Modern Human Phylogenetics

Week 1.
--The Four (4) Causes of Evolution
--Reading and building trees

Week 2.
--Mutation rates and dating trees
--Phylogeography

(Aaack, missing pop gen metrics...)

Week 3.
--mtDNA and y-chromosome trees

Weeks 4
--The Coalescent!

Weeks 5...more to come

Human nuclear genome - about 3 billion nucleotides, with about 3 million of them variable among any two random humans (99.9% identity); most variants probably have no phenotypic effects (are 'neutral')

Some jargon:
Marker - generic name for bit of DNA used to infer something...
SNP - single nucleotide polymorphism (2 or more bases at a site)
Allele - one of a number of variants of a marker
Haplotype - linear combination of SNPs or other markers on a chromosome such as C...C....A.T (haplotype 1), C...G....A.T (haplotype 2); sets of linked bases tend to be inherited together -- form flanked 'blocks'
Copy number variation - variation in number of copies of large sections of genome, including one or more genes (large deletions, duplications)

Must consider uni-parentally-inherited vs. bi-parentally-inherited genes (aka "markers")

Critically different dynamics

The four causes of organic evolution:

1. **MUTATION**: fuel of evolution, random
2. **SELECTION**: only known cause of design
3. **DRIFT**: powerful retarding force, random
4. **MIGRATION**: quick way to get more fuel

The data used to infer MHP (variation in the uni-parental and biparental genetic markers among individuals and populations) is created by the interaction of these four processes, and by nothing else.
The four causes of organic evolution:

1. **MUTATION**: fuel of evolution, random
2. **SELECTION**: only known cause of design
3. **DRIFT**: powerful retarding force, random
4. **MIGRATION**: quick way to get more fuel

Fairly important facts about mutation relevant to MHP:

1. There are different types of mutation (which we will get to).
2. Mutations have deleterious effects in the great majority of cases, so selection should minimize the mutation rate, subject to constraints (repair ability, time constraints in replication)
3. The larger the population, the greater the scope and potential for mutations that turn out to be adaptive. Drift (which we will also get to) is also weaker/slower in larger populations, and so selection is more efficient.

The theory around using genetic data to infer MH demographic and phylogenetics patterns seems based primarily on the drift dynamics in small populations.

What is mutation?

"A sudden heritable change in the genetic material, most often an alteration of single gene by duplication, replacement, or deletion of a number of base pairs"

Or, a duplication, replacement or deletion of the entire gene

Or, the …. of (part of) the entire genome!

What is mutation?

Any change in the nucleotide sequence of DNA

Mutations at level of the gene

Point mutations form new alleles

a single (or few) base pair substitutions

Can be caused by either

- External factors that damage DNA, polymerases make mistake when repairing.
- Errors during DNA replication, again by polymerase enzymes messing up.

Can be synonymous or nonsynonymous
Synonymous mutations were considered neutral...subject to no selection

However, new evidence suggests otherwise - different species use different codons preferentially: so, different tRNAs are used – selection?

Synonymous mutations away from common triplet (and so common tRNA) may be subject to selection (and be deleterious)

Mutations that create new genes: gene duplication

Cause:
Unequal crossover of chromosomes during meiosis

Consequences:
A. Produces an additional copy of the 'parent' locus; may become nonfunctional (pseudogene), OR may gain a new function through mutation and selection
B. Gene duplication now recognized as a major evolutionary force (last decade only)
40,000 human genes, of which 15,000 produced by 'recent' duplications

Gene duplication in the douc langur (Pygathrix nemaeus)

Langurs eat leaves
Bacteria ferment leaves
Langurs digest bacteria
Most primates have one copy of RNASE1, which digests dietary RNA and degrades double stranded RNA (viruses)

Douc Langur has two RNASE genes
RNASE1 + RNASE 1B
pH7.4  pH6.3
defense  digestion
New gene (note, not new allele) is more efficient at digesting food, can no longer cut viral RNA

* see 'cow'

