Lecture 13
Population Genetics

Slides credit:
Abhisheek Sarkar
<table>
<thead>
<tr>
<th>Project</th>
<th>PSets</th>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Loc</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Describe your previous research, areas of interest in computational biology, type of project that best fits your interests. Post in a profile that lets your classmates know you and find potential partners. <strong>Project profile due Tue 9/26</strong></td>
<td>PS1 out on L1-L5</td>
<td>1</td>
<td>Thu, Sep 7</td>
<td>Introduction</td>
<td>L1</td>
<td>Intro: Biology, Algorithms, Machine Learning, Course Overview</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Sep 8</td>
<td></td>
<td></td>
<td>R1 Recitation: Biology and Probability Review</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Tue, Sep 12</td>
<td>Module I: Aligning and Modeling Genomes</td>
<td>L2</td>
<td>Alignment I: Dynamic Programming, Global and local alignment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Sep 14</td>
<td></td>
<td></td>
<td>R2 Recitation 2: Deriving Parameters of Alignment, Multiple Alignment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Sep 15</td>
<td></td>
<td></td>
<td>L3 Alignment II: Database search, Rapid string matching, BLAST, BLOSUM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Tue, Sep 19</td>
<td>Foundations</td>
<td>L4</td>
<td>Hidden Markov Models Part 1: Evaluation/Parsing, Viterbi, Forward algorithms</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Sep 21</td>
<td></td>
<td></td>
<td>L5 Hidden Markov Models Part 2: Posterior Decoding, Learning, Baum-Welch</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Sep 22</td>
<td></td>
<td></td>
<td>No classes - student holiday</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>Mon, Sep 25</td>
<td>Project Intro: about the projects, self introductions, mentor intro, example projects, teamwork 32D-507</td>
<td></td>
<td>No Classes - Columbus Day Holiday</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>Tue, Oct 6</td>
<td>Foundations</td>
<td>L6</td>
<td>Expression Analysis: Clustering/Classification, K-means, Hierarchical, Bayesian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 6</td>
<td></td>
<td></td>
<td>R3 Recitation 3: Affinity Propagation Clustering and Random Forest Classification</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>Tue, Oct 10</td>
<td>Module I: Gene Expression and Epigenomics</td>
<td>L7</td>
<td>Transcript structure: GenScan, RNA-seq, Mapping, De novo Assembly, Diff Expr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 12</td>
<td></td>
<td></td>
<td>L8 Epigenomics: ChIP-Seq, Read mapping, Peak calling, IDR, Chromatin states</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 13</td>
<td></td>
<td></td>
<td>L9 Three-dimensional chromatin interactions: 3C, 5C, HIC, CHIA-Pet</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 20</td>
<td></td>
<td></td>
<td>R4 Recitation 4: ENCODE, Epigenome Roadmap, ChromHMM, ChromImpute</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>Tue, Oct 17</td>
<td>Foundations</td>
<td>L10</td>
<td>Regulatory Motifs: Discovery, Representation, PBMs, Gibbs Sampling, EM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 19</td>
<td></td>
<td></td>
<td>R5 Recitation 5: Gapped Motif Discovery, DNASHape, PBMs, Selex</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 20</td>
<td></td>
<td></td>
<td>L11 Network structure, centrality, SVD, sparse PCA, L1/L2, modules, diffusion kernels</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>Tue, Oct 24</td>
<td>Module IV: Population Genetics and Disease Genomics</td>
<td>L12</td>
<td>Deep Learning, Neural Nets, Convolutional NNs, Recurrent NNs, Autoencoder</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Oct 26</td>
<td></td>
<td></td>
<td>R6 Recitation 6: Networks review, Recommendation systems, EHR, PeheWAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Oct 27</td>
<td></td>
<td></td>
<td>Project feedback: Prepare 2-3 slide presentation of your term project for your mentor, 32D-507 at 4:30pm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>Tue, Oct 31</td>
<td>Foundations</td>
<td>L13</td>
<td>Population genetics: Linkage disequilibrium, pop struct, 1000genomes, allele freq</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 2</td>
<td></td>
<td></td>
<td>L14 Disease Association Mapping, GWAS, Organismal phenotypes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 3</td>
<td></td>
<td></td>
<td>R7 Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>Tue, Nov 7</td>
<td>Module V: Comparative Genomics and Evolution</td>
<td>L15</td>
<td>Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, IMWAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 8</td>
<td></td>
<td></td>
<td>L16 Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 10</td>
<td></td>
<td></td>
<td>R8 Recitation 8: Rare Variants, ExAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 17</td>
<td></td>
<td></td>
<td>Project feedback: Prepare 2-3 slide presentation of your term project for your mentor, 32D-507 at 4:30pm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>Thu, Nov 14</td>
<td>Foundations</td>
<td>L17</td>
<td>No Recitation, Veterans Day</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>No more sets! work on your final project</td>
<td>Quiz</td>
<td>No Recitation, Veterans Day</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 21</td>
<td></td>
<td></td>
<td>No lecture, thanksgiving break - Thu Nov 26, 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 23</td>
<td></td>
<td></td>
<td>No recitation, thanksgiving break</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Nov 24</td>
<td></td>
<td></td>
<td>No lecture, thanksgiving break - Thu Nov 26, 2015</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>Thu, Nov 28</td>
<td>Module VI: Current Research Directions</td>
<td>L21</td>
<td>Single-cell genomics: technology, analysis, microfluidics, applications, insights</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Nov 30</td>
<td></td>
<td></td>
<td>L22 Mining human phenotypes, PeheWAS, UK Biobank, meta-phenotypes + imputation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 1</td>
<td></td>
<td></td>
<td>R10 Recitation 10: Project Feedback, results, interpretation, directions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>Tue, Dec 5</td>
<td>Foundations</td>
<td>L23</td>
<td>Cancer Genomics, Single-cell Sequencing, Tumor-Immune Interface ]]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Dec 7</td>
<td></td>
<td></td>
<td>L24 Genome Engineering with CRISPR/Cas9 and related technologies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Thu, Dec 8</td>
<td></td>
<td></td>
<td>R11 Recitation 11: Presentation Tips - Intro, discussion, Slides, Presentation skills</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fri, Dec 18</td>
<td></td>
<td></td>
<td>Project feedback: Prepare 2-3 slide presentation of your term project for your mentor, 32D-507 at 4:30pm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>Tue, Dec 21</td>
<td>Foundations</td>
<td>L25</td>
<td>Final Presentations - Part I (11am), 32G-G8 reading room</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tue, Dec 25</td>
<td></td>
<td></td>
<td>Final Presentations - Part I (1pm), 32-141</td>
</tr>
</tbody>
</table>

