Population genetic inference

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>28th Jan</td>
<td>Modelling population genetic data</td>
<td>GM</td>
</tr>
<tr>
<td>4th Feb</td>
<td>Inference</td>
<td>GM</td>
</tr>
<tr>
<td>11th Feb</td>
<td>Model testing</td>
<td>GM</td>
</tr>
<tr>
<td>18th Feb</td>
<td>Extensions of the neutral model</td>
<td>GM</td>
</tr>
<tr>
<td>25th Feb</td>
<td>Recombination</td>
<td>SM</td>
</tr>
<tr>
<td>4th March</td>
<td>Recombination and haplotype structure</td>
<td>GM</td>
</tr>
<tr>
<td>11th March</td>
<td>Fine-scale mapping</td>
<td>JH</td>
</tr>
<tr>
<td>18th March</td>
<td>Genome-wide association studies</td>
<td>JM</td>
</tr>
</tbody>
</table>

The unusual nature of population genetic data

- Conventional statistical inference
  - Many independent data points
  - Sample space of low dimensions
  - Analytical formulations for inference using all possible information often possible

- Population genetic data
  - Typically a single draw from the evolutionary process
  - Sample space of many dimensions
  - Analytical formulations for inference using ALL possible information usually impossible to derive

Statistical inference

- Define model
- Explore properties
- Estimate parameters from data
- Test goodness-of-fit
- Refine model

Issues

- summary statistics
- graphical representation
- stochastic simulation
- moment methods
- likelihood
- Bayesian inference
- outliers
- heterogeneity
- comparison of estimators
- add parameters

Estimating parameters

- Moment methods
  - Equate observed quantity with theoretical expectation
  - Derive variances (analytical or by simulation)

- Likelihood
  - Calculate relative probability of observing data given different values of parameters

- Bayesian inference
  - Use prior information about parameters to influence estimate

Data set

\[ n = 10 \]
\[ S = 14 \]
Summary statistics

- Good properties of summary statistics
  - Include most (all) information in the data
  - Different statistics should use different information
  - Expectations and variances should have simple relationship to model parameters

**Statistic** | **Symbol** | **Expectation**
--- | --- | ---
Average pairwise diversity | $\pi = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij}$ | $\theta$
Number of segregating sites | $S$ | $\theta \sum_{i=0}^{n-1} 1/i$
Number of haplotypes | $K$ | $\theta \sum_{i=0}^{n-1} 1/(i + \theta)$ (no recombination)

$\theta = 4N_c\mu$

The frequency spectrum of polymorphism under the coalescent

$$E[\xi_i] = \frac{\theta}{i}$$

Moment estimators of $\theta$

- **Pairwise diversity**
  $$E[\pi] = \theta$$
  $$\hat{\theta}_e = \frac{1}{n\theta(n-1)} \sum_{i<j} \pi_{ij}$$
  $$\text{Var}(\hat{\theta}_e) = b_1\theta + b_2\theta^2$$
  Tajima (1983)

- **Number of segregating sites**
  $$E[S] = a_1$$
  $$\hat{\theta}_w = S / a_1$$
  $$\text{Var}(\hat{\theta}_w) = \theta / a_1^2 + \theta^2 a_2 / a_1^3$$
  Watterson (1975)

Likelihood estimation of $\theta$

- Using the recursion of Tavaré (1984)
  - Number of segregating sites only
  $$\hat{\theta} = 5.2$$
  $$2U = 1.9 - 14.4$$

- Using the full-likelihood method of Griffiths and Tavaré (1994)
  - Implemented in GENETREE software
  $$\hat{\theta} = 4.7$$
  $$2U = 1.9 - 12.2$$
Stochastic simulation

- Generate random samples from Fisher-Wright population
  - Joint simulation of genealogies and mutations (Griffiths & Tavaré, 1994)
  - Time until next event = exponentially distributed with rate equal to the sum of all possibilities
    \[ \phi(\tau) = \lambda e^{-\lambda \tau} \]
    \[ \lambda = \frac{n(n-1)}{2} + n\theta \]
  - Probability next event is a mutation = probability of mutation divided by total probability of events
    \[ \Pr(\text{mutation}) = \frac{\theta}{\theta + n-1} \]

Active lineages

<table>
<thead>
<tr>
<th>Event</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coalescence</td>
<td>( \frac{n(n-1)}{2} )</td>
</tr>
<tr>
<td>Mutation</td>
<td>( \frac{n\theta}{2} )</td>
</tr>
</tbody>
</table>

Conditional stochastic simulation

- Simulate histories proportional to the probability of the history given the data and model parameters
  - Choose next event according to probability under coalescent, or an approximation to this probability
    - weight simulation likelihood by probability under proposal scheme
      \[ L(Data) = \frac{1}{\prod L(Guess_i)} \cdot P(Guess_t | \text{Coalescent}) \cdot \frac{P(Guess_t | \text{Proposal})}{P(Proposal)} \]

Active lineages

Estimating ages of events

- Estimating ages of events
  - MRCA
  - Mutations
  - Haplotypes
  - estimated by GENETREE
An example: β-globin