Also, the number of copies of a gene may affect your phenotype!
100’s of multiple copy-number regions found in human genome
-- very many regions of high copy number variation found (CNV)
-- contain many genes that seem to be evolving rapidly
-- CNVs may be an important form of mutation

Natural selection of gene copy number in human evolution

Diet and the evolution of human amylase gene copy number variation

Nature Genetics, 2007
Mutations above gene level
Inversions caused by faulty repair, flips section of chromosome

Structural chromosomal variation between humans and chimps

‘Geometric’ mutations
Consequence:
Entire segment is inherited as a unit, is "linked" selection acts on whole set as one (because linked) - can retard efficacy of Natural Selection

Uniparentally-inherited markers are generally non-recombining segments (mtDNA or NR sections of y-chromosome)

Mutations at level of the genome
Polyploidy - chromosomal doubling
Production of 2n gametes at meiosis
Self fertilization leads to tetraploid (4n) "F1"

Consequence: creates new species through immediate reproductive isolation (because backcrosses are triploid and infertile)

Not relevant for MHP, far as I know.

Mutation Rates
There are at least four different quoted rates
1. per base position per replication or per generation
2. per locus per generation
3. per chromosome per generation
4. new phenotypic variation per generation

Mutation rate per base pair
-Grow C. elegans asexually for many known generations in the most benign environment possible (little selection)
-Freeze individuals at known time (gen 0, 280, 353, 396, etc)
-Sequence ‘a lot’ of DNA from everyone and count mutations.

\[ \text{mutation rate} = 2 \times 10^{-8} \text{ mutations/site/generation} \]

\( \approx \text{2 new mutations in every new worm in the world (a lot)} \)
(Q. how big is the C. elegans genome?)
recent estimate:
Human per base-pair per generation nuclear mutation rate = 
$12.85 \times 10^{-9}$ per site per generation (s.e. 1.95)

This is considered a poisson process: mean=variance, so a very noisy process (important for dating, which comes later)

(‘indels’ and gene copy number rates likely higher and more important as fuel for natural selection.)

Lynch, PNAS 2010

Mutation rate for different markers varies greatly, and is a topic for entire afternoon

Mutations: what are their effects?

One thing mutation does (on its own) is erode the frequency of any particular allele....but very, very slowly and not relevant to MHP

Random with respect to current ‘need’ (changes that would increase relative fitness in current environment)
Not random at nucleotide level (e.g. transitions > transversions)
Most mutations are bad or neutral in effect (M. Kimura)

The four causes of organic evolution:

1. **MUTATION**: fuel of evolution, random
2. **SELECTION**: only known cause of design
3. **DRIFT**: powerful retarding force, random
4. **MIGRATION**: quick way to get more fuel

Migration (very straightforward)

What is it?
Movement of alleles between populations

What causes it?
Movement of individuals followed by mating, or movement of gametes followed by fertilization
Migration

Migration: Island model

\[ p_{r1} = mp_{m} + (1-m)p_{r0} \]

- proportion migrants \(*\) proportion of migrants that carry focal allele
- frequency of focal allele on island after 1 generation of immigration

Effect of MIGRATION (gene flow):

- HOMOGENIZATION of allele, genotype frequencies, across populations (makes them more similar to one another)

What are effects of migration?

- equilibrium is reached extremely rapidly, and the population ends up with the source allele frequency!

Case study: song sparrows on Mandarte Island BC

- bottleneck: population crashed
- \( \approx 1 \) immigrant per year

Keller et al 2001

Proc Roy Soc 268: 1387-1394

The four causes of organic evolution:

1. MUTATION: fuel of evolution, random
2. SELECTION: only known cause of design
3. DRIFT: powerful retarding force, random
4. MIGRATION: quick way to get more fuel

The rule of thumb: \( \approx 1 \) migrant per generation homogenizes populations.
Now, let $p =$ proportion of one allele (A) at a locus
$q=1-p =$ proportion of another allele (a)

If phenotype of Aa = phenotype of AA, we say A is “dominant” and a is “recessive”,

if phenotype of Aa is intermediate, we say the alleles are “codominant”

Hardy Weinberg

How fast will selection work?

