Some basic (and not so basic) statistics used in population genomics

OEB 230 Week 4

1

Absolute measures of divergence

Average number of differences between two sequences per site within a population, "heterozygosity" π_W

e.g. in humans $\pi \approx 0.1\%$
in *Drosophila melanogaster* $\pi \approx 3\text{-}4\%$

D_{XY} is the equivalent *absolute* measure of divergence:

$$D_{XY} = \pi_b$$

Average no. diffs. per site for two sequences chosen between two populations

e.g. humans – Neanderthals, $D_{XY} \approx 0.15\%$
humans – chimps, $D_{XY} \approx 1.2\%$

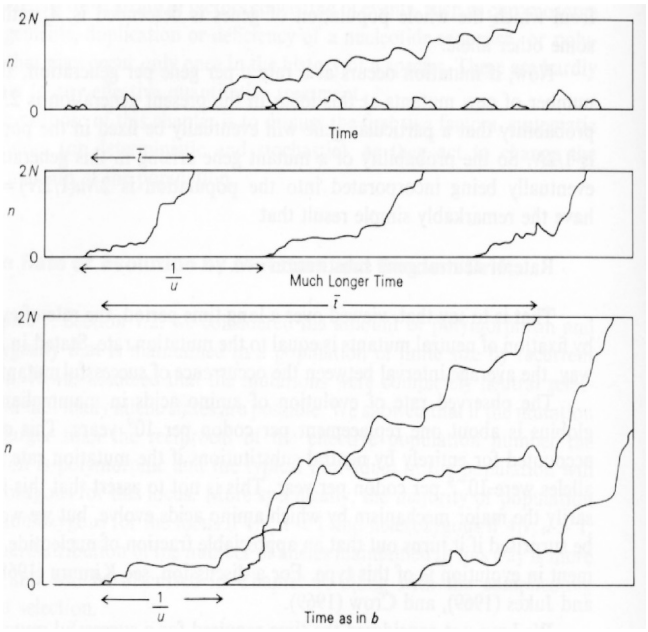
The average human is more closely related to the average chimpanzee than two wild *Drosophila melanogaster* are to each other! How could this be?

2

Kimura's 1968
Neutral Theory

"Evolutionary rate at the molecular level"

"Calculating the rate of evolution in terms of nucleotide substitutions seems to give a value so high that many of the mutations must be neutral ones"



3

$2N_e$ alleles

time $t-1$

Wright-Fisher model: every allele has an equal chance of reproducing

coalescence in a single generation

$2N_e$ alleles

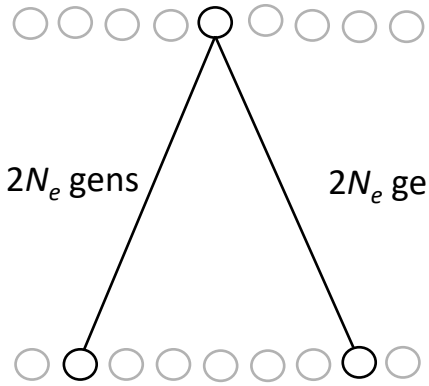
time t

Probability of two randomly selected alleles coalescing in the previous generation = $1/2N_e$

$1/2N_e$ is the rate per generation. It will therefore take on average ~ $1/(1/2N_e)$ gens. for a pair of alleles to coalesce, = $2N_e$ gens!

4

Expectation of polymorphism under neutrality



Tracing the track of two randomly selected alleles; how many mutations per site? $\theta = (2N_e + 2N_e)\mu = 4N_e\,\mu$

Assuming complete neutrality, we expect average level of polymorphism, $\pi \approx \theta = 4N_e\,\mu$

