

ECOL 600A FUNDAMENTALS OF EVOLUTION

BIOLOGICAL DIVERSITY, POPULATION AND EVOLUTIONARY GENETICS

Lecturer:
Bill Birky
218 Biological Sciences West
birky@u.arizona.edu
626-6513

Background reading: Futuyma Chap. 1, 2 (already assigned).
For people whose genetics background is old or weak, and for basic genetics reference:
Futuyma Chap. 3; Graur & Li Chap 1.

1. Biological Diversity and Diversity of Life Cycles

Read: Futuyma pp. 169-72

Unifying theme of EEB is biodiversity and biodiversification.

Biodiversity seen in the Tree of Life = phylogenetic tree based on sequences of a gene found in all organisms (e.g. small-subunit ribosomal RNA = SSUrRNA) from many organisms.

Structure of TOL is uncertain.

Gene trees often differ from traditional taxonomy based on morphology, more in groups that are farther from vertebrates. "All those microscopic organisms look alike." Gene trees more reliable at reconstructing evolutionary history than morphological trees for organisms other than some animals and plants in the crown.

Where is the diversity?

- Diversity of unfamiliar microscopic organisms is probably under-represented.
- Biological diversity may be greater in bacteria than in eukaryotes. We don't know.
- Diversity may be greater in eukaryotic protists than in crown organisms. We don't know.
- Among animals, is greater in beetles than in all the vertebrates. But diversity in microscopic invertebrates may be greater. We don't know.

Much of what we know about genetics and evolution has been learned from a small group of model systems, mainly animals, plants, fungi, all in the "crown", and *E. coli*.

Genomics is an increasingly important tool for molecular evolution, but it is done mainly on organisms that are important for human health or food (crown, bacteria).
A few exceptions emerging. e.g. Fungal Genome Initiative will sequence genomes of 21 additional fungi, selected partly to sample evolutionary diversity.

Structure of the TOL is determined by:

Macroevolution: tree topology, numbers of branches, etc. are determined by net rate of speciation = speciation rate - extinction rate

Microevolution: rates and patterns of change in sequences (hence of phenotypes) and branch lengths is province of evolutionary genetics; rates of molecular evolution = rates of base pair substitution.

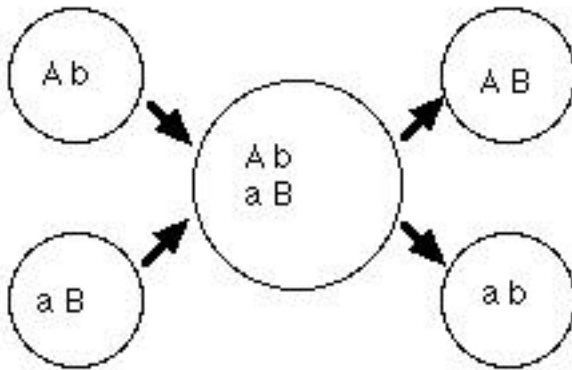
Models in population genetics depend on mode of reproduction: sexual vs. asexual, haploid vs. diploid. Diversity of life is reflected in a diversity of genetic systems.

Asexual reproduction = each individual receives genes from only one parent.

Eukaryotes: mitosis

Prokaryotes: genetically same as mitosis with one chromosome

Sexual reproduction = genes from two individuals put in one cell and recombined into new genotypes.



Prokaryotes

Conjugation, transduction, transformation; sporadic, infrequent

Eukaryotes

Meiosis and fertilization; frequency ranges never -> sporadic -> regular -> obligate

Eukaryotes may have asexual reproduction as haploids, diploids, or both.

Multicellular during haploid or diploid stage or both.

Genes can be counted and modeled in haploid or diploid stages.

2. Measuring Genetic Diversity within Species

Read: Futuyma Chap. 9 pp. 231-247, 253-263

Phenotypic variation may be quantitative (often determined by multiple loci and/or significant environmental effects), or discrete (usually determined by one or a few genes with minor or easily separated environmental effects). This section focuses on discrete variation.

A population is natural or convenient group of individuals of the same species.

Populations may be strongly subdivided into local populations, subpopulations, colonies, or demes.

Diversity is measured at the phenotypic, chromosomal, or increasingly, at the molecular level. Molecular methods used to detect variation in DNA sequences that is hidden at the phenotypic level.

