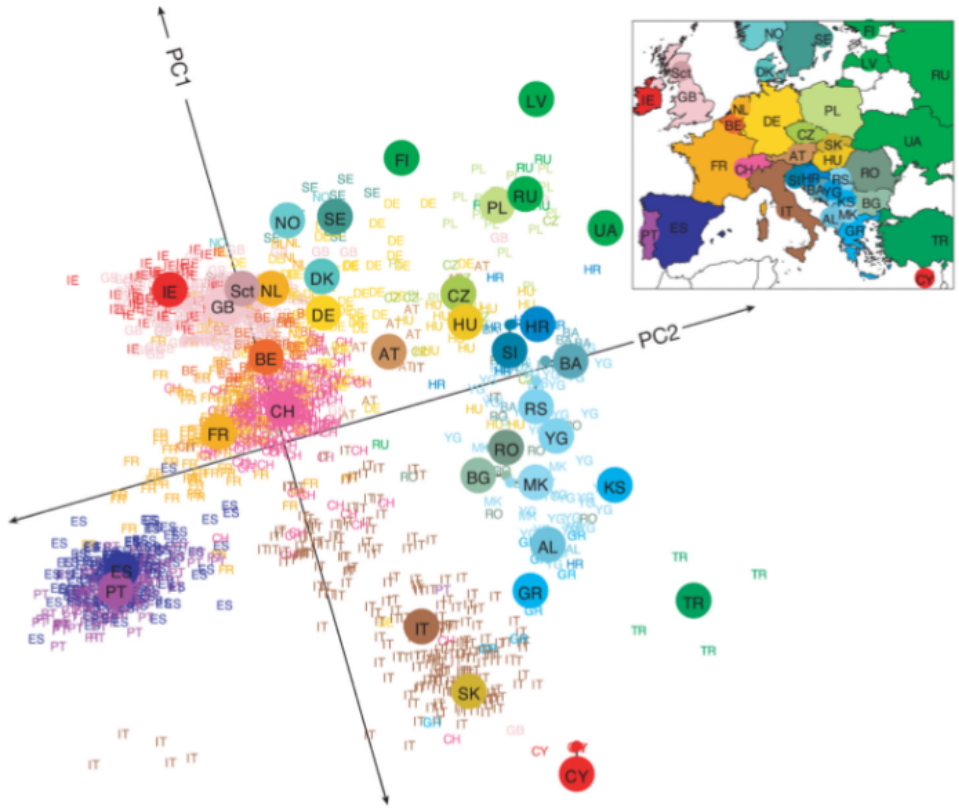

Figure: PCs 1 and 2 describe approximate NS and EW clines in European populations

PCA: references

- James, G., D. Witten, T. Hastie, and R. Tibshirani. n.d. “An Introduction to Statistical Learning.” Springer.
  - Chapters 6 and 10
- Lay, David C., Steven R. Lay, and Judi J. McDonald. 2016. Linear Algebra and Its Applications. Pearson.
  - Section 7.5
- Search for blogs (Towards Data Science) or StackExchange

Multidimensional scaling

This is a less popular alternative to PCA.

Given dissimilarity (distace) matrix $\mathrm{D} = (d_{ij})$, find matrix Z such that

$$d_{ij} \approx ||\mathrm{z}_i - \mathrm{z}_j||_2$$

Keep q much less than p of the Z columns.

- For classical MDS, $\approx$ is $=$, so this is the same as PCA.
- Can handle non-Euclidean distances.
- PCA for R: example 1, example 2
- Example R code for MDS

Matrix methods: recap

- PCs can summarize continental ancestries, longitude versus latitude within continent
- PCs do not say an African-American individual is 70% West African ancestry and 30% European ancestry

- Use PCs in GWAS to adjusting for confounding due to population structure
- Challenge: how many PCs do we use?
  - Famously, the *K* problem
  - How do we visualize more than 2 PCs?
  - How do we interpret clusters?