5

We're usually interested in a sample…

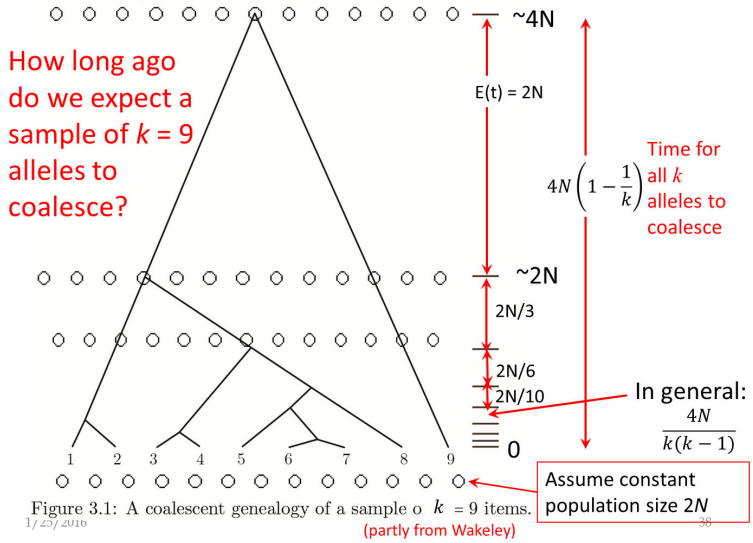


Figure 3.1: A coalescent genealogy of a sample o k = 9 items.

1/25/2016 38

(partly from Wakeley)

The coalescent

6

Sequence-based definition of F_{ST}

F_{ST} is the "fixation index", a *relative* measure of divergence, i.e. relative to diversity within sites:

$$F_{ST} = 1 - \frac{\bar{\pi}_w}{\pi_b}$$

Average no. diffs. per site within each population

Hudson & Slatkin, 1992

Average no. diffs. per site for two sequences chosen between two populations

D_{XY} is the equivalent *absolute* measure of divergence:

$$D_{XY} = \pi_b$$

Average no. diffs. per site for two sequences chosen between two populations

7

Gene flow/drift balance

$$F_{ST} \approx \frac{1}{1+4N_e m} \sim \frac{1}{2N_e} \Big/ m$$

Genetic divergence at locus *i*

Rate of fixation of drift per gen.

Rate of gene flow per gen.

Sewall Wright, 1931

8

Gene flow/selection balance

Genetic divergence at locus *i*

$$F_{ST} \sim \frac{s_i}{m}$$

Strength of divergent selection on locus *i* ~ postzygotic isolation per gen.

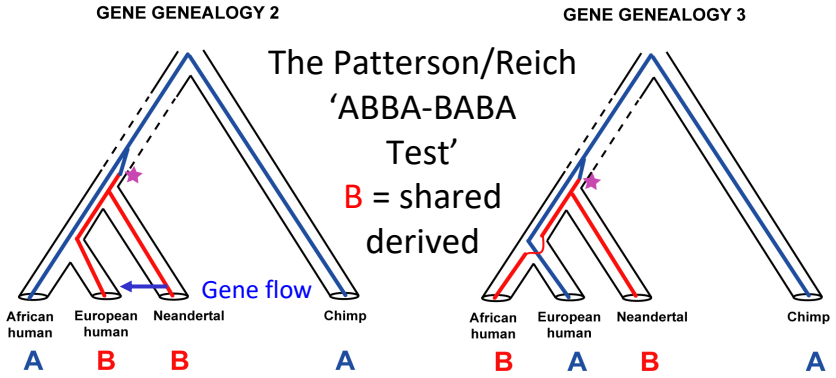
Rate of genome-wide gene flow, ~ disassortative mating per gen.

~ reproductive isolation

J.B.S. Haldane, 1930

9

Genomic test for DNA transfer between species



If ancestral polymorphism, EXPECT:
50% ABBA nucleotide sites
50% BABA sites

OBSERVE:
103612 ABBA sites
94029 BABA sites

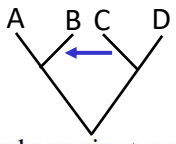
$$Patterson's\ D\ statistic = \frac{ABBA - BABA}{ABBA + BABA}$$

Therefore, *Europeans have more Neanderthal DNA than Africans!*

Green, R.E. et al. 2010. Science

10

Patterson's f_4 statistic



(i) Motivation

A key question is whether Native Americans today descend from a single ancient gene flow event from Asia, or alternatively harbor ancestry from multiple streams of Asian gene flow. To address this, we began by performing *4 Population Tests*[1] using the statistic f_4(*Southern Native American*, *Test Population; Outgroup1*, *Outgroup2*) where the statistic is defined as:

$$f_4(A, B; C, D) = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)(c_i - d_i) \qquad \text{(S6.1)}$$

Here, a_i, b_i, c_i and d_i are the variant allele frequencies at SNP i in populations A, B, C and D respectively. The statistic is proportional to the correlation in allele frequencies differences *(Southern Native American - Test Population)* and *(Outgroup1 - Outgroup2)* over all SNPs. It has an expected value of zero if the *Southern Native American* and *Test Population* are sister groups that descend from a homogeneous ancestral population. By using a Block Jackknife standard error, we obtain an approximately normally distributed Z-score that serves a formal test for whether the 4 populations are consistent with the unrooted tree.