Methods of detecting sequence variation

(a) Protein electrophoresis to detect electrophoretic variants of proteins (in 1955 the wide application of this technique to *Drosophila* by Hubby and Lewontin, and to humans by Henry Harris, began the modern era of molecular investigations of genetic diversity)

Adh gene in *Drosophila melanogaster*: most or all populations have two alleles, *Adh^F* and *Adh^S* (fast and slow). If we looked only at a monomorphic population, we wouldn't know there were two alleles; also a very small sample might have only one.

Limitations: Only applicable to some proteins; does not detect differences between alleles that do not change an amino acid which in turn changes the charge of the protein; can miss some of these. But methods is fast, cheap, and good enough for some applications.

(b) Restriction fragment analysis. Detects DNA sequence variation in restriction sites, called restriction fragment length polymorphisms = RFLP's = variation in restriction patterns

RAPDs and related methods

PCR-amplify DNA using single primers which amplify regions between inverted repeats, or sets of random primers, get different patterns of fragments from different individuals due to variation in regions that are usually anonymous.

(d) Sequencing

e.g. Marty Kreitman cloned and sequenced the *Adh* gene that codes for alcohol dehydrogenase, from 11 different strains of *Drosophila melanogaster* collected at different sites around the world. Each was homozygous so there is only one sequence from each. Included 5 *Adh^F* and 6 *Adh^S* alleles. The sequences were aligned and compared.

Graur & Li p. 59, Fig. 2.9, Table 2.1 summarize results.

- Strains 8, 9, and 10 have identical sequences, so there are 9 different alleles.
- Some but probably not all of these could be detected by RFLP analysis, because some of the differences are not in sites recognized by any restriction enzyme. Fig. 9.10 in Futuyma includes data from later RFLP analysis of 87 gene copies and found additional variable sites.
- Only two alleles are recognized by enzyme electrophoresis.
- All variation within exons, except at *AdhF/AdhS*, site, is synonymous (changes a codon to a synonymous codon) = silent (doesn't change amino acid).
- There are more variable sites in introns than exons, except for exon 4 near *AdhF/AdhS*.

Gene and genotype frequencies

The concepts of gene (allele) and genotype frequencies are the most general and important contributions of Hardy and Weinberg (1908), and are central to population genetics.

gene frequency = frequency of a particular allele of a gene in the population

Simple example: *AdhF* and *AdhS*, (F and S for short) in 100 *Drosophila*.

numbers of genotypes	40 FF	40 FS	20 SS
genotype frequencies	0.40	0.40	0.20
numbers of genes (alleles)	80 F	40 F 40 S	40 S
allele frequencies	0.4 F	0.2 F 0.2 S	0.2 S

Can calculate allele frequencies in two ways:

(1) Out of $2 \times 100 = 200$ genes in the sample gene pool, $80 + 40 = 120$ are F and $40 + 40$ are S, so the frequency of F = $f(F) = 120/200 = 0.6$ and the frequency of S = $f(S) = 80/200 = 0.4$.

(2) $f(F) = f(FF) + f(FS)/2 = 0.4 + 0.2 = 0.6$; $f(S) = f(SS) + f(FS)/2 = 0.2 + 0.2 = 0.4$

Note: The real population probably consisted of > 100 individuals, so these are gene and genotype frequencies in the sample, and *estimates* of the gene and genotype frequencies in the population, or the "gene pool".

Parameters of genetic diversity within populations: gene level

(1) The observed heterozygosity of a gene in a population is the frequency of individuals that are heterozygous for the gene.

(2) The expected heterozygosity of a gene is the probability that two copies of the gene, drawn at random from the population, are different alleles.

We can calculate the probability of drawing different genotypes (pairs of alleles) using basic probability theory. Imagine the alleles in an urn; draw allele, return it, draw another (sampling with replacement, binomial sampling):

$$\begin{aligned} P(\text{draw A}) &= f(A) = p \\ P(\text{draw a}) &= q & p + q &= 1 \\ P(\text{draw A \& A}) &= p^2 \\ P(\text{draw a \& a}) &= q^2 & p^2 + 2pq + q^2 &= 1 \\ P(\text{draw A \& a}) &= 2pq \end{aligned}$$

Note that this is equivalent to a real population that is random mating. Hardy-Weinberg law: so long as p and q are constant, genotype frequencies are given by the above.

$$\text{expected heterozygosity} = h = 2pq = 1 - (p^2 + q^2)$$

More than two alleles: use x_1, x_2, \dots, x_n for frequencies of n alleles.

$$h = 1 - \sum_{i=1}^{i=m} x_i^2 = 1 - (x_1^2 + x_2^2 + \dots + x_m^2)$$

Inbreeding: observed heterozygosity < expected

Expected heterozygosity is a better measure of diversity; a population with many different alleles and genotypes could have zero observed heterozygosity if it was strongly inbred.