Bayesian data analysis

Characterize uncertainty in parameter θ using probability

- Assume data X comes from model $\mathcal{M}_\theta$
- Assume θ comes from prior π
  - π may have "hyperparameters"
- Analogy: updating your initial beliefs with evidence as you collect more data

Apply Bayes rule to get posterior for θ

$$P(\theta|X) = \frac{P(X|\theta) \times \pi(\theta)}{P(X)}$$
$$\propto P(X|\theta) \times \pi(\theta)$$

Sample from $P(\theta|X)$ to report summaries (means, quantiles)

Bayesian data analysis

Ignore the normalizing constant $P(X)$. For conjugate priors, we observe that $P(X|\theta)\pi(\theta)$ looks like model we are familiar with.

- Normal-normal
- Beta-binomial
- Multinomial-Dirichlet

Inference of $\theta \in \mathbb{R}^p$ for $p \geq 2$ may require sampling from marginals of θ.

- Here the $\propto$ is useful
- Search “Gibbs sampling”

BDA: normal-normal conjugacy

$$X|\theta \sim N(\theta, \sigma_1^2)$$
$$\theta \sim N(\mu_2, \sigma_2^2)$$

$$\begin{aligned} P(\theta|X) &= \frac{P(X|\theta) \times \pi(\theta)}{P(X)} \\ &\propto P(X|\theta) \times \pi(\theta) \\ &\propto \exp((X-\theta)^2/(2\sigma_1)) \times \exp((\theta-\mu_2)^2/(2\sigma_2)) \\ &\propto \exp((\theta-\mu)^2/(2\sigma)) \end{aligned}$$

$$\theta|X \sim N(\mu, \sigma^2)$$

BDA: Poisson-gamma conjugacy

$$
\begin{aligned}
X|\theta &\sim \text{Poisson}(\theta) \\
\theta &\sim \text{Gamma}(a, b)
\end{aligned}
$$

$$
\begin{aligned}
P(\theta|X) &= \frac{P(X|\theta) \times \pi(\theta)}{P(X)} \\
&\propto P(X|\theta)\pi(\theta) \\
&= \frac{\exp(-\theta)\theta^X}{X!} \times \frac{b^a \theta^{a-1} \exp(-b\theta)}{\Gamma(a)} \\
&\propto \exp(-(b+1)\theta)\theta^{X+a-1}
\end{aligned}
$$

$$
\theta|X \sim \text{Gamma}(X + a, b + 1)
$$

BDA: beta-binomial conjugacy

$$
\begin{aligned}
X|\theta &\sim \text{Binomial}(n;\theta) \\
\theta &\sim \text{Beta}(a,b)
\end{aligned}
$$

$$
\begin{aligned}
P(\theta|X) &= \frac{P(X|\theta)\times\pi(\theta)}{P(X)} \\
&\propto P(X|\theta)\pi(\theta) \\
&= \binom{n}{X}\theta^X(1-\theta)^{n-X}\times\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \\
&\propto \theta^{X+a-1}(1-\theta)^{n-X+b-1}
\end{aligned}
$$

$$
\theta|X \sim \text{Beta}(X+a, n-X+b)
$$

BDA: multinomial-Dirichlet conjugacy

$$
\begin{aligned}
(X_1, X_2, \ldots, X_q)|p_1, p_2, \ldots, p_q &\sim M(n; p_1, p_2, \ldots, p_q) \\
(p_1, p_2, \ldots, p_q) &\sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_q) \\
P(\theta|X) &= \frac{P(X|\theta) \times \pi(\theta)}{P(X)} \\
&\propto P(X|\theta)\pi(\theta) \\
&\propto \prod_{i=1}^{q} p_i^{X_i} \times \prod_{i=1}^{q} p_i^{\alpha_i - 1} \\
&= \prod_{i=1}^{q} p_i^{X_i + \alpha_i - 1}
\end{aligned}
$$

$$
(p_1, p_2, \ldots, p_q)|(X_1, X_2, \ldots, X_q) \sim \text{Dirichlet}(X_1 + \alpha_1, X_2 + \alpha_2, \ldots)
$$

BDA: recap

- Connect priors beliefs with empirical data
- Posterior means can be posed as convex combination of sample mean and prior mean
  - For example, the posterior mean in the beta-binomial case:

$$\mathbb{E}[\theta|X] = \frac{X+a}{n+a+b} = \frac{a+b}{n+a+b}\frac{a}{a+b} + \frac{n}{n+a+b}\frac{X}{n}$$

where $a/(a+b)$ and X/n are prior and sample means.