* readings refer to chapters in compiled 2016 scribe notes, available in the materials folder on Stellar
** recitation topics will be adjusted to respond to lecture and student needs
### Module 4: Population and Disease Genetics

<table>
<thead>
<tr>
<th>Module IV: Population and Disease Genetics</th>
<th>L13: Population genetics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>L13</td>
<td>Linkage disequilibrium, pop struct, 1000genomes, allele freq</td>
</tr>
<tr>
<td>L14</td>
<td>Disease Association Mapping, GWAS, organismal phenotypes</td>
</tr>
<tr>
<td>R7</td>
<td>Recitation 7: Linkage Disequilibrium, Haplotype Phasing, Genotype Imputation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frontiers</th>
<th>L15: Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>L15</td>
<td>Missing Heritability, Complex Traits, Interpret GWAS, Rank-based enrichment</td>
</tr>
<tr>
<td>R8</td>
<td>Recitation 8: Rare Variants, ExAC</td>
</tr>
</tbody>
</table>

• **L13: Population genetics:**
  - Measuring and understanding human variation

• **L14: Disease association mapping:**
  - Molecular basis of human phenotypic variation and disease

• **L15: Quantitative trait mapping:**
  - Intermediate phenotypes bridging the genotype-phenotype gap

• **L16: Heritability:**
  - Whole-genome disease association beyond top hits
Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   - SNPs, indels, copy-number, structural, short tandem repeats.
   - Detection: 1000Genomes project, ExAC, rare/common/neutral.
   - Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   - Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   - 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering (FST). PC correction

4. Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales
Today: population genetics

1. **Identifying and measuring genetic variation**
2. Recombination and patterns of genetic variation
3. Human relatedness and ancestry painting
4. Human history and demography
5. Human evolution
## Types of genetic variation

- **99% of DNA is shared** between two individuals
- Variation in the remainder explains all our **predisposition** differences
- **Remaining** phenotypic variation: environmental/stochastic differences

<table>
<thead>
<tr>
<th>Name</th>
<th>Example</th>
<th>Frequency in one genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single nucleotide polymorphisms (SNPs)</td>
<td>GAGGAGAACG[C/G]AACTCGCCG</td>
<td>1 per 1,000 bp</td>
</tr>
<tr>
<td>Insertions/deletions (indels)</td>
<td>CACTATTC[C/CTATGG]TGTCTAA</td>
<td>1 per 10,000 bp</td>
</tr>
<tr>
<td>Short tandem repeats (STRs)</td>
<td>ACGGCAAGTCGTCGTCGTCACCGTAT</td>
<td>1 per 10,000 bp</td>
</tr>
<tr>
<td>Structural variants (SVs) / Copy Number Variants (CNVs)</td>
<td>Large (median 5,000 bp) deletions, duplications, inversions</td>
<td>1 per 1,000,000 bp</td>
</tr>
</tbody>
</table>
Representing and storing genetic variants