Expected heterozygosity varies among organisms and genes, 0 - 0.5.

Parameters of genetic diversity within populations: sequence level

Suppose that we sequence a gene or other segment of DNA, or sample the sequence by restriction analysis. We do this for a sample of individuals from the population. Calculating the parameters from restriction data is complicated. It is easier to understand the use of complete sequences.

(1) proportion of polymorphic sites

Problem with parameter is that it depends heavily on sample size.

(2) nucleotide diversity = $P(\text{a site has a different bp in 2 random copies of a gene})$
= proportion of sites different in 2 random copies of a gene

If x_i and x_j are the frequencies of the i^{th} and j^{th} types of DNA sequences, and d_{ij} is the proportion of bps different in the i^{th} and j^{th} types, then

$$d_{ij} = x_i x_j$$

For small sample sizes this is corrected by multiplying by $n/(n-1)$ where n is the sample size.

To calculate D , a matrix is made which shows the proportion of bases different between each pair of allele. Ignore any site at which there is a gap in any sequence. For Kreitman's data:

$$= [n/(n-1)][f(1)f(2) \binom{n-2}{2} + f(1)f(3) \binom{n-2}{2} + \dots] = (11/10)[(1/11)(1/11)(0.0013) + (1/11)(1/11)(0.0059) + \dots] = 0.007$$

Diversity is high at the sequence level in most species, and varies between species: (data averaged over several genes)

<i>Homo sapiens</i>	1×10^{-3}
<i>Drosophila melanogaster</i>	5×10^{-3}
<i>D. simulans</i> and <i>D. pseudobscura</i>	20×10^{-3}

The ITS in three species of the mosquito *Culex* had values of 0, 2, 3, 6, 16, and 28×10^{-3} .

The mitochondrial *coxI* gene in three species of the mussel *Potamilis* had values of 0, 2.6, and 3.9×10^{-3} . (The last two studies used small samples so I multiplied by $n/(n-1)$).

Patterns of genetic diversity within populations

Diversity varies among organisms

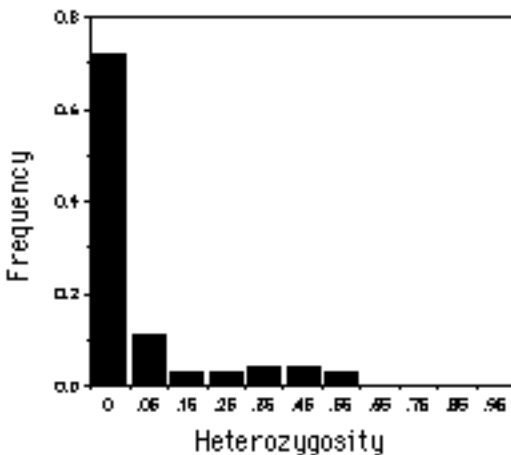
Futuyma Table 9.2 and Figure 9.9 gives examples for electrophoretic variants.

In contrast to most organisms, the cheetah is virtually monomorphic for all genes: in the South African subspecies, $h = 0.0004$ for 49 enzymes and 98 individuals; in the East African subspecies, $h = 0.01$ for 49 enzymes and 30 individuals. This animal is also monomorphic for the major histocompatibility locus, so that skin grafts are usually accepted among "unrelated" individuals.