~2% advantage for B1 allele (~300 generations)

Q. is the B1 allele dominant, recessive or codominant?

How fast is Natural Selection?

An allele with a 2% selective advantage will sweep through a large population (from a very low initial frequency) in less than 300 generations!

2% selective advantage for an allele:

So, if 50/100 individuals carried the new allele (as heterozygotes), and if TWO of these had 3 instead of 2 babies, this would be more than a 2% selective advantage for the allele!

Time to fixation for beneficial mutations

- Dominant mutations increase in frequency rapidly (are seen), but approach fixation slowly (selection against rare recessive is slow - they’re hidden)

- new recessive mutations increases very slowly in frequency (rarely exposed in homozygous form); once common, recessive alleles go to fixation quickly (selection against deleterious dominant is fast)

Inbreeding (avoidance)
Nonrandom mating can affect how often alleles are exposed to selection

- **What is it?**
  - **Assortative mating:** individuals choose mates phenotypically more similar than expected by chance
    - Can cause inbreeding: mating between relatives, more frequent than by chance
    - common for phenotypes in humans
  - **Disassortative mating:** individuals choose mates phenotypically less similar than expected by chance
    - Can cause outbreeding
    - rare in mammals

What are effects of inbreeding?

1. Increased homozygosity in population.
2. No change in allele frequencies per se.

- If you could self-fertilize, a greater proportion of your offspring are homozygous than the proportion from individuals that outcross
- Depending on the amount of genetic variation, this produces homozygous recessives at many loci in your offspring
- If a significant proportion of mutations are deleterious, these will be expressed...

A greater proportion of your offspring will have lower fitness than offspring from outcrosses -> **Inbreeding Depression**

This is why one just doesn't marry close relatives in almost every society studied

![Fig. 6.28](image)

That was straight selection for (or against) particular alleles: **erodes** genetic variation within a population

48
Selection can also maintain variation

Heterozygote advantage has been suggested for immune genes, the Human Leukocyte antigen (HLA) superlocus in humans.

These are genes that produce proteins that present antigens from pathogens to the immune system.

S. Amerindians - too few homozygotes at HLA loci

Table 3. The observed and expected numbers of homozygotes for each of the alleles at HLA-A and HLA-B and the ratio of observed to expected (O/E) for each allele.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Allele</th>
<th>Observed</th>
<th>Expected</th>
<th>O/E</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-A</td>
<td>02</td>
<td>41</td>
<td>53.00</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>32</td>
<td>44.75</td>
<td>0.723</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>34</td>
<td>46.00</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td>68</td>
<td>7</td>
<td>12.25</td>
<td>0.571</td>
</tr>
<tr>
<td>HLA-B</td>
<td>15</td>
<td>25</td>
<td>32.00</td>
<td>0.781</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>46</td>
<td>61.75</td>
<td>0.745</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td>21</td>
<td>24.50</td>
<td>0.857</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>11</td>
<td>17.00</td>
<td>0.647</td>
</tr>
<tr>
<td></td>
<td>51</td>
<td>7</td>
<td>9.25</td>
<td>0.737</td>
</tr>
</tbody>
</table>

How selection can maintain variation directly

Frequency dependent selection

- Selection coefficient is a function of the frequency of the phenotype in the population

Here, \( s \) on a genotype is not constant

cannot write "\( s_{AA} = k \)"

but rather a function of its frequency!
Negative frequency dependence
*Reward-less* orchids

![Graph showing negative frequency dependence](attachment:image)

Explanation: Bees learn, and so yellow individuals are visited less when common

Gigord et al PNAS 98:6253-6255

How selection can maintain variation directly

**Sex ratio in mammals (ie humans) (Fisher, 1930):**

Given every offspring has a mother and a father, it turns out that sex ratio is a frequency dependent character - there is selection to produce the rarer sex.

