Population Structure in Genome

02-223 Personalized Medicine:
Understanding Your Own Genome
Fall 2014
Genetics So Far...

• Recombination
  – Haplotypes, linkage disequilibrium etc.

• Mutation
  – Population structure in genomes
Minor Allele Frequency

• Minor allele for a SNP: the less frequently occurring allele

• Given genotype data for $N$ individuals
  – For each locus, we can define **minor allele frequency** as follows:
    \[
    \text{(Minor allele frequency)} = \frac{\text{(the number of minor alleles in the population)}}{\text{(total number of alleles in the population)}}
    \]
  – Typically, SNPs with a very low minor allele frequency are discarded, since they don’t contain sufficient information about genetic diversity.
## Minor Allele Frequency

- What is the minor allele frequency at each SNP locus?

<table>
<thead>
<tr>
<th>Individuals</th>
<th>SNP1</th>
<th>SNP2</th>
<th>SNP3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Allele Frequency and Mutations

• When a new mutation arises in a population for the first time, what is the minor allele frequency of this mutation?

• How would this allele frequency increase or decrease over time?
Overview

• Begin by considering a **single population**
  – Hardy-Weinberg Equilibrium
  – Genetic drift

• **Multiple populations**
  – Admixture
  – Clustering algorithm
  – Inferring population structure from genotype data by clustering individuals based on genotype data
Hardy-Weinberg Equilibrium

• Hardy-Weinberg Equilibrium
  – Under random mating, both allele and genotype frequencies in a population remain constant over generations.
  – Assumptions of the standard random mating
    • Diploid organism
    • Sexual reproduction
    • Nonoverlapping generations
    • Random mating
    • Large population size
    • Equal allele frequencies in the sexes
    • No migration/mutation-selection
Hardy-Weinberg Equilibrium

• $p, q$: allele frequencies of two alleles, $A$ and $a$

• Then, in the current generation, the genotype frequencies are
  – $D = p^2$
  – $H = 2pq$
  – $R = q^2$
  where $D$, $H$, $R$ are genotype frequencies for $AA$, $Aa$, $aa$, respectively.
Hardy-Weinberg Equilibrium: What Happens in the Next Generation?

- $D, H, R$: genotype frequencies for $AA, Aa, aa$, respectively.
- $p, q$: allele frequencies of $A$ and $a$
- $D', H', R'$: genotype frequencies for the next generation after random mating

<table>
<thead>
<tr>
<th>Mating</th>
<th>Frequency of mating</th>
<th>Offspring genotype frequencies</th>
</tr>
</thead>
</table>
| $AA \times AA$ | $D^2$              | $\begin{array}{ccc}
AA & Aa & aa \\
1 & 0 & 0
\end{array}$                |
| $AA \times Aa$ | $2DH$             | $\begin{array}{ccc}
AA & Aa & aa \\
\frac{1}{2} & \frac{1}{2} & 0
\end{array}$                |
| $AA \times aa$ | $2DR$             | $\begin{array}{ccc}
AA & Aa & aa \\
0 & 1 & 0
\end{array}$                |
| $Aa \times Aa$ | $H^2$             | $\begin{array}{ccc}
AA & Aa & aa \\
\frac{1}{4} & \frac{1}{2} & \frac{1}{4}
\end{array}$                |
| $Aa \times aa$ | $2HR$             | $\begin{array}{ccc}
AA & Aa & aa \\
0 & \frac{1}{2} & \frac{1}{2}
\end{array}$                |
| $aa \times aa$ | $R^2$             | $\begin{array}{ccc}
AA & Aa & aa \\
0 & 0 & 1
\end{array}$                |
Hardy-Weinberg Equilibrium: What Happens in the Next Generation?