- Many modern analyses (GWAS, eQTL) focus on SNPs/indels
- Often have only two alleles (states)
- Identified as reference SNP clusters (rsid)
- Submitted sequences containing a variant are clustered to build a database (dbSNP)
- To date, >100 M known variants in dbSNP

rs189107123

GAGGAGAACG[C/G]AACTCCGCCG
Representing and storing genetic variants

Distinguishing the two alleles:
• Matching the human reference sequence (reference/alternate)
• Being more frequent in the population (major/minor)
• Matching the most recent common ancestor between human and chimpanzee (ancestral/derived)

Classifying variants by minor allele frequency:
• Common (> 5%)
• Low frequency (0.5-5%)
• Rare (< 0.5%)
• Private/de novo (one observation)

rs189107123
GAGGAGAACG[C/G]AACTCCGCCG

Reference allele: C
Minor allele: G (frequency 0.03 in Europeans)
Ancestral allele: unknown (why?)
Common alleles typically have small effects.
Representing and storing genetic variants

- Humans are **diploid** organisms
- Each individual carries two **homologous** copies of each chromosome
- Therefore, they carry two copies of each variant (called the **maternal/paternal allele**)
- Variants co-occur in **haplotypes** which are inherited as a unit

**Haplotypes**

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(maternal)  
(paternal)

**Genotypes**

| 0 | 1 | 2 | 0 | 1 | 2 | 0 |
Representing and storing genetic variants

- Experimentally possible, but infeasible, to directly measure haplotypes over the whole genome.
- Cheaper and more efficient to measure **genotypes** (counts of minor allele).
- Genotyping loses information, which we need algorithms and statistical models to recover (**phasing, imputation**).

**Haplotypes**

0 0 1 0 1 1 0 (maternal)
0 1 1 0 0 1 0 (paternal)

**Genotypes**

0 1 2 0 1 2 0
Cataloguing common human variation

• Step 1: Sequence lots of individuals to discover variants
• Step 2: Catalogue common variants & haplotype blocks
• Step 3: Genotype much larger set of individuals
• Step 4: Estimate population-specific properties

• Key projects
  • HapMap (2003-2009)
  • 1000 Genomes (2008-present)
Discovering genetic variation: sequencing

We have previously seen high throughput sequencing applied to measure molecular phenotypes such as gene expression and histone modification.

What happens when reads align, but only with a mismatch? Positions which don’t match could be sequence variants.

Statistical methods needed to distinguish true variants from errors (variant calling).
Whole genome variant calling: GATK HaplotypeCaller

1. Use heuristic to find mismatches not explained by noise

2. Use assembly graph to identify possible haplotypes

3. For each haplotype, estimate $P(\text{read} \mid \text{haplotype})$ using probabilistic sequence alignment
   - Hidden Markov Model
   - States: insertion, deletion, substitution
   - Emissions: pairs of aligned nucleotides/gaps
   - Transitions: equivalent to insertion/deletion/gap penalties from Smith-Waterman algorithm (DP alignment)
   - Get $P(\text{read} \mid \text{haplotype})$ using forward-backward algorithm

4. Use Bayes rule to get $P(\text{haplotype} \mid \text{read})$

5. Assign genotypes to each sample based on the max a posteriori haplotypes

http://gatkforums.broadinstitute.org/gatk/discussion/4148/hc-overview-how-the-haplotypecaller-works
Exome variant calling: atlas2

- Motivation: the exome has different sequence properties than the rest of the genome (e.g., substitution rates, GC content).

- Train **logistic regression classifier** to predict which mismatches are errors and which are variants
  - Training data: 1KG Exome project sequencing reads where >2 reads align with a mismatch
  - True positives: Reads where mismatch is also discovered in 1KG Exon pilot project
  - True negatives: Remaining reads
  - Features: mismatch quality score, flanking quality score, whether neighboring nucleotides were swapped, normalized distance to 3’ end of the read

- Much faster than full Bayesian model (e.g. HaplotypeCaller), lower false positive rate in validation data

Bamshad et al. *Nat Rev Genet* 2011
Cataloguing genetic variants: International Hapmap Project

- Goal: understand the haplotype structure of human populations (shared variation within and between groups of humans; discussed in the next section)
- Fundamental knowledge enabling GWAS, eQTL, etc.
- Systematically catalog 3.1M SNPs which exist in the human population
- Study 11 subpopulations spanning the globe
- Infer haplotypes (co-inherited genetic variation) in the population
- Genotyped 270 individuals using DNA microarrays

<table>
<thead>
<tr>
<th>Population</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW</td>
<td>87</td>
</tr>
<tr>
<td>CEU</td>
<td>165</td>
</tr>
<tr>
<td>CHB</td>
<td>137</td>
</tr>
<tr>
<td>CHD</td>
<td>109</td>
</tr>
<tr>
<td>GIH</td>
<td>101</td>
</tr>
<tr>
<td>JPT</td>
<td>113</td>
</tr>
<tr>
<td>LWK</td>
<td>110</td>
</tr>
<tr>
<td>MXL</td>
<td>86</td>
</tr>
<tr>
<td>MKK</td>
<td>184</td>
</tr>
<tr>
<td>TSI</td>
<td>102</td>
</tr>
<tr>
<td>YRI</td>
<td>203</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1397</strong></td>
</tr>
</tbody>
</table>