- Pay attention to kernel (main part of density function), and simplify with algebra
- For conjugate priors, we can efficiently sample with well-known distributions

A population structure model

Notation

- X : matrix of genotypes of all individuals
- Z : populations of individuals at each marker
  - Consider K populations
- Q : **admixture proportions** of individuals
- F : population-specific **allele frequencies** per marker
- X observed, Z unobserved, (F, Q) to be estimated
- Assume loci are **unlinked**

Likelihood

$$L(F, Q) = P(X|F, Q) = \sum_Z P(X|Z, F, Q)P(Z|F, Q)$$
$$= \sum_Z P(X|Z, F)P(Z|Q)$$

structure and other methods

structure

- Use conjugate Dirichlet priors for F, Q
- Derive a Gibbs sampler
- Inference is based on samples from posterior distr.

ADMIXTURE

- Fast numerical methods to maximize likelihood

FRAPPE

- Derive an EM algorithm to find local optimum F, Q

Another way to solve incomplete data problems

We have emphasized EM and HMM methods for analysis problems where some data is unobserved.

structure introduces another way: data augmentation. In an MCMC sampling routine, we augment the data with values for the unobserved Z. This is the BDA approach.

The frequentist approach is EM, to impute the unobserved Z with expected values.

In general,

- Frequentist methods are faster
- Bayesian methods are more flexible

Thrush data in *structure* paper

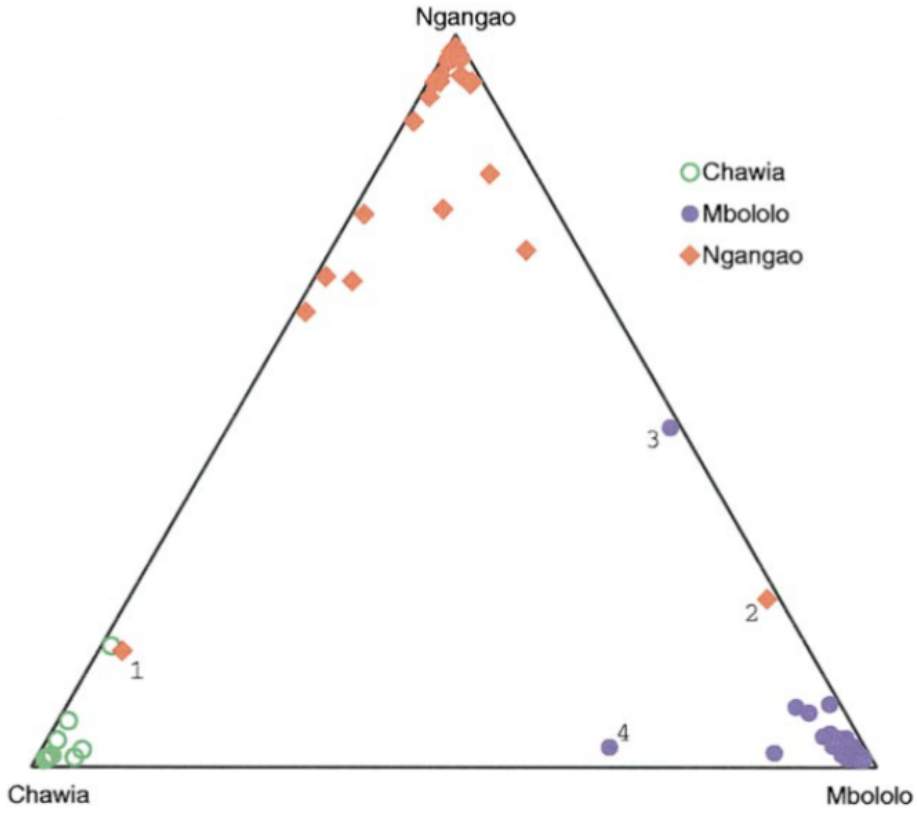

Figure: Triangle plots for posterior mean admixture proportions

Global ancestry: Bryc, et al. (2010)