- $D, H, R$: genotype frequencies for $AA$, $Aa$, $aa$, respectively.
- $p, q$: allele frequencies of $A$ and $a$
- $D', H', R'$: genotype frequencies for the next generation after random mating

<table>
<thead>
<tr>
<th>Mating</th>
<th>Frequency of mating</th>
<th>Offspring genotype frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AA \times AA$</td>
<td>$D^2$</td>
<td>$\begin{array}{ccc} AA &amp; Aa &amp; aa \ 1 &amp; 0 &amp; 0 \end{array}$</td>
</tr>
<tr>
<td>$AA \times Aa$</td>
<td>$2DH$</td>
<td>$\begin{array}{ccc} AA &amp; Aa &amp; aa \ \frac{1}{2} &amp; \frac{1}{2} &amp; 0 \end{array}$</td>
</tr>
<tr>
<td>$AA \times aa$</td>
<td>$2DR$</td>
<td>$\begin{array}{ccc} AA &amp; Aa &amp; aa \ 0 &amp; 1 &amp; 0 \end{array}$</td>
</tr>
<tr>
<td>$Aa \times Aa$</td>
<td>$H^2$</td>
<td>$\begin{array}{ccc} AA &amp; Aa &amp; aa \ \frac{1}{4} &amp; \frac{1}{2} &amp; \frac{1}{4} \end{array}$</td>
</tr>
<tr>
<td>$Aa \times aa$</td>
<td>$2HR$</td>
<td>$\begin{array}{ccc} AA &amp; Aa &amp; aa \ 0 &amp; \frac{1}{2} &amp; \frac{1}{2} \end{array}$</td>
</tr>
<tr>
<td>$aa \times aa$</td>
<td>$R^2$</td>
<td>$\begin{array}{ccc} AA &amp; Aa &amp; aa \ 0 &amp; 0 &amp; 1 \end{array}$</td>
</tr>
</tbody>
</table>

Totals (next generation)

\[
\begin{array}{c}
D' = D^2 + 2DH/2 + H^2/4 = (D + H/2)^2 = p^2 \\
H' = 2DH/2 + 2DR + H^2/2 + 2HR/2 = 2(D + H/2)(R + H/2) = 2pq \\
R' = H^2/4 + 2HR/2 + R^2 = (R + H/2)^2 = q^2
\end{array}
\]

Genotype frequencies remain the same between current and next generations.
Hardy Weinberg Equilibrium

• Under Hardy Weinberg Equilibrium, when a new mutation arises in a population, how does the allele frequency of this mutation change over time?
Hardy-Weinberg Equilibrium in Practice

- HWE often does not hold in reality because of the violation of the assumptions (i.e., random mating, no selection, etc.)

- Even when the assumptions for HWE hold, in reality, allele frequencies change over generations because of the random fluctuation – **genetic drift**!
Genetic Drift

• The change in allele frequencies in a population due to random sampling

• All mutations eventually drift to allele frequency 0 or 1 over time

• Neutral process unlike natural selection
  – But genetic drift can eliminate an allele from the given population.
Genetic Drift and Population Size

• The effect of genetic drift is larger in a small population

• How does the allele frequency change over time for a large population and a small population?
Scenarios of How Populations Evolve

When there is a population divergence, how do the allele frequencies change?
Overview

• Begin by considering a **single population**
  – Hardy-Weinberg Equilibrium
  – Genetic drift

• **Multiple populations**
  – Admixture
  – Clustering algorithm
  – Inferring population structure from genotype data by clustering individuals based on genotype data
What is Population Structure?