International HapMap 3 Consortium *Nature* 2010
Cataloguing genetic variants: Thousand Genomes Project

- 2,504 whole genome sequences at low depth (4x) across 26 subpopulations spanning the globe
- Develop sophisticated statistical tools (phasing, imputation) to account for noise, known patterns of variation (linkage disequilibrium; next section)

Thousand Genomes Consortium Nature 2016
Measuring known genetic variation: genotyping

- Key insight: Most genetic variants in an individual are recurrent in the population. Once they’ve been discovered/catalogued, build a common array for measuring them.

- DNA microarrays were the key technological advance of the 1990s.

- Idea: fragments of sample DNA containing SNPs will hybridize (reverse complement) to array probes (engineered DNA fragments).

- Tag fragments with fluorescent compound, use intensity to recover which probes were bound, which alleles were present in the sample.

- Today, still the fundamental technology used in large-scale population genetic assays (GWAS, 23andMe).

- We now study disease associations across populations, requiring new array designs due to differences in polymorphisms, LD across populations.

Image credit: Wikimedia Commons
Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   - SNPs, indels, copy-number, structural, short tandem repeats.
   - Detection: 1000Genomes project, ExAC, rare/common/neutral.
   - Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   - Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   - 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering (F_{ST}). PC correction

4. Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales
Linkage disequilibrium: D and D’

- Genetic variants do not segregate independently.
- Linkage disequilibrium measures the degree of departure from Mendel’s laws.
- \( D_{AB} = P_{11}P_{00} - P_{10}P_{01} = 0.07 \).
- If independent, then \( D_{AB} = 0 \) \( (P_{11}P_{00} = P_{10}P_{01}) \).

How to interpret non-zero values?
- Relative to \( D_{AB_{\text{max}}} \), which depends on frequencies of individual alleles at A, B.
- \( D_{AB_{\text{max}}} = P_{0*}P_{*1} - P_{1*}P_{*0} = 0.138 \).
- \( D' = D / D_{\text{max}} = 0.51 \).

\( \Rightarrow \) 51% of max possible disequilibrium.

### Haplotype AB

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Marginal allele frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0*</td>
<td>0.54</td>
</tr>
<tr>
<td>1*</td>
<td>0.46</td>
</tr>
<tr>
<td>*0</td>
<td>0.30</td>
</tr>
<tr>
<td>*1</td>
<td>0.60</td>
</tr>
</tbody>
</table>

### Expected and Observed Values

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0.162</td>
<td>0.24**</td>
</tr>
<tr>
<td>01</td>
<td>0.324</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>0.138</td>
<td>0.07**</td>
</tr>
<tr>
<td>11</td>
<td>0.276</td>
<td>0.39**</td>
</tr>
</tbody>
</table>
Linkage disequilibrium: $r^2$

- Define

$$r^2 = \frac{D^2}{P(A=0)P(B=0)P(A=1)P(B=1)} = 0.37$$

- This really is the squared Pearson correlation of the two SNPs

- In practice, Pearson correlation is efficiently computed for all SNPs in windows as $X'X/n$

- This is a fundamental quantity for modeling GWAS z-scores

<table>
<thead>
<tr>
<th>Haplotype AB</th>
<th>Marginal allele frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0*</td>
<td>0.54</td>
</tr>
<tr>
<td>1*</td>
<td>0.46</td>
</tr>
<tr>
<td>*0</td>
<td>0.30</td>
</tr>
<tr>
<td>*1</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0.162</td>
<td>0.24</td>
</tr>
<tr>
<td>01</td>
<td>0.324</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>0.138</td>
<td>0.07</td>
</tr>
<tr>
<td>11</td>
<td>0.276</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Key property: $r^2$ correlation for individual SNPs is exactly the $r^2$ of the GWAS association summary statistics of these SNPs
r^2 and recombination events across regions/populations

- Recurrent recombination events occur at hotspots

- $r^2$ correlations between SNPs depend on historical order in which they arose (not in their physical order on the chromosome)
Meiosis, recombination & double-stranded breaks

• Recombination is thought to be crucial for lining up chromosomes during meiosis for gamete formation.

• Recombination starts with a double-stranded break (DSB), which is repaired by strand invasion of the homologous chromosome.

• Repair can lead to either:
  • Gene conversion, via strand displacement annealing (SDSA), which transfers a segment of one homologous chromosome into the other, or
  • Recombination via cross-over repair of a double-stranded break, leading to

• is thought to be the fundamental selective advantage for sexual reproduction

Meiosis and recombination

- Recombination does not happen uniformly over each chromosome.
- Recombination hotspots occur once every 100kb, and recombinations occur hundreds of times more often in hotspots.
- Mouse studies revealed the role of PRDM9 in demarcating hotspots.