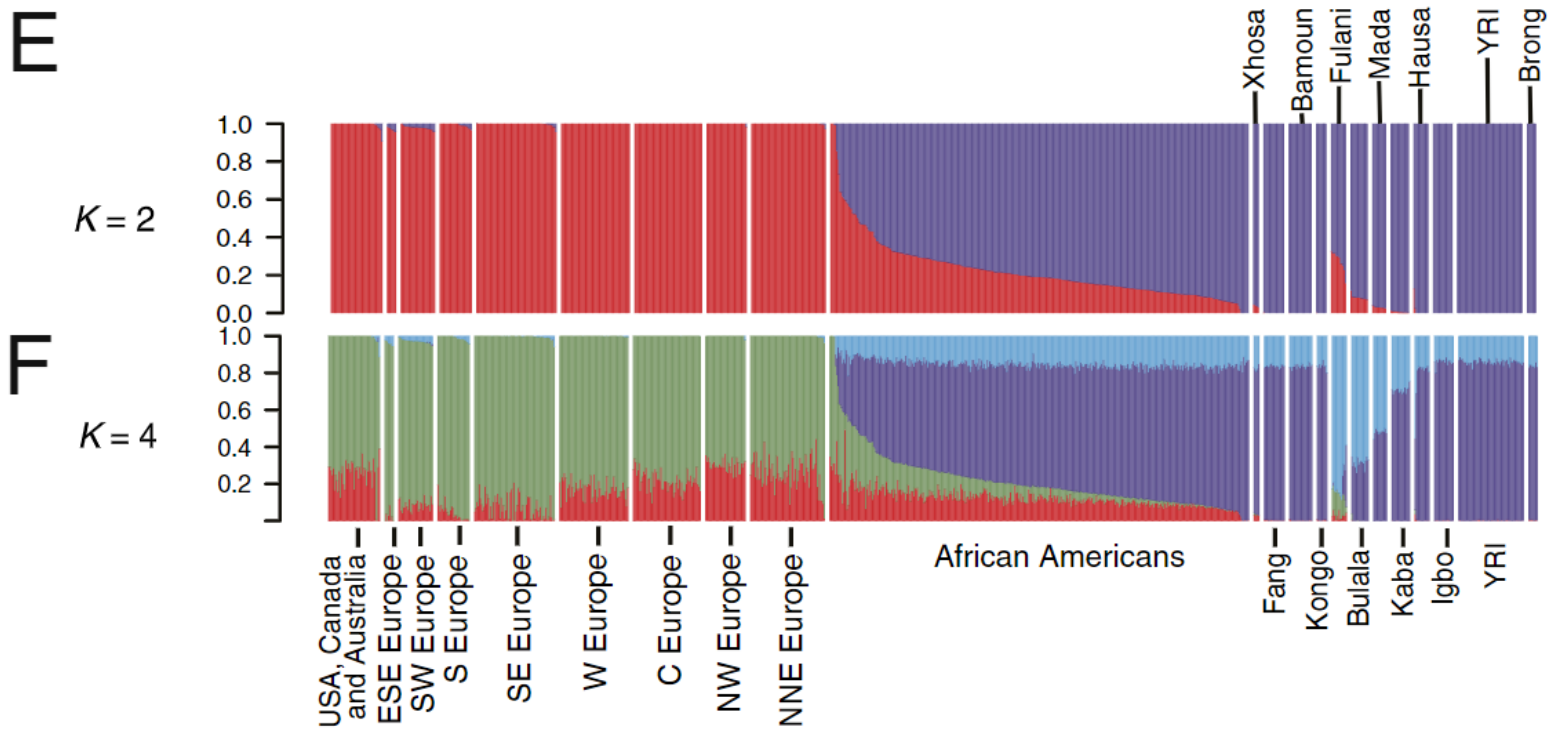

Figure: *structure* analyses for Europeans, Africans, and African-Americans.

1. Motivation
    1.1 Genetic drift
    1.2 Equity in research
2. Global ancestry
    2.1 Principal component analysis
    2.2 Multidimensional scaling
    2.3 Primer: Bayesian data analysis
    2.4 *structure* and other methods
3. Local ancestry
    3.1 RFMix and other methods
    3.2 Admixture mapping

Local ancestry interval (LAI)

Model-based clustering, PCA treat markers as independent. Ancestry occurs in chunks, **similar to IBD**.

If admixture occured G generations ago, post-admixture crossovers occur every $1/G$ Morgans.