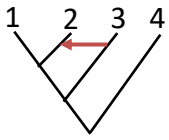
Reich, D. et al. 2012. Nature

11

Simon Martin's f_D statistic, a variant of f

Green et al. (2010) also proposed a related method to estimate *f*, the fraction of the genome shared through introgression (Green et al. 2010; Durand et al. 2011). This method makes use of the numerator of equation 1, the difference between sums of ABBAs and BABAs, which is called S. In the example described above, with $((P_1,P_2),P_3),O)$, the proportion of the genome that has been shared between P_2 and P_3 subsequent to the split between P_1 and P_2 can be estimated by comparing the observed value of S to a value estimated under a scenario of complete introgression from P_3 to P_2. P_2 would then resemble a lineage of the P_3 taxon, and so the denominator of equation 1 can be estimated by replacing P_2 in equations 2 and 3 with a second lineage sampled from P_3, or by splitting the P_3 sample into two,

$$\hat{f}_G = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_{3a}, P_{3b}, O)} \qquad (4)$$



12

Linkage disequilibrium (LD)

When two genes deviate from the expected two locus equilibrium, the genes are said to be in *linkage **dis**equilibrium* (or *gametic disequilibrium*).

The strength of this deviation is measured by the linkage disequilibrium (LD) coefficient, ***D***. Suppose we have two diallelic loci, A/a and B/b

Observed two locus gametic frequencies		random expectation		deviation
p_{AB}	=	$p_A p_B$	+	D
p_{Ab}	=	$p_A(1-p_B)$	-	D
p_{aB}	=	$(1-p_A)p_B$	-	D
p_{ab}	=	$(1-p_A)(1-p_B)$	+	D

$\Sigma = 1$

D can vary between a maximum of +0.25 and a minimum of –0.25, but the range is usually smaller because the frequency of alleles is not exactly 0.5.

(Because $p_{ij} \geq 0$!).

Lewontin & Kojima, 1960, Evolution

13

Factors that affect LD

1. Decay

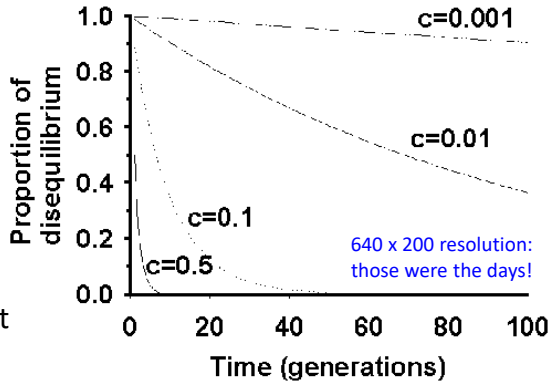
Disequilibrium declines by a fraction given by the recombination rate every generation

If c = recombination rate between genes, then: $D_t = D_{t-1}(1 - c)$

So, after many generations (***t***):

$D_t = D_0(1 - c)^t$

D can therefore decline *by at most 50% in each generation.*



2. Build-up

a) Epistatic selection, b) Genetic drift

14

Standardization of LD

Frequency of gamete AB, $p_{AB} = p_A p_B + D$

To get an idea of the fraction of maximal possible disequilibrium, D is *standardized* in various ways. Often used is the

correlation coefficient: $-1 \le R \le 1$:

$$R_{AB} = \frac{D_{AB}}{\sqrt{p_A p_B (1-p_A)(1-p_B)}}$$

Can also use R_{AB}^2, which measures ~ fraction of variance explained by correlation between the two genes.

(Another common one is $D' = D/D_{max}$ - *which I think is a bit silly!*)

15