PRDM9, recombination, and selection (aka. The tragic love story of PRDM9)

- PRDM9 is a zinc finger protein which binds to specific DNA motifs, methylates H3K4 surrounding the binding site, and recruits double-strand break enzymes.
- PRDM9 is under strong constraint, but the DNA-binding zinc finger array has high mutation rate and is under **positive selection**.
- More than 40 known PRDM9 alleles, each with different DNA-binding specificity.
- The repaired double strand break no longer contains the PRDM9 motif, leading to evolutionary competition between the protein and its motif.

Baudat et al. *Nat Rev Genet* 2013
Mutational history of multiple haplotypes

- Example region: 36 SNPs spanning 9kb
- In principle: $2^{36}$ possible allele combinations (haplotypes)
- Sample 120 parental European chromosomes.
- In practice: only 5 recurrent haplotypes seen (and 2 singleton haplotypes)
• Relatively few haplotypes exist in the human population (consider 10M SNPs: we don’t see $2^{10^M}$ haplotypes!)
• Implies high level of genotype sharing even for unrelated individuals

Daly et al Nat Genet 2001
Haplotype phasing

• Goal: resolve genotypes to the underlying haplotypes
• Requires auxiliary information: parent genotypes, LD
Trio phasing

- Requires parental genotypes
- Generalizes to larger pedigrees
Trio phasing

• Homozygous sites can be trivially phased
Trio phasing

- If (at least) one parent is homozygous → No ambiguity left
- If both parents are heterozygous → Need LD to resolve remaining ambiguous sites
Phasing unrelated individuals

- Modern analyses often consider collections of unrelated individuals
- We don’t have pedigree information, so we can only use patterns of LD
- Input: phased haplotypes in a reference panel, observed genotypes in a population sample

Delaneau et al. *Nat Genet* 2012
Phasing unrelated individuals

- The unobserved haplotypes underlying the observed genotypes can be traced back to a common ancestor with the reference panel (ancestral recombination graph, L20)

- Directly fitting the ARG is intractable for large samples, so approximate it

- Generate each of the unobserved haplotypes by copying segments from the reference haplotypes, such that the resulting genotypes match the observations

Delaneau et al. *Nat Genet* 2012
Phasing unrelated individuals

- Recombination corresponds to switching which reference haplotype we copy from
- The model needs to infer where recombination events occur
- Trying all possible sets of recombination events is intractable, so slide a window over the genome and consider two possibilities:
  - Recombination at the right edge
  - No recombination
- Compress the redundancy in the reference haplotypes by representing as paths through a graph

Delaneau et al. Nat Genet 2012
Phasing unrelated individuals

- Only some paths through the haplotype graph are consistent with the observed genotypes.
- We know that the sum of the unobserved haplotypes must equal the observed genotypes.

Delaneau et al. *Nat Genet* 2012
Phasing unrelated individuals

- Use a Hidden Markov Model to compute the posterior probability of all feasible paths through the haplotype graph (forward-backward algorithm)
  - Hidden state: at each position, for each observed individual, the two reference haplotypes to copy
  - Emissions: for each individual, unobserved haplotypes (allowing substitutions not seen in the reference individuals)
  - Transitions: probability of a recombination event (switching which haplotype to copy)

Delaneau et al. *Nat Genet* 2012
### Genotype imputation

- **Goal:** infer the genotypes at unobserved SNPs
- **Requires** phased reference haplotypes

```
0 0 1 1 1 0 0 1 1 0 0 0 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 0 0 1
1 1 1 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 1 1 1 1 0 0 1
1 . . . 2 . 0 . . . . 0 1 . 1
1 . . . 1 . 0 . . . . 0 . 0
0 . . . 1 . 1 . . . 1 0 . 1
```

- **Goal:** infer the genotypes at unobserved SNPs
- **Requires** phased reference haplotypes
Genotype imputation

- Intuition: same haplotype copying model for phasing applies
- Phase alleles for genotyped SNPs
- Also copy the alleles for the unobserved variants
- Recover genotypes by summing the inferred haplotypes

\[
\begin{array}{cccccccccccccccc}
0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\
\end{array}
\]

\[
\begin{array}{cccccccccccccccc}
1 & 1 & 2 & 2 & 2 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & \ldots & 1 & . & 0 & \ldots & \ldots & 0 & . & 0 \\
0 & \ldots & 1 & 1 & \ldots & 1 & 0 & . & 1 \\
\end{array}
\]
Fine-mapping disease associations: (1) Epigenomics / functional data (next lecture)

- **Association mapping** refers to identifying variants/gene associated with disease
- This is confounded by LD
- Many variants are strongly correlated to the true causal variant, and will show nearly as strong associations
- Use estimated correlations to explain correlated associations and recover the true underlying effects