- Admixture in Americas is $G \approx 10$, so LAIs ≈ 10 cM
- Not all crossovers switch ancestry, so LAIs should be longer

Popular methods

- RFMix
- LAMP-LD
- HapMix

Use cases for LAIs

- Population genetics on ancestral groups of the admixed
  - Learn historical demography of indigeneous peoples
- Determine identity of ancestral populations
  - Which part of Africa is African ancestry from?
- Estimate timing of admixture events
- Admixture mapping
  - Do LAIs correlate with trait values?

RFMix and other methods

Previous methods work for unlinked loci. To **model LD**, use a Li + Stephens (2003) type model with the augment (ancestry + copied reference haplotype).

- HapMix : does the above; 2 pops only; does phasing simulaneously
- LAMP-LD : use a haplotype frequency models with fewer states (faster)
- RFMix : use conditional random field, similar to HMM, but w/o modeled haplotype frequencies

LAIs for AAs

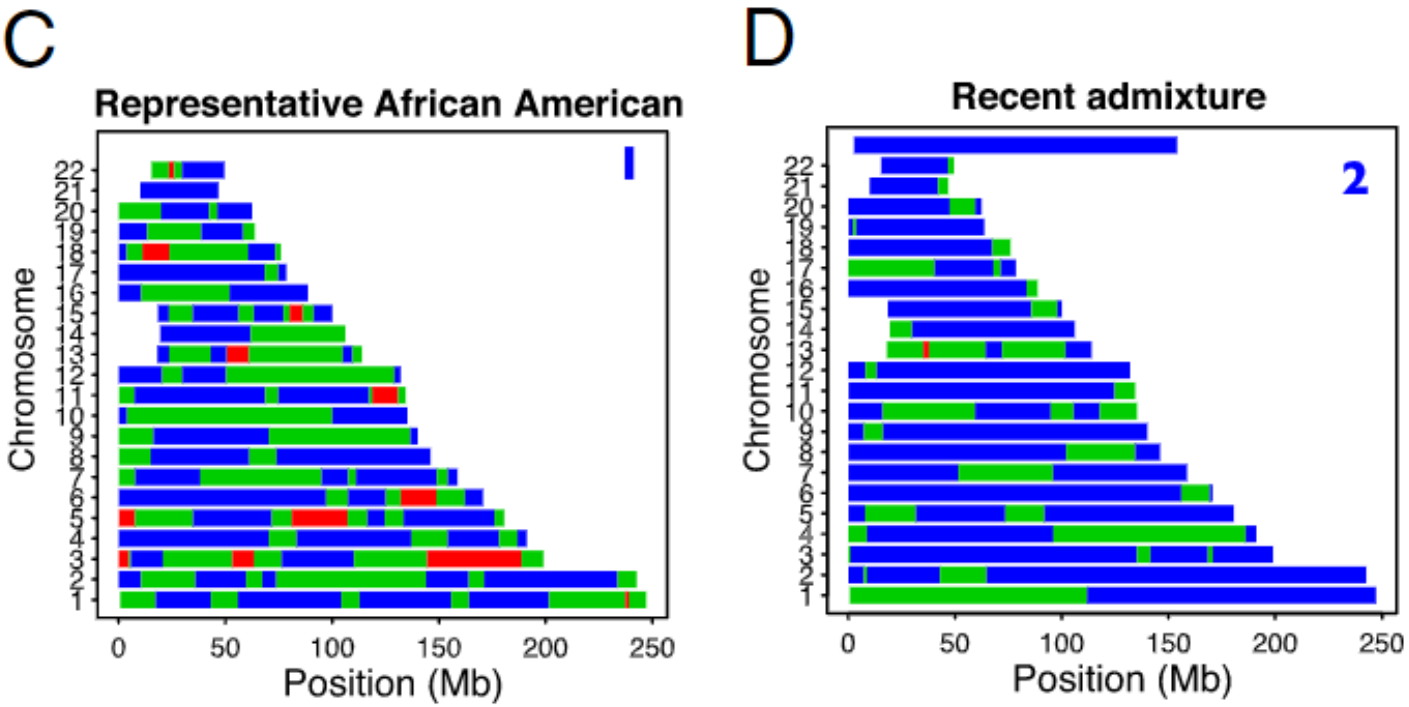

Figure: Example local ancestry intervals for African-Americans and recent African + European admixture. Blue = African ancestry, red = European ancestry, green = shared