- Population Structure
  - A set of individuals characterized by some measure of genetic distinction
  - A “population” is usually characterized by a distinct distribution over genotypes – allele frequencies
  - Example

```
<table>
<thead>
<tr>
<th>Genotypes</th>
<th>aa</th>
<th>aA</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
# HapMap Phase 3 Samples

<table>
<thead>
<tr>
<th>label</th>
<th>population sample</th>
<th># samples</th>
<th>QC+ Draft 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW*</td>
<td>African ancestry in Southwest USA</td>
<td>90</td>
<td>71</td>
</tr>
<tr>
<td>CEU*</td>
<td>Utah residents with Northern and Western European ancestry from the CEPH collection</td>
<td>180</td>
<td>162</td>
</tr>
<tr>
<td>CHB</td>
<td>Han Chinese in Beijing, China</td>
<td>90</td>
<td>82</td>
</tr>
<tr>
<td>CHD</td>
<td>Chinese in Metropolitan Denver, Colorado</td>
<td>100</td>
<td>70</td>
</tr>
<tr>
<td>GIH</td>
<td>Gujarati Indians in Houston, Texas</td>
<td>100</td>
<td>83</td>
</tr>
<tr>
<td>JPT</td>
<td>Japanese in Tokyo, Japan</td>
<td>91</td>
<td>82</td>
</tr>
<tr>
<td>LWK</td>
<td>Luhya in Webuye, Kenya</td>
<td>100</td>
<td>83</td>
</tr>
<tr>
<td>MEX*</td>
<td>Mexican ancestry in Los Angeles, California</td>
<td>90</td>
<td>71</td>
</tr>
<tr>
<td>MKK*</td>
<td>Maasai in Kinyawa, Kenya</td>
<td>180</td>
<td>171</td>
</tr>
<tr>
<td>TSI</td>
<td>Toscans in Italy</td>
<td>100</td>
<td>77</td>
</tr>
<tr>
<td>YRI*</td>
<td>Yoruba in Ibadan, Nigeria</td>
<td>180</td>
<td>163</td>
</tr>
</tbody>
</table>

| Total  | 1,301 | 1,115 |

* Population is made of family trios
HapMap: Allele Frequencies in Different Populations

- Comparison of allele frequencies for individuals from pairs of populations
- The red regions show that there are many SNPs that have similar low frequencies in each pair of analysis panels/populations.
- CHB (Chinese) and JPT (Japanese) have similar allele frequencies
Motivation

• **Reconstructing individual ancestry:** The Genographic Project

• **Studying human migration**
  – **Out of Africa**
  – **Multi-regional hypothesis**

• **Study of various traits**
  – Lactose intolerance
  – Origins in Europe?
  – Infer from
    • Migration studies
    • Mutation studies in populations
Lactose Persistence/Intolerance
Motivation

- **Reconstructing individual ancestry:** The Genographic Project
  - [https://genographic.nationalgeographic.com/genographic/index.html](https://genographic.nationalgeographic.com/genographic/index.html)
- **Studying human migration**
200,000 years ago

50,000 years ago

30,000 years ago

10,000 years ago

https://genographic.nationalgeographic.com/genographic/index.html
Scenarios of How Populations Evolve

Single population  Divergence  Admixing
Admixture

each person’s genome is a mosaic of genomes from different populations
Inferring Population Structure from Genome Data

• Assume the genome data from a large number of individuals are given

• Assume the ethnicity of each individual is unknown

• Then, can we figure out from the genome data:
  – which individuals belong to the same population (or ethnic) group?
  – How many population (or ethnic) groups are there?
What is Clustering?

- Organizing data into *clusters* such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects.
K-means Clustering Algorithm: Partitional Clustering

• Each object is placed in exactly one of $K$ non-overlapping clusters.

• The user has to specify the desired number of clusters $K$.

• We use the similarity measures between an observed sample and the cluster center (mean).
K-means Clustering: Initialization

- For a pre-defined number of clusters $K$, initialize $K$ centers randomly.
K-means Clustering: Iteration 1

- Iterate between the following two steps
  - Assign all objects to the nearest center.
  - Move a center to the mean of its members.
K-means Clustering: Iteration 2

- After moving centers, re-assign the objects...
K-means Clustering: Iteration 2

- After moving centers, re-assign the objects to nearest centers.
- Move a center to the mean of its new members.
K-means Clustering: Finished!

- Re-assign and move centers, until no objects changed membership.
Algorithm $k$-means

1. Decide on a value for $K$, the number of clusters.
2. Initialize the $K$ cluster centers randomly.
3. Decide the cluster memberships of the $N$ objects by assigning them to the nearest cluster center.
4. Re-estimate the $K$ cluster centers, by assuming the memberships found above are correct.
5. Repeat 3 and 4 until none of the $N$ objects changed membership in the last iteration.
Algorithm $k$-means

1. Decide on a value for $K$, the number of clusters.

2. Initialize the $K$ cluster centers (randomly, if necessary).

3. Decide the cluster memberships of the $N$ objects by assigning them to the nearest cluster center.

4. Re-estimate the $K$ cluster centers, by assuming the memberships found above are correct.

5. Repeat 3 and 4 until none of the $N$ objects changed membership in the last iteration.

Use one of the distance / similarity functions, sum of squared differences between cluster mean and data point

Average / median of cluster members
**Soft-Clustering of Individuals into Three Clusters with Mixture Model**

<table>
<thead>
<tr>