Li and Kellis *BiorXiv* 2016
Fine-mapping disease associations
(2) Multi-ethnic analysis

Case 1: LD boundaries differ

Case 2: allele frequencies differ

• Allele frequencies and LD patterns can differ between populations
• Currently, disease associations are biased for discovery in European cohorts
• As we begin conducting association studies in Asia/Africa, there is a pressing need to develop statistical methods which can account for population genetic differences

Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   - SNPs, indels, copy-number, structural, short tandem repeats.
   - Detection: 1000Genomes project, ExAC, rare/common/neutral.
   - Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   - Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   - 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering (F_{ST}). PC correction

4. Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales
Identity by descent in related individuals

- Parents share 50% of their DNA with their children
- Siblings share 50% of their DNA with each other
- These are not the same 50%!
Identity by descent in related individuals

Parent-child
(100% of the genome is 50% identical)

Siblings:
50%: half the genome
100%: quarter of genome
0% quarter of genome

Figure credit: Lucile Kellis
Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   - SNPs, indels, copy-number, structural, short tandem repeats.
   - Detection: 1000Genomes project, ExAC, rare/common/neutral.
   - Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   - Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   - 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering (FST). PC correction

4. Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales
Number of variants varies greatly by population

- Over 100 million observed variants: 4-5M positions differ between each of us and the human reference
- Each of us carries 2-3K structural variants affecting 20mb of sequence
- Each of us carries hundreds of protein truncating variants, 10Ks of non-synonymous mutations
- African individuals have more variation in their genomes (why?)

Thousand Genomes Consortium Nature 2016
Population size, bottlenecks and expansion

- **Effective population size**: number of individuals needed in idealized model to recapitulate population properties
- Here, recapitulate the **coalescent time** (next section, L20): time to most recent common ancestor
- **Pairwise Markov sequential coalescent model** with population splits/growth enables comparison within vs. between populations
- 1KG suggests shared history beyond 150 kya
- Non-African populations sustained a **bottleneck** 15-20 kya (migration out of Africa)
- After migration, rapid population expansion (with interesting exceptions: Finland, Peru, Mexico)

Ancestry painting (e.g. admixed individual)

Ancestry Composition tells you what percent of your DNA comes from each of 31 populations worldwide. This analysis includes DNA you received from all of your recent ancestors, on both sides of your family. The results reflect where your ancestors lived before the widespread migrations of the past few hundred years.

- **79.0%** Sub-Saharan African
- **72.3%** West African
- **2.9%** Central & South African
- **3.8%** Broadly Sub-Saharan African
- **18.4%** European
  - **Northern European**
  - **2.5%** British & Irish
  - **0.2%** Scandinavian
  - **11.4%** Broadly Northern European
  - **0.6%** Ashkenazi
  - **Southern European**
  - **0.5%** Broadly Southern European
  - **3.3%** Broadly European
- **1.9%** East Asian & Native American
- **0.8%** Native American
- **0.8%** Southeast Asian
- **0.2%** Broadly East Asian & Native American

0.7% Unassigned

100% TL Dixon

Which segments of a genome are shared with what populations
Ancestry painting using Latent Dirichlet Allocation (LDA)

- Originally proposed for document classification
- Words are associated with multiple topics
- Documents are associated with multiple topics, depending on which words they contain
- Here, individuals are documents, variants are words, and populations are topics
Latent Dirichlet Allocation (LDA)

- Assume each variant $l$ is independent
- The observed genotype $G$ arises from two unobserved haplotypes
- Each individual’s genome has probability (mixture proportion) $Q$ of coming from each ancestral population
- That means each haplotype $Z$ (per individual) has the same mixture proportion
- The allele frequency $P$ of each variant could differ in each ancestral population

The observed genome of an individual depends on:
(a) their ancestry mixture
(b) the genotype frequency in the corresponding populations

Challenging model to fit:
- ADMIXTURE (maximum likelihood)
- fastSTRUCTURE (variational)
- Terastructure (stochastic variational)
RFMix

• Idea: use random forest classifiers to predict the source population of each local segment

• Input: Observed allele frequencies in reference populations

• Output: which population did the observed segment come from

• Use a conditional random field to model the correlations between neighboring segments

• Intuition: switching ancestry every segment is unlikely (like HMM)

Maples et al. Am J Hum Genet 2013
Ancestry painting: population-level

**Goal:** infer ancestry of segments of the genome, population structure (patterns of relatedness between ancestry groups)

Sharing of genetic variants enables ancestry painting of individual genomes

The history of migration, settlement, conquest is written on our genomes

Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   – SNPs, indels, copy-number, structural, short tandem repeats.
   – Detection: 1000Genomes project, ExAC, rare/common/neutral.
   – Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   – Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   – 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   – Admixture, ancestry painting. PCA/MDS clustering (FST). PC correction

4. Human demographic history.
   – Population size, dynamics, bottlenecks, expansion, effective pop size.
   – Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales
Genetic relatedness and geography

- Can we decompose genetic variation into the major forces shaping it?