Schematic for ancestry + LD model

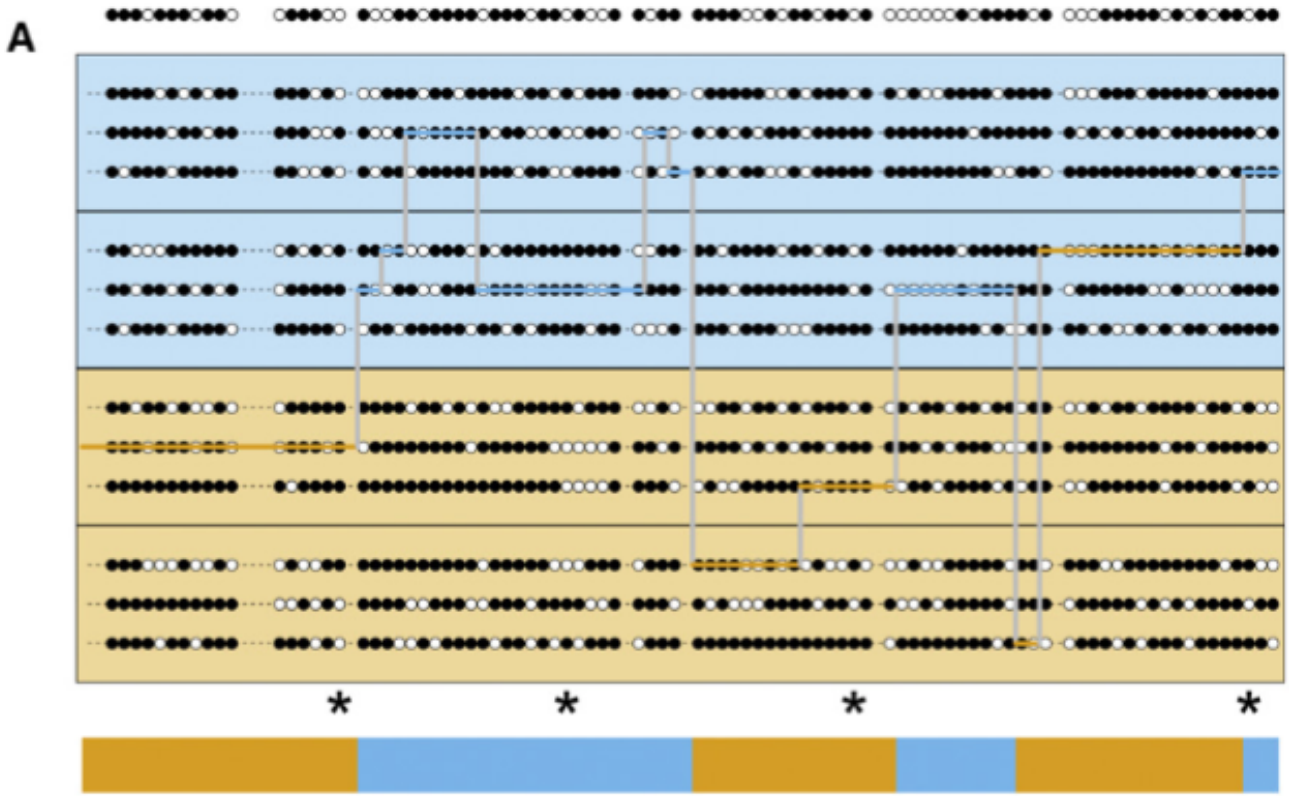

Figure: Copying from reference haplotypes. Blue and orange denote distinct ancestral populations. Asterisks denote ancestry switches (MOSAIC paper, 2019)

Admixture mapping

Goal: check for association between local ancestry and trait

- When population-specific allele frequencies differ a lot at causal locus

Procedure

1. Infer local ancestry from (phased) genotype data
2. Test if # copies of ancestry (0/1/2) is associated with trait
   - One ancestry at a time, or all ancestries at once
3. Apply appropriate genome-wide threshold (not 5e-8)
   - See Grinde, et al. (2019)

Reference: Shriner, Daniel. 2017. “Overview of Admixture Mapping.” Current Protocols in Human Genetics. 94(1): 1.23.1–1.23.8.

Admixture mapping: the math

Notation

- Y is trait
- X_1 is number of copies of ancestry at locus
- $X_2, \ldots$ are other covariates (age, sex, etc.)