Diversity varies among genomes

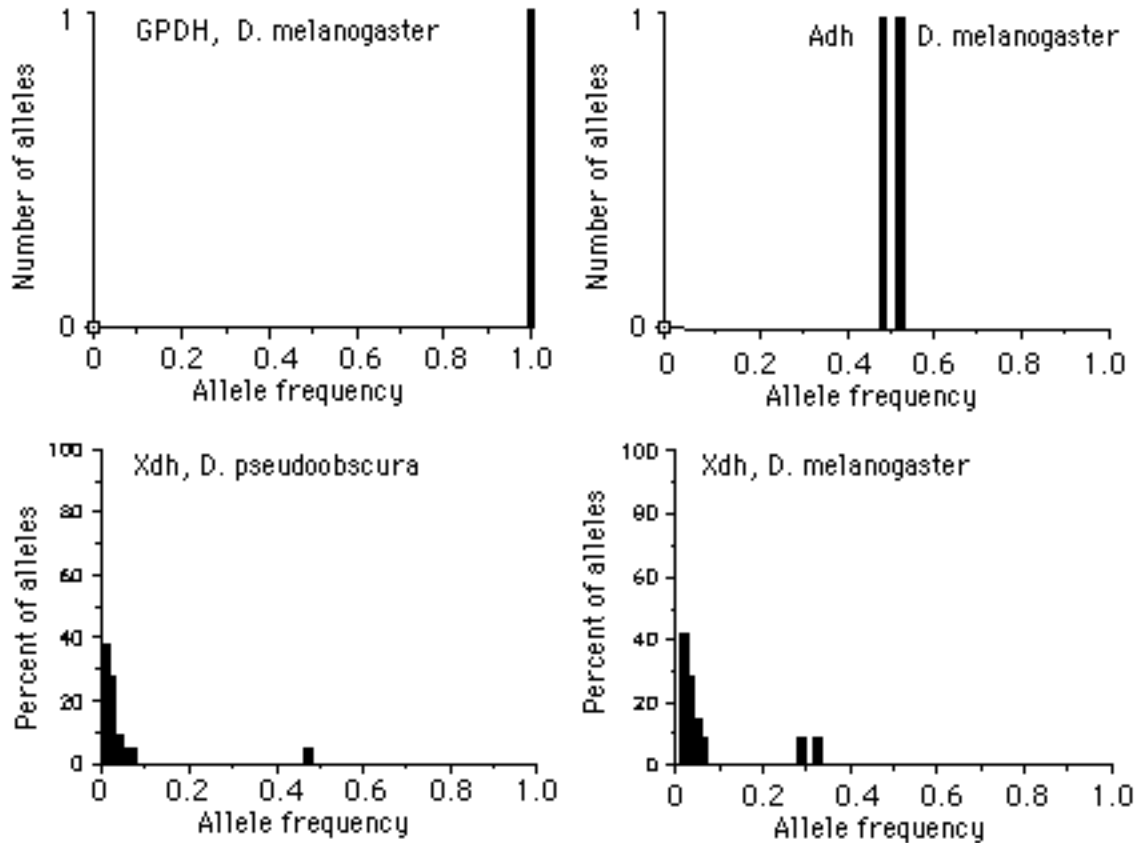
hominids nuclear < mitochondrial
plants nuclear > chloroplast

Diversity varies among genes and other sequences



Distribution of numbers of loci having different values of expected heterozygosity. Humans, 71 loci.

Different genes have different patterns of allele number and allele frequencies:



Frequency distributions of allele frequencies at 4 different loci in *Drosophila*, illustrating Lewontin's 4 classes of loci with respect to diversity.

Genes and other regions also vary in nucleotide diversity. Some protein-coding genes are almost invariant, with the same in almost all individuals, while others are much more variable.

Some non-gene sequences have many different alleles. Extreme examples are the VNTR loci = Variable Number of Tandem Repeats loci.

VNTR's and other highly variable loci are used in:

- forensic medicine, to identify source of blood stains found at the scene of a crime;
- ecological and behavioral studies, to identify parents of individual animals in a social group, e.g. birds or lions;
- genetic markers for recessive genes causing human hereditary diseases.

Diversity varies among regions of a gene

Drosophila melanogaster, *Adh* gene, combined sequence and RFLP data:

	<u>5' untranscribed</u>	<u>transcribed, not translated</u>	<u>exons</u>
polymorphic sites in <i>D. melanogaster</i>	0.024	0.019 (leaders, introns, trailers)	0.013

synonymous polymorphism >> nonsynonymous polymorphism

3. Explaining Genetic Diversity within Species

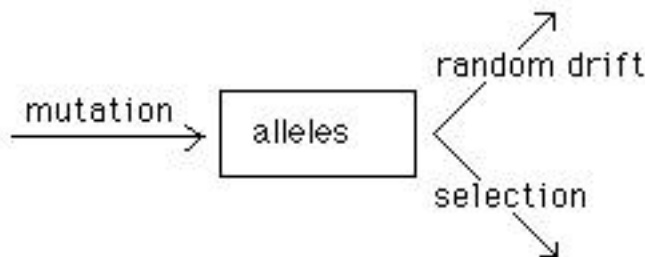
Read: Futuyma: Chap. 10 pp. 267-283; Chap. 11 pp. 297-307; Chap. 12, Chap. 13.
Graur & Li: Chap. 2

