Agenda

• Population Genetic Inference
  – Mutation
  – Selection
  – Recombination
The Coalescent Process

- “Backward in time process”
- Discovered by JFC Kingman, F. Tajima, R. R. Hudson c. 1980
- DNA sequence diversity is shaped by genealogical history
- Genealogies are unobserved but can be estimated
- Conceptual framework for population genetic inference: mutation, recombination, demographic history
2 sample coalescent

MRCA

sequence1 sequence2

$T_2$

N = population size of diploid individuals

n = sample size of haploid chromosomes

MRCA = most recent common ancestor

$T_2 = \text{coalescence time for 2 chromosomes}$
2 sample coalescent

Probability that the time of MRCA is $t$ generations ago

$$P(T_2 = t) = \left(1 - \frac{1}{2N}\right)^{t-1} \left(\frac{1}{2N}\right)$$

If we “blur” our eyes a bit (as $N$ gets very large) this becomes

$$P(T_2 = t) = \left(\frac{1}{2N}\right)^t e^{-\left(\frac{1}{2N}\right)t}$$
2 sample coalescent

If we consider $t' = t/2N$, call that “coalescent time”

Then $E(T_2) = 1$
There are \( \binom{n}{2} = \frac{n(n-1)}{2} \) possible pairs, each coalescing at rate \( 1/2N \)

\[
P(T_n = t) = \binom{n}{2} \binom{2N}{n} e^{-\binom{n}{2} / 2N} t
\]

\[
E(T_n) = \frac{2N}{\binom{n}{2}}
\]
\[ E(\text{TMRCA for } n \text{ chromosomes}) = T_2 + T_3 + T_4 \ldots + T_n \]
\[ = 2(1 - 1/n) \text{ coalescent units} \]
In humans, it is known that “appropriate” values for $N_e$ are surprisingly small. This approximation is called the “effective population size”:

- $N_e \approx 10,000$ in Europe
- $N_e \approx 9,500$ in East Asia
- $N_e < 25,000$ for all human populations, highest in Africa
**$N_e$ and coalescence times in humans and other animals**

The mean coalescence time for two lineages is just $E(T_2) = 1$ in units of $2N_e$ generations, so if we have $G = 22$ years per generation, the average ancestry depth for 2 human chromosomes is

$$1 \times 2N_e \times G \text{ in years}$$

$$(20,000-50,000) \times 22 = 440,000-1,100,000 \text{ years}$$

$N_e$ varies **widely** across species (Charlesworth, Nature Reviews Genetics 2009):

- 25,000,000 for E.coli
- 2,000,000 for fruit fly D. Melanogaster
- <100 for Salamanders (Funk et al. 1999)
Adding mutations

For neutral models, can separately model the genealogical process (the tree) and the mutation process (genetic types)

- Infinite sites mutation model
What is expected number of mutations between 2 chroms?

\[ t \sim \text{Expo} \left( 2N \right) \]
\[ \pi \sim \text{Pois} \left( 2t\mu \right) \]

\[ \mu = \text{mutation rate per bp per generation} \]
\[ \pi = \text{number of sequence changes between 2 chroms} \]

\[ E(t) = 2N \quad E(\pi | t) = 2t\mu \]

\[ E(\pi) = 2\mu E(t) = 4N\mu \]

4Nu comes up repeatedly in population genetics, often referred to as theta

\[ \theta = 4N\mu \]
Number of segregating sites in a sample of size \( n \)

The total time in the tree for a sample of \( "n" \) chromosomes is

\[
L = 4t_4 + 3t_3 + 2t_2
\]

or in general:

\[
L = \sum_{i=2}^{n} i \cdot t_i
\]

Watterson Estimator for population scaled mutation rate

\[
\hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}
\]
Estimators of Theta

Watterson estimator is just one approach, uses summary of the data

\[ \hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}} \]

There is also the Tajima estimator – “Average # of pairwise differences”

\[ \hat{\theta}_T = \frac{n}{n-1} \sum_{i=1}^{S} 2p_i(1-p_i) \]

\( p_i = \) allele frequency of mutation \( i \)
\( S = \) number of polymorphic sites
The site frequency spectrum

Under standard neutral model, the site frequency spectrum is a beta distribution with parameters $\theta/2, \theta/2$.