Model

$$g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

Test

- $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$
- Likelihood ratio test

Another approach: ancestry informative markers (AIMs)

Some markers have very divergent population-specific allele frequencies.

Procedure

1. Collect a list of AIMs
2. Genotype admixed individuals at the AIMs
3. Test if AIMs are associated with trait
   - May have less power to detect effects
   - Simpler than LAI inference

Admixture mapping: Freedman, et al. (2006)

- Incidence of prostate cancer 1.6 fold higher in AAs
- 1300 AIMs genotyped
- GWAS implicates SNPs in 8q24 region

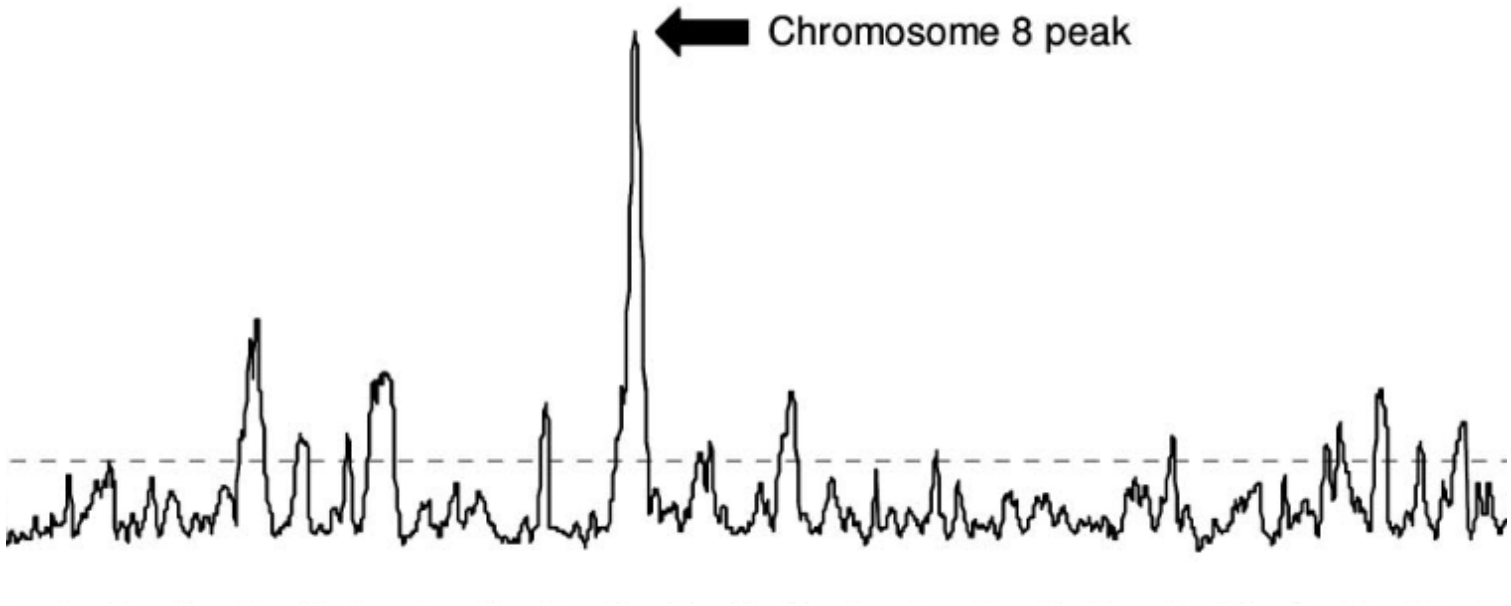

Figure: Interpretation same as GWAS Manhattan plots, except significance threshold may be different

Admixture mapping: Brown, et al. (2017)

- More urine albumin excretion in US Hispanics
- 12k Hispanics/Latinos from HCHS/SOL study
- RFMix to get LAIs
- Associated with indigenous ancestry in chromosome 2