- 326 sequences of a 2.67 kb region (Harding et al. 1997)
- Nuclear loci expected to have greater coalescent depth than mitochondria or Y chromosome
- Assume
  - Infinite-sites model
  - No recombination within region (exclude known hot-spot)

![β-globin gene family on Human Chromosome 11](image)

20 segregating sites found in a sample of 349 sequences from a globally distributed sample of individuals

Bayesian estimation

- May have prior information on the effective population size, or rate of mutation that we wish to incorporate
- Calculate the probability of θ given the data

Prior probability for θ  Likelihood

![Bayesian estimation diagram](image)

Posterior probability for θ  Probability of data

\[ P(θ | D) = \frac{P(θ)P(D | θ)}{P(D)} \]

Bayesian v likelihood?

- Likelihood is intuitive (what is the probability of the data given these parameter values?), but making statements about confidence can be difficult
  - e.g. 2-unit support interval requires asymptotic conditions that are rarely met in population genetics
- Bayesian methods are also intuitive (what is the probability that my parameter has a certain value given these observations?), but does require statements about prior probability.
- **Strengths of Bayesian inference**
  - Natural representation of uncertainty
  - Can combine information from different sources
- **Difficulties of a Bayesian world**
  - Model comparison can be difficult
  - Major computational challenges in exploring parameter space efficiently for population genetic data
Example: Haplotype estimation

- Diploid genotype data consists of a series of unobserved haplotypes
  - How can we reconstruct haplotypes from genotype data?
    - Parental trios (but still any sites heterozygous in both parents unresolvable)
    - Start with full homozygotes and build up picture iteratively (Clark algorithm)
    - Maximise likelihood of data with EM algorithm
    - Estimate posterior probabilities of phasings with Bayesian approach
      (Stephens and Donnelly 2001)

- Advantages of Bayesian approach
  - Represents uncertainty of reconstructions
  - Can easily combine different sources of data

The PHASE algorithm

- Start with initial guess (e.g., EM solution)
- Sample novel reconstructions from the posterior distribution of phasings given a coalescent approximation
- Probability nth chromosome given rest of sample

- If many samples are drawn from the posterior, the degree of confidence in any single reconstruction can be assessed

Other approaches to estimation

- Where estimation of full likelihood is computationally intractable, approximations can be made which perform the estimation procedure well, but do not attempt to use the full information in the data
  - Composite likelihood: The full likelihoods of subsets of the data are calculated and combined as if they were independent
  - Summary-statistic likelihood: full likelihoods are estimated for one or a few summary statistics
  - Approximations to the full coalescent: Simpler probabilistic models that capture some key aspects of the coalescent are employed

Example: composite likelihood gene mapping

- Has been used for some time in pedigree-based studies to deal with families with >2 sibs
- Fine-scale mapping using the coalescent (Hudson)
  - To calculate the composite likelihood that the disease mutation is at a given site (with some degree of relative risk) calculate the coalescent log likelihood for sites 1, 2...n and add them together
  - Over-estimates the strength of signal due to non-independence
**Example: Summary-statistic likelihood and human expansion**

- Human population has expanded dramatically in last few thousands of years, but when did this expansion start?
- Full-likelihood analysis computationally difficult, look at summaries of the data
- Pritchard et al. (1999): 445 individuals at 8 Y chromosome microsatellites. Data summarised as
  - Mean variance in repeat number
  - Mean effective heterozygosity
  - Number of haplotypes
- Use rejection scheme to fit model
  - Strong evidence for growth with MRCA estimated to be in last 120,000 years.

**Example: Urn-model representation for recent demographic events**

- For recent demographic events, subsequent mutation will have had little influence on patterns of diversity
- If the mutation process is ignored, differentiation between populations is simply due to coalescent events
  \[ Pr(a | t, n) = \sum_{n=0}^{t} e^{-t/(2n+1)} \left( 1 - \left( 2(1-a)^{n + 1} \right) \right) \]

**Tree-based approaches to population genetic estimation**

- All inference considered so far considers infinite-sites models of sequence evolution
  - All polymorphic sites are the result of a single mutation
- BUT for many organisms (viruses, bacteria, mitochondria) this assumption is very unlikely
- If such genomes do not recombine, a combination of traditional phylogenetic and coalescent analyses can be used
  - Felsenstein group (Kuhner et al. 1998)
  - Drummond et al. (2002)
  - Pybus et al. (2001)
- The coalescent can be thought of as a Bayesian prior for the shape of genealogies

**Example: Non-contemporaneous viral samples**

- Coalescent theory can readily be adapted to consider samples taken at different time points
  - Time-stratified samples allow us to estimate the effective population size and the mutation rate independently.
  - The very high mutation rate in some viruses, such as HIV1 and HCV allow good resolution of the phylogenetic structure
  - Care should be taken to exclude recombination in population samples
  - Drummond et al. (2002). Posterior for the mutation rate in HIV1 env sequences from a single patient with 5 sampling dates over 3 years.