Plot of theoretical SFS for 1Mb
The \textit{sfs} under neutrality and selection
The *sfs* of genic variants

splice-disrupting 621
stop-gain 1,654
non-synonymous 84,358
synonymous 61,155

Daniel MacArthur, Suganti Balasubramaniam
Sample-based estimators of $\Theta$ using the sfs

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Sensitivity</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_w = \frac{1}{n-1} \sum_{i=1}^{n-1} \xi_i$</td>
<td>low</td>
<td>Watterson (1975)</td>
</tr>
<tr>
<td>$\theta_{\pi} = \frac{1}{n^2} \sum_{i=1}^{n-1} i(n-i) \xi_i$</td>
<td>intermediate</td>
<td>Tajima (1989)</td>
</tr>
<tr>
<td>$\theta_{\xi_e} = \xi_e = \xi_1$</td>
<td>singleton</td>
<td>Fu and Li (1993)</td>
</tr>
<tr>
<td>$\theta_H = \frac{1}{2n} \sum_{i=1}^{n-1} i^2 \xi_i$</td>
<td>high</td>
<td>Fay and Wu (2000)</td>
</tr>
<tr>
<td>$\theta_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i \xi_i$</td>
<td>high</td>
<td>Zeng et al. (2006)</td>
</tr>
</tbody>
</table>

**Sensitivity** = the frequency of observed polymorphisms that makes estimates using a given estimator large relative to the others.
Tajima’s $D$

$$D_t = \frac{\theta_\pi - \theta_W}{C}$$

$D_t = 0$ neutral evolution

$D_t > 0$ balancing selection, more intermediate variants

$D_t < 0$ positive selection
a Genealogies

Locus under positive selection

Locus under balancing selection

Locus under no selection (neutral)

b Haplotypes

c Site frequency spectra

\[ \pi = 0.063 \quad D = -0.89 \]

\[ \pi = 0.085 \quad D = 0.40 \]

\[ \pi = 0.076 \quad D = 0.06 \]
A typical population genomics study design for detecting positive selection.

1. Sample loci and calculate statistic ($T_j$)

2. Construct empirical distribution

3. Identify “outlier” loci

Akey J M Genome Res. 2009;19:711-722
Are humans still experiencing adaptive evolution?
A Map of Recent Positive Selection in the Human Genome

Benjamin F. Voight®, Sridhar Kudaravalli®, Xiaoquan Wen, Jonathan K. Pritchard*

Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America

iHS(x), a homozygosity-based statistic
- function of test allele frequency
- summarizes extent of haplotype homozygosity of derived allele chromosomes compared to ancestral
- built-in control for local recombination rate

PLoS Biology 2006
Table 2. p-Values for Enrichment of GO Categories among Genes Showing Evidence for Partial Sweeps

<table>
<thead>
<tr>
<th>GO Nesting</th>
<th>GO Category</th>
<th>ASN</th>
<th>CEU</th>
<th>YRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>21-1</td>
<td>Chemosensory perception</td>
<td></td>
<td>0.0006</td>
<td>0.0004</td>
</tr>
<tr>
<td>21-1-1</td>
<td>Olfaction</td>
<td></td>
<td>0.0006</td>
<td>0.0008</td>
</tr>
<tr>
<td>22-2</td>
<td>Gametogenesis</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22-2-2</td>
<td>Spermatogenesis and motility</td>
<td>0.02</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>22-3</td>
<td>Fertilization</td>
<td>0.004</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>1-11</td>
<td>Other carbohydrate metabolism</td>
<td>0.0002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Electron transport</td>
<td></td>
<td></td>
<td>0.0002</td>
</tr>
<tr>
<td>4-13</td>
<td>Chromatin packaging/remodeling</td>
<td></td>
<td>&lt;0.0001</td>
<td>0.01</td>
</tr>
<tr>
<td>16-1-1</td>
<td>MHC-I-mediated immunity</td>
<td></td>
<td>&lt;0.0001</td>
<td>0.02</td>
</tr>
<tr>
<td>3-2</td>
<td>Steroid metabolism</td>
<td></td>
<td></td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>3-5</td>
<td>Lipid and fatty acid binding</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-4-2</td>
<td>mRNA transcription initiation</td>
<td></td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>5-3</td>
<td>Protein modification</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7-5</td>
<td>Vitamin/cofactor transport</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Phosphate metabolism</td>
<td>0.002</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>13-4</td>
<td>Peroxisome transport</td>
<td></td>
<td></td>
<td>0.002</td>
</tr>
</tbody>
</table>

(a) East Asians, rs6060371 (in SPAG4), $p_d = 0.742$, 2.3 cM/Mb
Conflicting evidence of population-specific selection

(A) Genic SNPs are more likely than non-genic SNPs to have extreme allele frequency differences

Conflicting evidence of population-specific selection

(B) Maximum allele frequency difference is well explained by average genetic distance between populations

Polygenic model of adaptation

Pritchard, et al Curr Biology, 2010
Linkage Disequilibrium

“nonrandom associations between alleles”

