Third Pavia International Summer School for Indo-European Linguistics, 7-12 September 2015

HISTORICAL LINGUISTICS AND MOLECULAR ANTHROPOLOGY

Brigitte Pakendorf, Dynamique du Langage, CNRS & Université Lyon 2

brigitte.pakendorf@cnrs.fr

LECTURE 2 : ANALYSIS OF GENETIC DATA

Factors that can influence genetic variation:

- genetic drift
- bottlenecks/founder events
- substructured ancestral population

Genetic diversity within populations:

- helps assess population size and population size changes:
  - smaller populations have lower diversity
  - population size reductions reduce diversity
- important for detecting effect of drift
- helps trace the origin of a mutation (haplogroup): area with highest diversity

Gene/haplotype diversity:

- Probability that two haplotypes picked at random from a sample are NOT identical
- can range from 0 to 1:
  - 0: everyone has same haplotype
  - 1: everyone has a different haplotype
- does not take into account difference in sequence (2 haplotypes differing by 1 mutation and 2 haplotypes differing by 10 mutations are equally not-identical)

Measures that take molecular differences into account:

- **Mean number of pairwise differences (MPD):** number of differences between every pair of haplotypes averaged over entire sample
- **Nucleotide diversity (π):** number of differences between every pair of haplotypes averaged over entire sample and divided by haplotype length

Genetic differences between populations:

- simplest measure: **p-distances**
  - average number of nucleotide differences between all pairs of sequences from two populations relative to length of sequence
- various more refined measures have been devised that take into account specific processes of DNA sequence mutation
- general measure: **F_{ST}**
- proportion of total genetic variance that is due to differences between populations
- applicable to any kind of marker:
  - allele frequency data = F_{ST}
  - sequence data = Φ_{ST}
  - STR data = R_{ST}
- can range from 0 to 1
  - 0 = no genetic difference between populations
  - 1 = maximum genetic difference

Displaying genetic differences between populations:

- Multidimensional Scaling analysis (MDS)
- Neighbour-joining (NJ) trees
- Correspondence Analysis (CA)
- Principle Components Analysis (PCA)

Exploratory tools to identify factors structuring genetic variation:

- **AMOVA**: Analysis of Molecular Variance
- analysis of genetic variance at different hierarchical levels:
  - within populations
  - between populations within a group (e.g. geographical region, language family)
  - between groups (e.g. geographical region, language family)
- **Mantel test:**
- assesses the significance of correlation between distance matrices, e.g. between matrix of genetic distances and matrix of geographic distances
- permutation test:
  - one matrix is kept constant, the values [illegible] ıbled, and correlation coefficient is calculated
  - repeated 1000 times or more
  - if observed correlation coefficient is lɛ [illegible] ɔserved correlation is significant

based on genetic distances

based on haplogroup/allele frequencies

Trees based on character-state information:

- avoids loss of information inherent in distance methods
- different methods available, e.g.:
    - Maximum Parsimony (MP)
    - Maximum Likelihood (ML)
    - Bayesian MCMC analysis
- often two different algorithms are used – if both reconstruct the same tree one can be confident of the phylogeny

- **Maximum Parsimony**: finds the tree that minimizes the number of mutations needed to account for observed data
- **Maximum Likelihood**: finds the tree that maximizes the likelihood of observing a set of sequences given a specific mutation model
- **Bayesian MCMC**: combines Markov Chain Monte Carlo (MCMC) method with Bayesian approach

MCMC = Markov Chain Monte Carlo

- a method that searches space of possible trees for region of highest likelihood
- starts at random at one point (or at a likely tree chosen by user), compares likelihood of current tree to previous one – if current tree has higher likelihood, then it will continue searching from this one
- samples a tree every 1000 steps and keeps them → set of posterior trees of which the most probable one is generally depicted

Bayesian analysis

- lets you include prior knowledge into your calculations → prior probabilities or "priors"
- to this adds the information from the observed data → posterior probabilities or "posteriors"

Time estimates based on trees:

- if mutation rate is known, then TMRCA of branches can be dated with any character-based approach
- but dating is implemented already in widely-used programme based on Bayesian MCMC analysis (BEAST = Bayesian Evolutionary Analysis Sampling Trees)

Tree algorithms

- assume ancestor is not included in the data → all individuals treated as tips
- cannot visualize conflicting evolutionary pathways in one tree