4.1 The forces determining both genetic diversity and evolutionary rates include mutation, selection, and random drift.

Mutation is the ultimate source of new alleles.

The fate of a new (mutant) allele is determined by random drift and selection. These forces cause the frequency of a new mutation to increase or decrease until eventually it is fixed in the population or lost.

- (1) mutation (rate)
- (2) selection (kind and strength)
- (3) random drift (effective population size)



Mutation.

Mutation is the ultimate source of genetic variation and differences between species. The probability that a particular base pair or even a gene will undergo mutation in a particular individual is very low, on the order of 10^{-8} to 10^{-9} per base pair or 10^{-4} to 10^{-6} per gene. But the total mutation rate in the population is the rate per base pair (or gene) per gamete times

the number of copies of the gene in the population. The number of copies is simply the number of individuals, times 2 for a diploid organism.

$$M = 2Nu$$

M = mutation rate per site (or gene) per generation

2N = number of successful gametes each generation

u = mutation rate per site (or gene) per gamete

Most mutations must be eliminated, otherwise genetic variation will accumulate until species identity is lost.

OR can have back-back mutations that restore original phenotype (rarely, original genotype).

Model, one locus, 2 alleles, deterministic with forward and back mutation

q = f(A1) wild type

p = f(A2) mutant

mutation rates u = forward rate A1 → A2 v = backward rate A1 ← A2

At mutation equilibrium, $q_e = u/(u+v)$

u >> v (why?)

Population goes to very high frequency of mutant alleles, mainly detrimental (why?)

Infinite alleles model: every mutation → new allele.

Realistic for long genes studied at molecular level.

Diversity increases indefinitely.

In real world, balanced by removal of alleles due to drift and selection.

Random genetic drift.

Why drift happens

Not all individuals in a population produce the same number of offspring, due to their genotype (selection) or chance (drift).

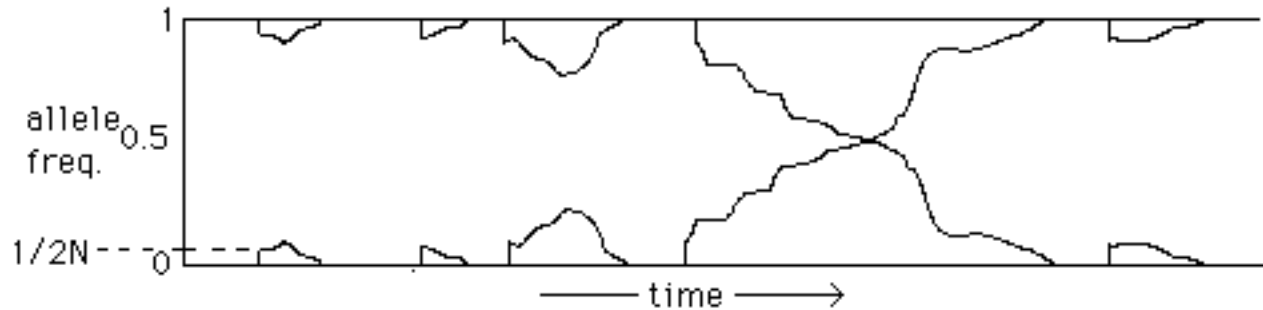
Wright-Fisher model of drift

Binomial sampling (sampling with replacement):

Each generation, draw N copies with replacement to make next generation. Same as drawing N copies without replacement from urn with infinite number of copies. Probability of drawing a specific number of copies of a specified allele follows the binomial distribution. Markov chain in which the probability of going from one state (number of copies of allele) to another is constant and independent of the previous state.

Drift leads to fixation or loss of alleles

Allele frequencies undergo random walk, ending in absorbing state of 0 (loss) or 1 (fixation; alternative alleles lost).