• Compare to HWE:
  – Under HWE, gametes unite at random. $\Pr(A,a) = \text{pr}(A) \times \text{pr}(a)$ where $A$ and $a$ are alleles at same locus

• LD statistics measure to what extent $\Pr(A,B) = \text{pr}(A) \times \text{pr}(B)$ when $A$ and $B$ are alleles at different loci

• Example applications: mapping genes and measuring recombination
Linkage Disequilibrium

Frequencies:

\[ P_A \quad P_B \]
\[ P_a \quad P_b \]

Define \( D = P_{AB} - P_A P_B = P_{AB} P_{ab} - P_{Ab} P_{aB} \)

- If \( D \neq 0 \) then we have LD
- \( r^2 = D/(P_A P_B P_a P_b) \)
Where does LD come from?

Final Aligned Data Set:
LD-based Case-Control Association Study

- locate disease locus
- Unlikely to be among our genotyped markers
  → Detect indirectly using available markers
Where does array content come from?

Publicly Funded Genomics Projects

- Human Genome Project
- Phase I HapMap Project
- CNV Project
- Phase II HapMap
- 1000 Genomes Project

Number of SNPs on an Affymetrix Chip

- 10k
- 500k
- 1M
The International HapMap Project

-HapMap project (2002-2008) cost $120 Million USD

-Measured variation at 4 million SNPs in 4 populations (90 Europeans, 90 Nigerians, 45 Chinese, 45 Japanese)

Results:

- Over 2.8 M SNP with > 5% allele frequency

-80% of variants within each population can be captured with 30% of the most informative SNPs, “tagSNPS”

-Nigerians require most tag SNPs, followed by Europeans and then Asians
HapMap tagSNPs are useful for other populations

Conrad, et al. (2006) genotyped 3000 SNPs in 52 populations across the globe (the “Human Genome Diversity Panel” or HGDP)

C+J: Asian, CEU: European, YRI: Nigerian
The serial bottleneck model
The recombination rate

Can vary hugely along a sequence

Determines association between loci in the population

Is hard to measure directly, because recombination occurs on average only $\sim 1$ in 100,000,000 meioses between any pair of successive nucleotides in the genome.

Can be measured indirectly, by parametric analysis of variation data. Researchers in Oxford, and elsewhere, have developed such parametric approaches (Li and Stephens, 2003; Ptak et al. 2005; Hudson 2001, McVean 2002, McVean et al. 2004)

Now, we are considering $\rho = 4Nr$
Marginal Trees

Time

Position A

Position B

Position C

Physical position
The Fine-Scale Structure of Recombination Rate Variation in the Human Genome

Gilean A. T. McVean, Simon R. Myers, Sarah Hunt, Panos Deloukas, David R. Bentley, Peter Donnelly

Calculated a “composite likelihood” for a sample of haplotypes:

\[ L(4Nr) = \sum_{ij} L(n_i, n_j, n_{ij} | 4Nr_{ij}) \]

- sum over pairs
- lookup table
- RJMCMC

Fig. 2. Comparison between estimates of local recombination rates from population genetic data (red) and sperm analysis (blue) in the HLA region; data from (3). To convert the male crossing-over rates to sex-averaged rates, we used the previous observation that the female crossing-over rate in this region is about four times that of males (42).
A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium∗
A common sequence motif associated with recombination hot spots and genome instability in humans

Simon Myers\textsuperscript{1,2}, Colin Freeman\textsuperscript{2}, Adam Auton\textsuperscript{2,3}, Peter Donnelly\textsuperscript{2,4} & Gil McVean\textsuperscript{2}
Drive Against Hotspot Motifs in Primates Implicates the \textit{PRDM9} Gene in Meiotic Recombination

Simon Myers,\textsuperscript{1,2}† Rory Bowden,\textsuperscript{1,2}§ Afidalina Tumian,\textsuperscript{1} Ronald E. Bontrop,\textsuperscript{3} Colin Freeman,\textsuperscript{2} Tammie S. MacFie,\textsuperscript{4}‡ Gil McVean,\textsuperscript{1,2}§ Peter Donnelly\textsuperscript{1,2}§

\textbf{A}

\begin{align*}
 TGCCAGCT & TTTCTC TT AAGGCCCT CCCA ACC ACC CCT CT \\
\end{align*}

\textbf{B}

\begin{align*}
 GAACACCAGACGACGAAGAAGAACC GCC GT AAACC ACC GAT \\
\end{align*}
PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans

Ingrid L Berg, Rita Neumann, Kwan-Wood G Lam, Shriparna Sarbajna, Linda Odenthal-Hesse, Celia A May & Alec J Jeffreys