  ➔ PCA/SVD decomposition

- First components correspond to population structure.

- Population structure is shaped by geography! (people near each other are more likely to mate)

- In Europe, First two components correspond to N-S and E-W migration axes

- Country neighbors & borders visible at the genetic level

Principal components analysis

- Idea: find a new basis for high dimensional data
- First principal component (axis) must explain most variation
- Second PC is orthogonal and explains maximum amount of the remaining variation
- PCs computed using **singular value decomposition**

Importance of PCA for GWAS

- Geography/demography causes genome-wide differences between individuals
- Confounds specific differences between individuals with different phenotypes
- Two main strategies to correct for population structure:
  - Include principal components as fixed covariates in linear model
  - Incorporate polygenic effect of population structure using linear mixed models

Measuring divergence between populations

- The **fixation index** $F$ measures the loss in heterozygosity in a population.
- Originally developed by Wright for studying livestock inbreeding.
- The special case $F_{ST}$ compares pairs within a subpopulation against pairs between subpopulations.

<table>
<thead>
<tr>
<th>Population</th>
<th>AA freq</th>
<th>AG freq</th>
<th>GG freq</th>
<th>Eff pop size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFR</td>
<td>0.16</td>
<td>0.48</td>
<td>0.16</td>
<td>10000</td>
</tr>
<tr>
<td>EUR</td>
<td>0.01</td>
<td>0.18</td>
<td>0.81</td>
<td>4000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0.117</td>
<td>0.394</td>
<td>0.489</td>
<td>14000</td>
</tr>
</tbody>
</table>

\[
F_{ST} = \frac{H_{total} - \bar{H}}{H_{total}} = \frac{0.394 - 1/2(0.48 + 0.18)}{0.394} = 0.16
\]
Measuring divergence between populations

- $F_{ST}$ equivalently compares the average coalescent time (time to most recent common ancestor) of pairs within a population to pairs between populations.
- The underlying model also allows us to estimate effective population sizes from observed levels of heterozygosity.
• Under suitable assumptions, Euclidean distance in the first principal components corresponds to the **mean coalescent time**
• Predict exactly where individuals will appear in principal component space
• Illustrates a deep connection between population genetics and evolution
Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   – SNPs, indels, copy-number, structural, short tandem repeats.
   – Detection: 1000Genomes project, ExAC, rare/common/neutral.
   – Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   – Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   – 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   – Admixture, ancestry painting. PCA/MDS clustering (F_{ST}). PC correction

4. Human demographic history.
   – Population size, dynamics, bottlenecks, expansion, effective pop size.
   – Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales
Population dynamics

- Many models in population genetics assume population sizes are constant (coalescent process)
- The pairwise sequentially Markov coalescent is used to model population growth

ARGs: Ancestral recombination graphs

• Idea: history of a locus across individuals described by a coalescent tree (like a phylogenetic tree)
• Loci on one haplotype are explained by the same tree
• Recombinations correspond to changing the tree between positions
• Model as pairwise sequentially Markov coalescent process
• Build an HMM where hidden states are coalescent trees (ancestral recombination graph)
• Integrate over all paths to estimate divergence times, population sizes

• Mitochondrial DNA is inherited only from the mother, doesn’t recombine appreciably
• Accumulation of mutations in mtDNA corresponds to population divergence

Stewart et al. Nat Rev Genet 2015
Evidence for **only one** out of Africa migration

Analysis of aboriginals in Australia showed that:

- they share **the same bottleneck** (~50k years ago) with Eurasians
- show divergence from Eurasians shortly thereafter (50-70k years ago)

⇒ same out-of-Africa migration event as Eurasians!

Recent migration into India

- North Indians show evidence of admixture with Eurasians, unlike South Indians
- Evidence suggests migration 1-4KYA, concordant with history/mythology
Recent migration into India

- Evidence of a **cline** (gradient) of admixture going from Europe, to North India, to South India.

Moorjani et al. *Am J Hum Genet* 2013
Present day South Americans are a mixture of Native American, European, and African ancestry.

Sex bias in inferred ancestry: increased Native American ancestry on the X chromosome (why?)
Conquest of South America

- Admixture of European and Native American populations obvious from PCA, ancestry painting
- Cline of Native American admixture following geography
- Cline of European admixture in Argentina recapitulates history (arrived through the Andes, spread to the Atlantic)

Recent admixture in African Americans

• Present day African Americans are an admixture of European, African, Native American ancestries
• Regional differences in proportion of European ancestry
• Possibly explained by one migration event (1802) or two events (1714, 1854).
• 2% of all non-African genomes derives from the Neanderthal genome
• The Ust’ Ishim man (45 KYA) also has 2% Neanderthal DNA!
• Interbreeding between *H. sapiens* and Neanderthal 50-60 KYA
South Asians have approx. 5% ancestry from Denisovans, another hominid subspecies.

