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 1 : INTRODUCTION TO MOLECULAR ANTHROPOLOGY

Literature :

Textbooks:

Stoneking, Mark (to appear 2016??) : *Introduction to Molecular Anthropology*. Wiley-Blackwell.

Jobling, Mark, Edward Hollox, Matthew Hurles, Toomas Kivisild, Chris Tyler-Smith (2013): *Human Evolutionary Genetics*. Garland Science.

Review articles:

Pakendorf, Brigitte & Mark Stoneking (2005): Mitochondrial DNA and Human Evolution. *Annual Review of Genomics and Human Genetics* 6: 165-183.

Kivisild, Toomas (2015): Maternal ancestry and population history from whole mitochondrial genomes. *Investigative Genetics* 6: 3. doi: 10.1186/s13323-015-0022-2

Jobling, Mark & Chris Tyler-Smith (2003): The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics* 4 (8): 598-612.

Stoneking, Mark & Johannes Krause (2011): Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics* 12 (9): 603-614.

Pugach, Irina & Mark Stoneking (2015): Genome-wide insights into the genetic history of human populations. *Investigative Genetics* 6: 6. doi: 10.1186/s13323-015-0024-0

Written for linguists:

Pakendorf, Brigitte (2007): Appendix 1 in *Contact in the Prehistory of the Sakha (Yakuts): Linguistic and Genetic Perspectives*. LOT Dissertation series 170. Utrecht: LOT.

---- (2014): Historical linguistics and molecular anthropology. In Bowern, Claire and Bethwyn Evans (eds): *The Routledge Handbook of Historical Linguistics*, Oxon, New York: Routledge: 627-641.

Some basic terminology :

- DNA = sequence of nucleotides: A(denine), T(hymine), C(ytosine), G(uanine)
- nucleotide ≈ base pair (bp)
- mutation: random change of DNA sequence, e.g. change of nucleotide at certain position
- polymorphism: variation between individuals in DNA sequence
- SNP: single nucleotide polymorphism
- locus: physical location of genes on a chromosome
- allele: particular variant at a polymorphic locus
- linkage: association of genes located close together on a chromosome
- gene pool: collection of alleles present in a population
- gene flow ≈ admixture

Types of DNA:

mitochondrial DNA (mtDNA):

- completely separate genome
- hundreds to thousands of practically identical copies per cell
- ~16,570 bp
- complete human mtDNA sequenced 1981 → Cambridge Reference Sequence (CRS); revised Cambridge Reference Sequence 1999 (rCRS)
- inherited solely from mother
- no recombination

nuclear DNA:

- 23 pairs of homologous chromosomes, i.e. 46 chromosomes per cell
- ~3 billion bp
- inherited from both parents
- Y-chromosome: inherited only in the paternal line, from fathers to sons
- recombination (exception: large parts of Y-chromosome = Non-recombining portion of the Y-chromosome, NRPY)

haploid vs. diploid DNA:

- haploid DNA = only one copy of each gene
  - mtDNA (hundreds of identical copies)
  - Y-chromosome
  - X-chromosome in men
- diploid DNA = two copies of each gene
  - autosomes
  - X-chromosome in women

Summary – types of DNA:

Human genome	nuclear DNA
Nuclear DNA	autosomes + sex chromosomes
Autosomes	chromosomes 1-22
Sex chromosomes	X + Y chromosome
Uniparental markers	Y-chromosome + mtDNA
coding DNA	determines amino acid sequence of a protein
non-coding DNA	regulatory function, or no function

Important concepts for understanding the characteristics of the different types of DNA:

Recombination: the exchange of parts of homologous chromosomes during meiosis

Linkage disequlibrium (LD):

- non-random association between genotypes at linked SNPs" (Stoneking to appear, ch.9)
- i.e. alleles at particular loci that occur together on the same chromosome
- will be broken down over time by recombination
- size of fragments in LD can be used to estimate population size or time of admixture

Effective population size (N_e):