<th>Probability of</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual 1</td>
<td>0.1</td>
<td>0.4</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Individual 2</td>
<td>0.8</td>
<td>0.1</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>Individual 3</td>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>Individual 4</td>
<td>0.10</td>
<td>0.05</td>
<td>0.85</td>
<td>1</td>
</tr>
<tr>
<td>Individual 5</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Individual 6</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Individual 7</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Individual 8</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Individual 9</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Individual 10</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

- Each individual can be assigned to more than one cluster with a certain probability.
- For each individual, the probabilities for all clusters should sum to 1. (i.e., each row should sum to 1.)
- Each cluster is explained by a cluster center variable (i.e., cluster mean)
Clustering Algorithm for Inferring Population Structure

• Cluster genome sequences into multiple population groups
  – A population label (cluster label) is assigned to each individual
  – The whole genome (all loci) of each individual are assigned to have come from the same population
  – It does not consider population admixture, where different parts of the genome can have different ancestry

• *Structure* extends the mixture model to an admixture model
  – Relax the assumption of one ancestor per individual in mixture model
  – Individuals can have ancestors in multiple different populations
Population Structure Inference as Clustering

• Genotype data from an endangered bird species, the Taita thrush

• Individual birds were sampled at four locations in southeast Kenya
  – Chawia (17 individuals)
  – Ngangao (54 individuals)
  – Mbololo (80 individuals)
  – Yale (4 individuals)

• The geographic samples are likely to represent distinct populations because each location corresponds to a distinct fragment of cloud forest

• Genotyped at seven microsatellite loci
Clustering Genome Data to Learn Population Structure

- Each point on the right corresponds to a bird.

- Clustering the birds based on genotype data with \( K (\text{#clusters}) = 3 \) with a mixture model.

- For each bird, the values of the three coefficients (cluster probabilities) in the ancestry vector \( (q_1, q_2, q_3) \) are given by the distances to each of the three sides of the equilateral triangle.

- The coloring shows the known regions where the bird was taken and is not the result of clustering.

Inference of Population Structure Using Multilocus Genotype Data. (Pritchard et al., Genetics, 2000)
23 pairs of chromosomes. One unique you.

Find out what percent of your DNA comes from populations around the world, ranging from East Asia, Sub-Saharan Africa, Europe, and more. Break European ancestry down into distinct regions such as the British Isles, Scandinavia and Italy. People with mixed ancestry, African Americans, Latinos, and Native Americans will also get a detailed breakdown.
From Mixture to Admixture Model

• Mixture model
  – The genotypes for all loci have the sample population membership.

• Admixture model
  – In reality, each person’s genome is a mosaic of genomes from different populations. Admixture!
  – Mixture model has been generalized to admixture model to allow different locus to have different population membership.
Population Structure with Admixing Inferred from Genome Data
Population Structure with Admixing Inferred from Genome Data
Personalized Medicine and Population Structure in Genomes

- *FY*O allele for resistance to P. vivax malaria
Social Issues on Genomes and Race

• Your race/ethnicity as reflected in genomes

• Ethical issues related to the genetics behind race
  – BiDil: racialized drug which works well for African Americans to reduce heart failure
  – http://www.radiolab.org/story/91653-race/
Social Issues on Genomes and Race

• Jews’ genetic similarity:
  – Genetic studies have often produced conflicting results on the question of whether distant Jewish populations in different geographic locations share greater genetic similarity to each other or instead, to nearby non-Jewish populations.
Social Issues on Genomes and Race

• Jews’ genetic similarity:
  – Ashkenazi Jews from Europe, Sephardic Jews from Spain, Portugal, and middle east
  – Studies of genomes of Jews revealed that there was more contact between Ashkenazim and Sephardim than suspected
  – Members of any Jewish community are related to one another as closely as are fourth or fifth cousins in a large population, which is about 10 times higher than the relationship between two people chosen at random off the streets of New York City
  – http://www.nytimes.com/2010/06/10/science/10jews.html?_r=0
Summary

• Genetic variation data can be used to infer various aspects of population history such as population divergence, admixture.

• HWE describes the theoretical allele frequencies in the ideal situation. In practice, because of genetic drift, the allele frequency for a new mutation eventually reaches 0 or 1 within a population.

• Clustering algorithms can be used to infer population structure from genome data.