They have larger Denisovan haplotypes than Neanderthal fragments, implying ordering of interbreeding events (why?).

More Denisovan, Neanderthal ancestry on the X chromosome.
Summary: re-writing human population history

- Bottlenecks/founder effects: rare alleles suddenly rise in frequency due to small population size
- Selective sweeps: rare alleles suddenly rise in frequency due to positive selection
- Loss of heterozygosity in non-African populations
- Admixture between previously isolated populations
- Interbreeding between ancient humans and other hominids
Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   - SNPs, indels, copy-number, structural, short tandem repeats.
   - Detection: 1000Genomes project, ExAC, rare/common/neutral.
   - Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   - Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   - 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering (F_{ST}). PC correction

4. Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales
Molecular evolution

- Fundamental question: how does allele frequency change over time?

- Under strong assumptions (no mutation, no migration, no selection, large population size, random mating), allele/genotype frequencies remain in **Hardy-Weinberg equilibrium**

<table>
<thead>
<tr>
<th>Allele</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.6</td>
</tr>
<tr>
<td>G</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$0.6 \times 0.6 = 0.36$</td>
</tr>
<tr>
<td>AG/GA</td>
<td>$2 \times 0.6 \times 0.4 = 0.48$</td>
</tr>
<tr>
<td>GG</td>
<td>$0.4 \times 0.4 = 0.16$</td>
</tr>
</tbody>
</table>
Molecular evolution

• Under strong assumptions (no selection, constant population size, random mating, no overlapping generations), change in allele frequency of a new mutation over time is described by the Wright-Fisher process.

• Without selection, main force on allele frequency is genetic drift (random chance).

• Natural selection changes the course of the process (either towards 0 or 1).
Molecular evolution

• Fundamental question: how have allele frequencies changed over time?
• Different methods give insight into change over different time scales
Purifying selection on coding sequences

- Selection against new mutations drives allele frequency to zero
- Exome Aggregation Consortium (60K samples!) has shed new insight into purifying selection on protein-coding sequences
- Idea: predict how many new mutations should fall in a gene given its properties, then compare to the observed number
- Discover excess of de novo mutations in rare disease cases (ASD)

Samocha et al. Nat Genet 2014
Purifying selection on non-coding sequences

- Selection against common mutations must be weaker (why?)
- Leads to difference in heterozygosity at SNPs in predicted regulatory regions

Ward and Kellis *Science* 2012
Purifying selection also operates on transcription factor binding sites

Ward and Kellis Science 2012
Purifying selection on non-coding sequences

- Purifying selection on putative functional elements also leads to reduction in derived allele frequency

Ward and Kellis *Science* 2012
Test for positive selection

• Mutations which were beneficial to *H. sapiens* specifically will not be conserved with other primates

• The **McDonald-Kreitman test** identifies sites which are under selection within a species

• Idea: compare the amount of variation within the species (polymorphism) with the amount between species (substitutions)

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymous</td>
<td>$D_s$</td>
<td>$P_s$</td>
</tr>
<tr>
<td>Non-synonymous</td>
<td>$D_n$</td>
<td>$P_n$</td>
</tr>
</tbody>
</table>

\[
\frac{D_n}{D_s} < \frac{P_n}{P_s} \text{ implies purifying selection} \\
\frac{D_n}{D_s} > \frac{P_n}{P_s} \text{ implies positive selection}
\]
Recent positive selection

- Signature of recent positive selection: common variants on long haplotypes (why?)
Positive selection on lactase persistence

- The gene LCT encodes the enzyme lactase, required to digest milk
- Evidence of positive selection on European LCT: the persistence haplotype is present in 77% of Europeans, extends 1MB

Sabeti et al. Science 2006
Positive selection on height

- Difference in height between northern and southern Europeans driven by selection
- Analysis of ancient Eurasians reveals selection for reduced height in Iberia, increased height in steppe populations

Goals for today: Population genomics

1. Genetic variation: detection, quantification, and initial insights
   - SNPs, indels, copy-number, structural, short tandem repeats.
   - Detection: 1000Genomes project, ExAC, rare/common/neutral.
   - Variant calling from sequencing reads, logistic regression, joint calling.

2. Haplotypes, Recombination, LD, Phasing
   - Biology: meiosis, recombination, hotspots, PRDM9 protein/motif evol.

3. Human relatedness and ancestry painting
   - 23andMe parent-child vs. sibling 50%. SNP sharing. IBD/heritability.
   - Admixture, ancestry painting. PCA/MDS clustering (FST). PC correction

4. Human demographic history.
   - Population size, dynamics, bottlenecks, expansion, effective pop size.
   - Ancient DNA, Neanderthals, Denisovans. Gorilla/Chimp introgression.

5. Measuring human selection at multiple timescales