Solution (for non-recombining data): networks

- "reconstructs ⋯ all maximum parsimony or MP trees)" (Network user guide 2012; http://www.fluxus-engineering.com/Network4611_user_guide.pdf)
- Median Joining (MJ): for multistate data (e.g. STRs)
- Reduced Median (RM): for binary data (e.g. sequences)
- allows one to weight sites differently → give more weight to sites that evolve slowly, less weight to sites that have high mutation rate

Exploratory unsupervised analyses:

- possible with dense genome-wide data → SNP chip data, whole genome sequences
- avoid a priori population definitions: unit of analysis is individual, colour coding done at later stage → can show up substructure in populations, variation, etc
- Principal Components Analysis
- STRUCTURE-type Analysis

STRUCTURE-type analysis:

- algorithm assumes individuals come from unknown number of groups
- each group is assumed to have different allele frequencies at genotyped loci → groups are assumed to be internally homogenous and maximally different from each other
- algorithm estimates allele frequencies for each ancestral group and also estimates proportion of ancestry that one or more groups contributed to each individual in analysis
- user defines number of possible ancestral groups, generally starts off with 2 (K=2) and sequentially defines more
- algorithm has "goodness of fit" measure to choose best number of groups

Important note: PCA and STRUCTURE-type analyses are **exploratory**

- ancestry components detected by STRUCTURE-type analyses CAN be the result of admixture, but need not be
- likewise, patterns found in PCA can be the result of long-distance migration and divergence, or of long-settled populations with gene flow on local geographical scale (isolation-by-distance)
- use formal tests of admixture and isolation by distance to distinguish

Dating divergence times (≈Time to the Most Common Recent Ancestor, TMRCA):

- simplest approach:
  - count number of mutational differences between species/populations/haplogroups
  - use mutation rate to estimate time of divergence
- requires:
  - knowledge of
    - number of mutational differences between species/populations
    - mutation rate
  - assumption
    - constant rate of mutation

- Estimating number of mutational differences between populations/haplogroups is relatively straightforward, more complicated between species.
- Estimating mutation rate is much more difficult, debate over which method is best:
  - **phylogenetic method**: use external calibration point, e.g. fossil evidence for colonization of an island or continent
  - **pedigree method**: count number of mutations that have occurred in deep-rooting pedigrees
  - pedigree rate often higher than phylogenetic rate → counts all mutations, not just fixed ones
- Rate of mutation is not always constant → use relaxed clock model in BEAST – but enormous confidence intervals

Further point to keep in mind when assessing molecular dates:

- mutation is a stochastic process → not regular like radioactive decay
- we sample a subset of individuals from a population to count our mutations → if we sampled a different subset, we would find a slightly different amount of mutations
- molecular dates will always have much larger confidence intervals than archaeological ones

Detecting (and dating) admixture events:

- detection can be relatively straightforward with characteristic uniparental haplogroups, if populations are genetically distinct
- but no good way to date the event
- with autosomal DNA, it is possible to date admixture events using patterns of linkage disequilibrium (LD)
- the longer the time since the admixture event, the shorter the LD fragments
- number and width of chromosomal blocks of different ancestry are used to estimate time since admixture
- works well with genetically distinct populations, is more problematic with related populations

Problems with analyses of admixture using autosomal data:

- need to identify parental populations, or reasonable proxies
- programmes that estimate admixture don't tell you whether your proxies are correct
- "If you want to model African Pygmies as the result of admixture between New Guinean Highlanders and Greenland Eskimos, these programs will do their best to give you an answer, no matter how nonsensical." (Stoneking to appear: ch. 12)

Population size:

- Patterns of Linkage disequilibrium (static)
- Bayesian Skyline Plots (BSPs; dynamic)

Patterns of Linkage Disequilibrium

- in large populations there are many reproductive events => many recombination events

- in small populations, there are less reproductive events => fewer recombination events
- recombination breaks down fragments in LD
- ➢ in smaller populations, LD fragments will be longer than in large populations

Bayesian Skyline Plots

- if you have a phylogenetic tree that is consistent with a molecular clock, you can estimate N_e at different times in the past from the number of branching events
- number of branching events is proportional to N_e: more branching events = bigger N_e
- from tree you can produce a plot of population size change over time with confidence intervals