A new mutation begins with a very low frequency of $1/N$ (haploid) or $1/2N$ (diploid) population of N individuals. Random drift is much more likely to eliminate a new mutation than to fix it.

For selectively neutral alleles:

$P(\text{fixation of a neutral allele of frequency } x) = x$

$P(\text{fixation of a new mutation}) = 1/N$ (haploid) or $1/2N$ (diploid)

$P(\text{loss}) = 1 - x = 1 - 1/2N$ or $1 - 1/N$

The strength of random drift depends on the population size

Intuitive explanation: In big populations there are more genes and hence more possible gene frequencies. To go an equivalent distance, the gene frequency must pass through more steps or intermediate frequencies, and the probability of going through more steps in one generation is smaller.

What is important is not the population size per se but:

1. The number of genes, which is $2N$ for diploids and N for haploid species and organelle genes.
2. The effective population size N_e .

In theory, N_e relates random drift in a real population or a more complex model population to random drift in a simpler model of a population. In real life, N_e depends on any factor that determines the variance in offspring number per individual.

Inbreeding effective population size

Model is equal sex ratio; look at effect of unequal sex ratio.

For diploids

$N_e = 4N_m N_f / (N_m + N_f)$ where N_m and N_f are the numbers of breeding animals

$N_e = N$ if $N_m = N_f$, otherwise $N_e < N$

Effective population size with variable N

$N_e =$ harmonic mean of population sizes in generations 1, 2, ... i ... n = $n / (1/N_i)$

This number is closer to the smallest N than to the largest N.

Population bottleneck: N becomes very small, then increases. Genetic diversity is reduced and recovers very slowly (depending on which parameter one uses) as mutations replenish the genetic diversity.

Variance effective population size

Model is Poisson distribution of offspring number. Look at effects of greater variation.

In nearly all cases, $N_e \approx N$.

Humans: $N_e \approx 0.8 N$ or less. Random drift occurs at the same rate that it would in an ideal population 80% of the size of the real population.

We could call N_e the effective number of genes in a haploid organism and $2N_e$ the effective number of genes for a diploid.

The time to fixation or loss

For a neutral mutation of frequency p in a diploid population:

If it is fixed, the mean time to fixation is

$$\bar{t}_1(p) = -4 N_e [(1-p)/p] \ln(1-p) \text{ generations}$$

If it is lost, the mean time to loss is

$$\bar{t}_0(p) = -4 N_e [p/(1-p)] [\ln(p)]$$

Mean time to fixation or loss (time it is segregating in population) is

$$\bar{t}(p) = -4 N_e [p \ln(p) + (1-p) \ln(1-p)] \quad \text{which is very close to the time for loss}$$

New mutation: substitute $1/2N$ for p , get mean time to fixation $\bar{t}_1 (1/2N) = 4N_e$ generations

This is also the time to the coalescent (= most recent common ancestor) of *all* genes in a population. Time to coalescent of any *two* copies of the gene is ca. $2N_e$.

For those many neutral mutations that are lost, the mean time to loss (in generations) is much shorter:

$$\bar{t}_0 (1/2N) = 2(N_e/N) \ln 2N$$

Take-home lesson: most new neutral mutations are lost, very fast.

Random drift always happens. But the process is slower in larger populations, and may be so slow that before a gene is fixed, something else happens. Or it may be negligible.

Combined effects of mutation and drift: neutral theory



If most alleles are neutral or effectively neutral, then the genetic diversity is larger when the total mutation rate Nu is large, and small when the population size is small. Likewise it is larger when N_e is larger, because new mutations remain in the population longer.

Populations left undisturbed a long time reach an equilibrium at which new alleles are eliminated by drift as fast as they are created by mutation. At this equilibrium, in diploids:

$$H = \frac{4N_e u}{1 + 4N_e u} = 4N_e u$$

This equation fits many real situations in which most genetic diversity is due to neutral alleles. Motoo Kimura: the majority of molecular data are explained by neutral theory.

The above equation is for alleles at the level of whole genes; for sequence data

$$\frac{4N_e \mu}{1 + (4/3)4N_e \mu} = 4N_e \mu$$

Of course in this case u is the mutation rate per site (or per base pair), not per gene).

Note that this result can be derived in a very simple way. If the mean time to the coalescent of two neutral alleles is $2N_e$ generations, then the total amount of divergence between the alleles will be approximately $2 \times 2N_e u$.

