Extended Haplotypes

• When a mutation first occurs, it is on a single chromosomal background and in \( lD' = 1 \) disequilibrium with all other pre-existing polymorphisms on that chromosome.

• It takes time to break up this disequilibrium, so the initial haplotype defined by a new mutation can be quite extensive, even spanning across recombination hotspots.

• Slatkin (Evolution 39:53-65, 1985) showed that new mutants can be informative indicators of recent gene flow & dispersal. Extended haplotypes are a way of identifying such new mutants.

• To detect such haplotypes, see Browning and Browning Genetic Epidemiology 31: 365-375, 2007; or Climer’s program BlocBuster.

Not All Haplotypes Are of Equal Length, and Some Can Span Across Recombinational Hotspots

BlocBuster: Identification of Haplotypes of Unequal Length (Sharlee Climer, WU)

<table>
<thead>
<tr>
<th>Genotypes for Ten Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP 1</td>
</tr>
<tr>
<td>SNP 2</td>
</tr>
</tbody>
</table>

\[
D = g_{AB}g_{ab} - g_{Ab}g_{aB} = (3/20)(4/20) - (10/20)(3/20) = -0.045
\]

D is a **global** measure of association involving all alleles. Do all alleles contribute to D in this case or just some? Impossible to say just from the value of D.
Not All Haplotypes Are of Equal Length, and Some Can Span Across Recombinational Hotspots

BlocBuster: Networks of Alleles.

<table>
<thead>
<tr>
<th>Genotypes for Ten Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNP 1</strong></td>
</tr>
<tr>
<td><strong>SNP 2</strong></td>
</tr>
</tbody>
</table>

The Network makes it clear that alleles a and B show no associations with other alleles, but A and b are associated.

Eliminate all lines in network that fall below a statistically determined threshold. Implies that A-b define a 2 site haplotype, a and B define single site haplotypes.
Not All Haplotypes Are of Equal Length, and Some Can Span Across Recombinational Hotspots

BlocBuster: Networks of Alleles. Easily extended to large numbers of loci or nucleotides because based on simple pairwise comparisons.


Geographic distribution of the Asian and American populations genotyped for this study

Visual genotypes (data set WW34), clustered by population, for individuals either homozygous or heterozygous for the 9-repeat allele

Implies that this extended haplotype is identical by descent in all Western Beringians and Native Americans & has experienced less recombination than most chromosome types in this region.


Population Affiliation

- As genetic resolution increased, Smouse and Spielman (Proc. 5th Internatl. Cong. Human Genetics, pg. 255-260, 1976) first proposed that it would be theoretically possible to reverse the $f_{st}$ analysis and assign individuals to populations based on their genotypes.

- **Population Affiliation** Attempts to Assign Individuals to Specific Populations (either pre-specified or not) on the basis of molecular genetic data.

- This technique is now much more valuable because of the abundance of molecular markers.

- Can be a powerful tool for looking at population structure, gene flow, and ecology.
Given a sample that represents the sum of two or more populations, can you reassign individuals to reconstruct the component subpopulations? In the above example, there would be many ways to assign individuals from the sum to create two HW subpopulations.

Note, this criterion works best when D=0 in all subpopulations: Therefore use unlinked or loosely linked markers.

When dealing with two or more loci, can also assign individuals from the total sample to create two or more subpopulations with no or minimal D.
Genotype Frequencies in Two Populations

Population 1 (p=0.6)
Population 2 (p=0.2)

Genotype Frequencies in Two Populations

Population 1 (q=0.3)
Population 2 (q=0.7)

2 Locus Genotype Frequencies in Two Populations

Assignment Programs

Table 1 Summary of four Bayesian clustering software packages and their underlying model

