Cryptic Relatedness and fine scale population structure
Learning objectives

• Define fine scale population structure and cryptic relatedness

• How is it identified
  • Identity-by-descent
  • Rare variation

• Why it can be important for association analyses, especially of rare variants.
Cryptic Population Structure

Lung Cancer Prevalence in Europe

Boyle and Ferlay (2005) Annals of Oncology
Identity by Decent (IBD): A method to find both distant and recent relationships

- Cousins
- Siblings
- IBD
- Recombination event

The smaller the segments in common, the more distant the relationship.
IBD length is correlated with historical relationships.

\[ E[g|l] \approx \frac{3}{2 \times l} \]


Palamara and Pe’er (2013) Bioinformatics
Identity-by-descent as a means to look at fine-scale structure over time

Harris et al. (2018) PNAS
Identity-by-descent as a means to look at fine-scale structure over time

Harris et al. (2018) PNAS
IBD can estimate effective population size over time.
IBD on a large scale

IBD on a large scale

Rare VS Common: Population Structure Simulations

- **Time of Separation**: $N_A = 10,000$
- **Time of Bottleneck**: $N_B = 1,000$
- **$N_F = 2,000,000$**

Pop A vs. Pop B

Rare VS Common: Assignment of Ancestry Proportions

Information Gain: how well a variant can distinguish between populations. (Rosenberg et al. 2003)

\[ I_n(Q; J) = \sum_{j=1}^{N} \left( -p_j \ln p_j + \sum_{i=1}^{K} q_i p_{ij} \ln p_{ij} \right) \]

Expected Information Gain
- Calculate for a specific site count
- Correct for missing data
- Weighted average to calculate across a range of frequency (rare or common)

\[ E(I_n \mid C, M) = \sum_{m \in M} \sum_{l=0}^{C} r_{lm} \times \sum_{j=1}^{N} \left( -p_{jlm} \ln p_{jlm} + \sum_{i=1}^{K} q_i p_{ijlm} \ln p_{ijlm} \right) \]

Rare Variants Identify Cryptic Populations

Common (MAF > 10%)

O’Connor et al. (2014)
Rare Variants Identify Cryptic Populations

Common (MAF > 10%)  Rare (MAF < 0.5%)

What is Their Geographic Ancestry?

O’Connor et al. (2014)
PCA of Global Diversity Including Cryptic Population

O’Connor et al. (2014)
PCA of Global Diversity Including Cryptic Population

O’Connor et al. (2014)
Population Average PCA with More Axes

- Unknown
- Ashkenazi
- Moroccan
- Sephardic
- Azerbaijani
- Bene Israel
- Cochin
- Ethiopian
- Georgia
- Iranian
- Iraq
- Uzbekistan
- Yemen

Population Average PCA with More Axes

- Unknown
- Ashkenazi
- Moroccan
- Sephardic

O'Connor et al. (2014)  
Trans-Omics for Precision Medicine (TOPMed) Cohorts

- $N \approx 18K$
- This data freeze has 15 cohorts, each with 100s of samples
- Predominantly African, Latino, and European American
  - Samoa
  - Amish
- All are well characterized for heart, lung, blood, and sleep phenotypes
Rare variant sharing across cohorts

- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity
Rare variant sharing across cohorts

- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity
Rare variant sharing across cohorts

- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity
Rare variant sharing across cohorts

- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity
fineStructure analysis of genome-wide ancestry
fineStructure analysis of genome-wide ancestry
fineStructure analysis of genome-wide ancestry
fineStructure analysis of genome-wide ancestry

- African
- Caucasian
- East Asian
- European
African American’s have more homogeneous ancestral proportions

- Calculated Euclidian distance between fineSTRUCTURE proportions
- African American cohorts have the shortest distance and the greatest rare variant sharing
African American’s have more homogeneous ancestral proportions

- Calculated Euclidian distance between fineSTRUCTURE proportions
- African American cohorts have the shortest distance and the greatest rare variant sharing
Estimated Effective Migration Surfaces (EEMS)

- Setting up the population grid.
- Samples are collected at known locations across a two-dimensional habitat; green and orange represent two species of African elephant, forest and savanna, respectively.
- A dense triangular grid is chosen to span the habitat.
- Each sample is assigned to the closest deme on the grid.
- Migration rates vary according to a Voronoi tessellation that partitions the habitat into 'cells' with constant migration rate; colors represent relative rates of migration, ranging from low (orange) to high (blue).
- Each edge has the same migration rate as the cell into which it falls. Cell locations and migration rates are adjusted, using Bayesian inference, so that expected genetic dissimilarities under the EEMS model match observed genetic dissimilarities.
- The EEMS is a color contour plot produced by averaging draws from the posterior distribution of the migration rates, interpolating between grid points.

Sampling schemes

- PCA: uniform
- PCA: barrier
- EEMS: barrier
- EEMS: uniform

Figure 2: Simulations comparing EEMS and PCA. For each method, we show results for two migration scenarios—representing uniform migration and a barrier to migration—and three different sampling schemes.

- The true underlying migration rates for the uniform scenario; colors represent relative migration rates.
- The three sampling schemes used; the size of the circle at each node is proportional to the number of individuals sampled at that location, and locations are color-coded to facilitate cross-referencing the EEMS and PCA results.
- PCA results.
- EEMS results. In contrast to PCA, EEMS is robust to sampling scheme and shows clear qualitative differences between the estimated effective migration rates under the two scenarios, reflecting the underlying simulation truth.

Petkova et al. (2015)
Nature Genet.
Assumptions: Stepping Stone Model

- Migration can only occur between adjacent demes.
- Migration rate between each deme is assumed to be equal.

Kimura and Weiss (1964)
EEMS: Migration and diversity within Peru

Harris et al. (2018) PNAS
EEMS captures long-term migration patterns

Richmond et al. (2015) Molecular Ecology
EEMS in Malaria Parasites of South East Asia

Shetty et al. (submitted)
Application to Malaria Parasites in W. Africa

Pf3K Version 5.1
Robustness of Sampling on EEMS
Concluding summary

- Fine-scale population structure is subdivisions of individuals on an ever increasingly granular scale.
- Identity-by-descent and sharing of rare variants are a powerful method of identifying recent relationships and can be scaled by time.
- Cryptic population structure arises with extended relationships within a cohort, unknown to the investigators.
- EEMS can visualize migration patterns on a fine-scale illustrating cryptic structure not observed with other methods.
Questions?