- number of individuals that have the potential of contributing to next generation
- ≠ census size → children are not yet participating in reproduction and might die before they reach maturity, old people are beyond reproducing, not all sexually mature individuals have children
- is also applied to molecules, differs between autosomes, X-chromosomes, and uniparental markers (mtDNA/Y-chromosome)

Advantages of uniparental markers:

- **mtDNA**: transmitted only in the maternal line
- **Y-chromosome**: transmitted only in the paternal line
  → highlight sex-specific interactions
- do not undergo recombination
  → we can reconstruct phylogenetic trees of mutations → haplogroups

Disadvantages of uniparental markers:

- limited to a tiny fraction of an individual's ancestors
  → restricted view of population history
- low effective population size
  → very susceptible to chance fluctuations (=genetic drift)

Advantages of autosomal markers:

- huge amounts of genetic information → can nowadays be mined with novel computational methods
- higher effective population size → less susceptible to genetic drift
- unbiased view of population prehistory → all the ancestors are taken into account
- linkage disequilibrium can be used to infer population history

Disadvantages of autosomal markers:

- no possibility to disentangle sex-biased gene flow

Sample collection:

- blood samples
- cheek swab samples
- saliva samples

Polymorphisms studied:

Classical markers = blood groups, blood proteins

disadvantages:

- potentially underlie selection, because gene products have important function
- variation very crude, even similar-looking phenotypes may have underlying genotypic differences

DNA-based markers

1. Mitochondrial DNA (mtDNA; >1990s); Cann/Stoneking/Wilson 1987 *Nature* 325(6099): 31-36; "mitochondrial Eve"
2. Y-chromosome (> 2000s); Underhill et al. 2000 *Nature Genetics* 26(3): 358-361
3. Autosomal markers (> 2005); Rosenberg et al. 2002 *Science* 298(5602): 2381-2385
4. Genome sequences (> 2010); The 1000 Genomes Project Consortium 2010, *Nature* 467(7319): 1061-1073

Main types of polymorphisms:

- Single nucleotide polymorphisms (SNPs)
- Length polymorphisms
  - Insertions and deletions (indels), e.g. "9bp-deletion" in mtDNA
  - Repetitive DNA, e.g. Short tandem repeats (STRs)

Single nucleotide polymorphisms:

- variation among individuals at particular position in DNA sequence (= "locus")
- alleles = nucleotide variants found at this site
- "singletons": allele found in only 1 individual in a sample
- generally rare mutation events, especially in nuclear DNA

Short tandem repeats (STRs):

- also known as microsatellites
- repeated elements between 2-6 bp long
- very high mutation rate
- allow for more detailed analyses of genetic variation on the background of particular SNPs
- very frequently determined on Y-chromosome (Y-STRs) → commercial genotyping kits available to type 12, 17 or 23 Y-STR loci at once
- but can also be typed on autosomes

Genotyping:

= determining character state of particular polymorphism in an individual

- SNPs:

  - determined by either sequencing stretch of DNA → unbiased detection of all polymorphic sites, but "inefficient", as many invariant sites included
  - or genotyping target polymorphisms → "efficient", only polymorphic sites included; autosomes = "SNP chips"; but ascertainment bias
- STRs:
  - always targeted

Haplotypes and haplogroups:

- haplotype = combination of character states at particular polymorphic sites (e.g. sequence of nucleotides or sequence of repeat numbers of STRs) → possible to establish for uniparental or autosomal DNA (if known which SNPs are from the same chromosome)
- haplogroup = group of haplotypes defined by shared mutation (≈ subgroups in historical linguistics that are defined by shared innovations) → only for uniparental DNA
- distinction between macro-haplogroup, haplogroup, and sub-haplogroup not neat → depends on SNPs typed or researchers' choice (sequencing)

Haplogroup labels:

- can be very cumbersome:
  - e.g. mtDNA B4a1a1a, widespread in Oceania
  - e.g. Y-chromosome E1b1a7a, common in sub-Saharan Africa (Bantu expansion)
- alternative labels:
  - haplogroup with defining mutation, e.g. E-U174 instead of E1b1a7a
  - not always possible if sequences rather than targeted SNPs
  - don't show phylogenetic relationship