<table>
<thead>
<tr>
<th>Software</th>
<th>Admixture</th>
<th>Parental populations</th>
<th>Rational</th>
<th>Prior distribution</th>
<th>Algorithm</th>
<th>Choice of K</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRUCTURE</td>
<td>Yes</td>
<td>Not required</td>
<td>Infers ancestry coefficients Minimizes departures from HW and LD disequilibria Delineates populations under Hardy-Weinberg equilibrium</td>
<td>Non spatial</td>
<td>MCMC</td>
<td>Multiple runs ln P(D</td>
</tr>
<tr>
<td>GENELAND</td>
<td>No</td>
<td>Not relevant</td>
<td></td>
<td>Colored Voronoi tiling</td>
<td>RJMCMC</td>
<td>Single run Reversible jump</td>
</tr>
<tr>
<td>BAPS</td>
<td>Yes</td>
<td>Required or provided by the mixture model</td>
<td>Seeks spatially smooth and genetically homogeneous clusters</td>
<td>Inspired from Markov Random fields (mixture) Non spatial (admixture)</td>
<td>Stochastic optimization</td>
<td>Multiple runs Split and merge</td>
</tr>
<tr>
<td>TESS</td>
<td>Yes</td>
<td>Not required</td>
<td>Models spatial trends and autocorrelation</td>
<td>Markov random field (mixture) Log-Gaussian random field (admixture)</td>
<td>MCMC</td>
<td>Multiple runs Information theoretic criterion</td>
</tr>
</tbody>
</table>

E.g., Human Population Subdivision As Shown Through An Assignment Programs

STRUCTURE analysis of 377 autosomal microsatellite loci in 1056 individuals from 52 populations. Within-population differences among individuals account for 93 to 95% of genetic variation; differences among major groups constitute only 3 to 5%. STRUCTURE found distinct subpopulations corresponding to major geographical areas:

Each individual is represented by a thin vertical line, which is partitioned into K colored segments that represent the individual's estimated membership fractions in K clusters. Black lines separate individuals of the 52 different populations sampled.

E.g., Human Population Subdivision As Shown Through An Assignment Programs

When similar data are used to calculate pairwise $f_{st}$ population genetic distances and then plotted against geographical distance, no distinct subdivisions are evident:
Serre, D., and S. Paabo. 2004. Genome Res 14:1679-85. Sampled from the same database as Rosenberg et al., but sampled individuals widely scattered geographically rather than by predefined populations. Ran STRUCTURE with $K=2$ on both sampling schemes:

As you increase the geographical resolution in humans, the “clusters” become less distinct.


Li et al. Science 319: 1100-1104, 2008


Lineage Sorting During Coalescence Can Yield Different Conclusions

BAPS clustering using 137 microsatellites in 569 Drosophila melanogaster from 21 localities around the world.

BAPS clustering using various subsets of the microsatellites in 569 Drosophila melanogaster from 21 localities around the world.


---

Measuring Gene Flow From Assignment Programs

Dispersal is inferred when individuals are assigned genetically to a population that is different from the one from which they are sampled.

Previous gene flow is inferred when individuals are inferred to be a genetic mixture of two or more populations.

These can be quantified.

Analyzed 177 juvenile panda clownfish (*Amphiprion polymnus*) in Papua New Guinea using 11 microsatellite loci (larval dispersal is extremely difficult to measure directly).

Works well if gene flow levels are low and have many highly polymorphic loci.

Inbreeding Effective Size for the Island Model with an Infinite Number of Demes

Common Gene Pool From All Local Demes

Let each local deme be “ideal”

What is the inbreeding effective size of the total (infinite) population?
When $m=0$ (isolation), $N_{\text{eff}} = N$, the local deme size. With $m>0$, $N_{\text{eff}}$ increases with increasing $m$.

$E.g., \mu = 10^{-5}$ and $m = 0.01$, then $N_{\text{eff}} = 1001N$

Consider a finite island model of $n$ local populations, each of ideal size $N$, with a gene flow rate of $m$. The total population size in this model is $nN$, and if the population were panmictic, its variance effective size would be $nN$. Wright (1943) showed:

$$N_{\text{evT}} = \frac{nN}{1 - f_{st}}$$

Note, as $m \rightarrow 1$, $f_{st} \rightarrow 0$, and $N_{\text{evT}} \rightarrow nN$

As $m$ decreases, $f_{st}$ increases, and $N_{\text{evT}} > nN$
Eastern Collared Lizard
*Crotaphytus collaris collaris*

Population Subdivision (decreasing gene flow) Decreases the Total Inbreeding Effective Size, With The Limit Under Total Isolation Being the Local Deme Size

Population Subdivision (decreasing gene flow) Increases the Total Variance Effective Size, With the Limit Under Total Isolation Being Infinity.
Warning!

• Although $f_{st}$ or $F_{st}$ can be interpreted in terms of the balance of gene flow and drift, it is only a description of the current spatial pattern of variation.

• E.g., suppose a population is split into two isolated fragments with $m = 0$. At equilibrium, $f_{st} = 1$, but this can take many generations to occur.