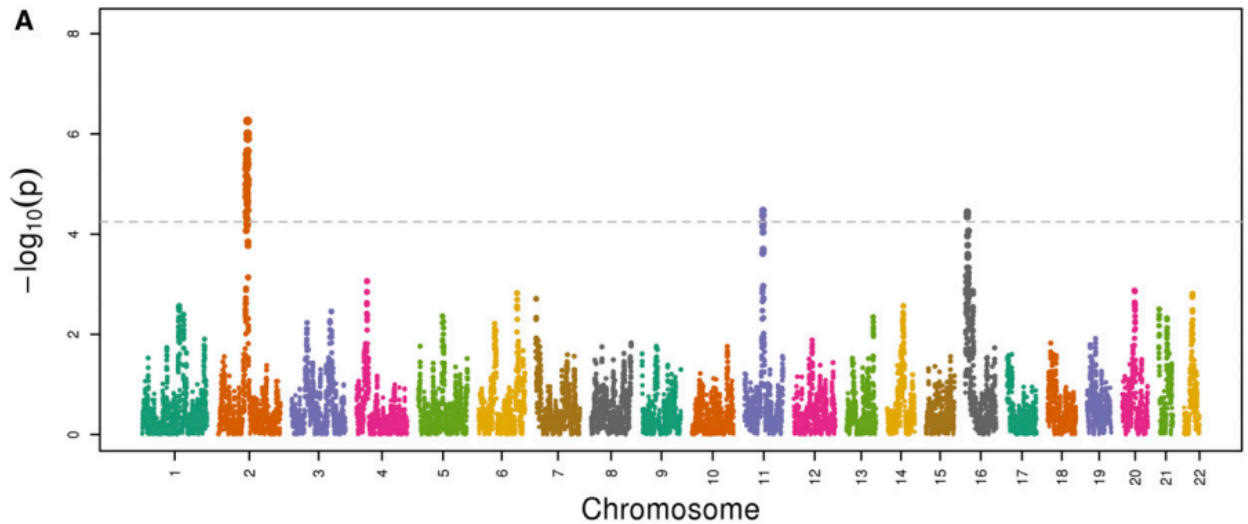

Figure: American ancestry associated with chromosome 2 hit

Admixture mapping: Horimoto, et al. (2021)

- Studied 2565 Caribbean Hispanics
- No GWAS signal found for Alzheimer's dementia
- American local ancestry in 3q13.11 has protective effect

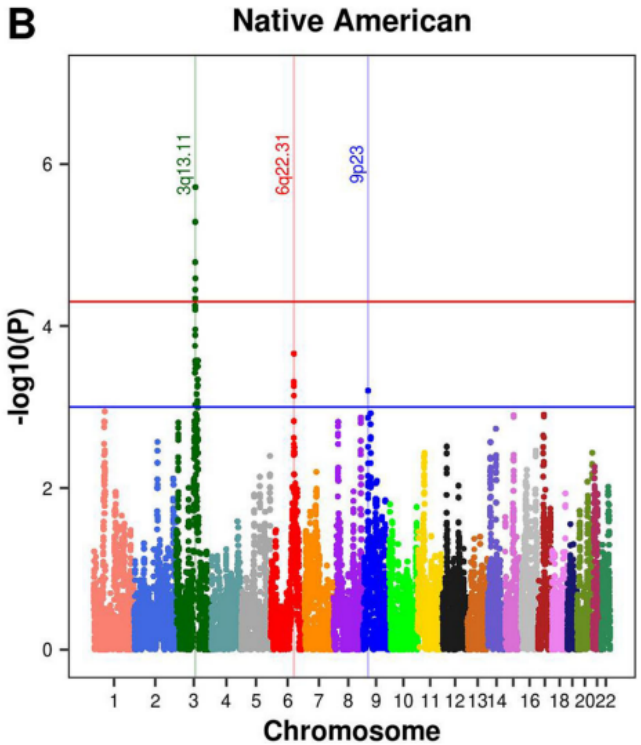

Figure: Manhattan plot for indigenous American ancestry

Admixture mapping: pros and cons versus GWAS

These are shared between admixture and linkage mapping.

Pros

- Can find signals from variants not genotyped
- Can find signals from variants not imputed well (structural variation)

Cons

- If causal marker is genotyped, indirect method $\rightarrow$ less power
- Any signal covers wide region

Admixture mapping: pros and cons versus linkage

Pros

- Signals cover tighter but wide regions
- Do not require families

Con

- Do not obtain the power of families to find Mendelian traits

Complex traits $\rightarrow$ admixture mapping, GWAS
Mendelian traits$\rightarrow$ linkage mapping