The four causes of organic evolution:

1. **MUTATION:** fuel of evolution, random
2. **SELECTION:** only known cause of design
3. **DRIFT:** powerful retarding force, random
4. **MIGRATION:** quick way to get more fuel

Genetic Drift

Variation in allele frequency due to chance ("sampling error") in finite populations, especially very small populations (but always there)

Example: any hand from a well-shuffled deck will never have equal numbers of each suite, and variance increases with smaller hands...

Causes of Genetic Drift

1. **Continuous Drift**
2. **Founder Effect**
3. **Population bottlenecks**

Continuous population fluctuation

Human cycles in Egypt
Huntington’s Disease & founder effect

Autosomal dominant genetic disease that damages nerve cells and leads to death. Onset is in mid 30’s

mutation is lengthening of CAG triplet in gene Huntingtin.

http://aklager.wordpress.com/

US: 4-8 per 100,000
Europe: 1.6 to 10.0 per 100,000
Tasmania: 17.4 per 100,000
Mauritius: 46 per 100,000

However, Lake Maracaibo region of Venezuela: 700 per 100,000, and in some villages 50% population (20 000 total have disease allele)

many villagers trace their ancestry back 200 years to a single woman (Maria Concepción Soto) with Huntington’s disease who had 10 kids

Very old pattern attributed to founder events
O-bloodtype increases in frequency as one moves south.

Will be very important as we look at MHP data

What are the effects of Drift?
• Random change in allele frequency
• Gradual loss* of genetic variation
• Differentiation among populations

* if not counteracted by selection, this loss is random

Drift in different populations can lead to fixation of different alleles (or at least very different p’s and q’s)
looks like evolution by NS, but is just evolution
As alleles drift, heterozygosity goes down as one allele begins to dominate

at any one locus, heterozygosity is maximized at p=0.5, and drift always eventually takes one away from this maximum

\[ p \] (loss of variation...)

Loss of heterozygosity due to drift

\[ \frac{1}{2N} \] per generation, where N is the population size

So if there are 50 individuals, lose 1/100 or 1% of heterozygosity per generation

\[ Het_{t+1} = Het_t (1-(2N)^{-1}) \]

S. Wright

‘Het’ is just 2pq, so one can use this to calculate change in p (and so q): something to try at home

Effects of drift: Buri’s flies (1950’s)

107 bottles of Drosophila, 8+8 individuals

Many populations fixed for different alleles

Figs 7.16, 7.17
Wright's equation is based on an "ideal" population (random mating, same expected number of offspring/adult).

Flies (indeed, most species) show:
1. real variance in male mating success
2. real variance in female fecundity

--so "census size" $(N=16)$ not equivalent to "effective population size" $(N_e \sim 9)$

How do we calculate $N_e$?

- Number of adults
- Sex ratio of breeders
- Fluctuations over time

$N_e = \frac{4 N_m N_f}{(N_m + N_f)}$

E.g. even if everyone is a successful breeder:
1 bull + 100 cows = $(4*1*100)/101 \sim 4$
2 bulls + 2 cows = $(4*2*2)/4 = 4$

Things that depress $N_e$ - make population genetically smaller than it "looks"

Variation in-
- population size
- sex ratio
- family size

$N_e$ is harmonic mean across generations $t$
$N_e$ goes down with fewer mating males ($m$) than females ($f$) or vice versa
$N_e$ is variation in family size.
Ideal $= 2$, i.e. under Poisson

And you have to integrate over all three sources of variation

What is ratio of $N_e/N$ in wild populations?

Old rule of thumb says $N_e/N \sim 0.33$

However, survey of 102 species: mean $N_e/N \sim 0.1$
(range $10^{-6}$ for pacific oysters to 0.994 for current humans!)

from multiple regression, following factors influence $N_e/N$:
- var(population size across generations)  Biggest factor by far
- var(family size)
- taxonomic group
- sex ratio
In small populations:

Drift *counteracts* selection - ‘negative s’ alleles retained by chance, ‘positive s’ alleles don’t increase.

Inbreeding also means deleterious recessives are expressed, which may mean lots of individuals with low fitness, which further decreases population size! (but this will be counter-acted by selection, of course)

Drift and selection continually ‘compete’, and drift is more likely to ‘win’ in smaller populations