Basic Population Genetics

Bruce Walsh lecture notes
Uppsala EQG course
version 28 Jan 2012

Allele and Genotype Frequencies

Given genotype frequencies, we can always compute allele frequencies, e.g.,

$$p_i = \mathrm{freq}(A_i) = \mathrm{freq}(A_i A_i) + \frac{1}{2} \sum_{i \neq j} \mathrm{freq}(A_i A_j)$$

The converse is not true: given allele frequencies we cannot uniquely determine the genotype frequencies

For n alleles, there are n(n+1)/2 genotypes

If we are willing to assume random mating,

$$\mathrm{freq}(A_i A_j) = \begin{cases} p_i^2 & \text{for } i = j \\ 2p_i p_j & \text{for } i \neq j \end{cases}$$

Hardy-Weinberg proportions

Hardy-Weinberg

- Prediction of genotype frequencies from allele freqs
- Allele frequencies remain unchanged over generations, provided:
  - Infinite population size (no genetic drift)
  - No mutation
  - No selection
  - No migration
- Under HW conditions, a single generation of random mating gives genotype frequencies in Hardy-Weinberg proportions, and they remain forever in these proportions

Gametes and Gamete Frequencies

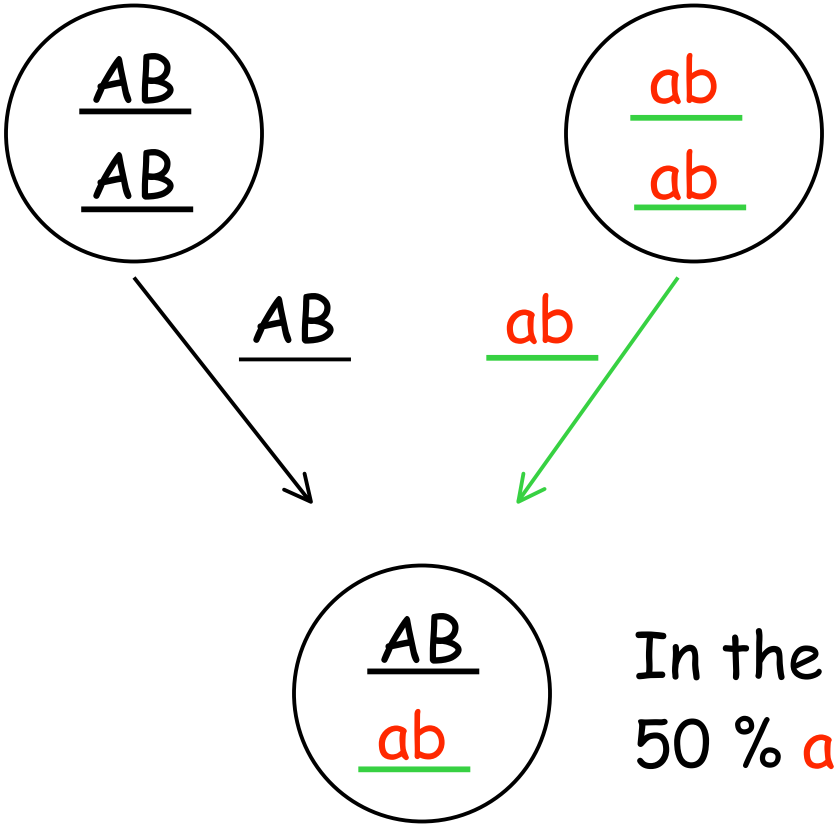
When we consider two (or more) loci, we follow gametes

Under random mating, gametes combine at random, e.g.

$$\text{freq}(AABB) = \text{freq}(AB|\text{father})\,\text{freq}(AB|\text{mother})$$

$$\begin{aligned}\text{freq}(AaBB) = {} & \text{freq}(AB|\text{father})\,\text{freq}(aB|\text{mother}) \\ & + \text{freq}(aB|\text{father})\,\text{freq}(AB|\text{mother})\end{aligned}$$

Major complication: Even under HW conditions, gamete frequencies can change over time



In the F_1, 50% AB gametes
50 % ab gametes

If A and B are unlinked, the F2 gamete frequencies are

AB 25% ab 25% Ab 25% aB 25%

Thus, even under HW conditions, gamete frequencies change

Linkage disequilibrium

Random mating and recombination eventually changes gamete frequencies so that they are in linkage equilibrium (LE). Once in LE, gamete frequencies do not change (unless acted on by other forces)

At LE, alleles in gametes are independent of each other:

$$\mathrm{freq}(AB) = \mathrm{freq}(A)\,\mathrm{freq}(B)$$
$$\mathrm{freq}(ABC) = \mathrm{freq}(A)\,\mathrm{freq}(B)\,\mathrm{freq}(C)$$

The disequilibrium between alleles A and B is given by

$$D_{AB} = \mathrm{freq}(AB) - \mathrm{freq}(A)\,\mathrm{freq}(B)$$

Linkage disequilibrium

When linkage disequilibrium (LD) present, alleles are no longer independent --- knowing that one allele is in the gamete provides information on alleles at other loci

$$D_{AB} = \mathrm{freq}(AB) - \mathrm{freq}(A)\,\mathrm{freq}(B)$$

Positive D: AB gamete more frequent than expected

Negative D: AB gamete less frequent than expected

The Decay of Linkage Disequilibrium

The frequency of the AB gamete is given by

$$\text{freq}(AB) = \boxed{\text{freq}(A)\,\text{freq}(B)} + \boxed{D_{AB}}$$

LE value

Departure from LE

If recombination frequency between the A and B loci is c, the disequilibrium in generation t is

$$D(t) = \boxed{D(0)}(1-c)^t$$

Initial LD value

Note that D(t) -> zero, although the approach can be slow when c is very small

Population Structure

Populations often show structure, with an apparently single random-mating population instead consisting of a collection of several random-mating subpopulations

Suppose there are n subpopulations, and let w_k be the probability that an random individual is from population k

Let p_{ik} denote the frequency of allele A_i in subpopulation k.

The overall frequency of allele A_i is $p_i = \sum_{k=1}^{n} w_k * p_{ik}$

The frequency of A_iA_i in the population is just

$$\mathrm{freq}(A_iA_i) = \sum_{k=1}^{n} w_k p_{ik}^2$$

Expressed in terms of the population frequency of A_i,

$$\begin{aligned}\mathrm{freq}(A_iA_i) &= p_i^2 - \left(p_i^2 - \sum_{k=1}^{n} w_k * p_{ik}^2\right) \\ &= p_i^2 + \mathrm{Var}(p_i)\end{aligned}$$

Thus, unless the allele has the same frequency in each population (Var(p_i) = 0), the frequency of homozygotes exceeds that predicted from HW

Similar logic gives the frequency of heterozygotes as

$$\mathrm{freq}(A_i A_j) = 2p_i p_j + \mathrm{Cov}(p_i, p_j)$$

Hence, when the population shows structure (as does the human population), homozygotes are more common than predicted from HW, while heterozygotes can be more (or less) common than expected under HW.

Population structure also generates disequilibrium

Again suppose there are k subpopulations, each in linkage equilibrium

The population frequency of A_iB_j gametes is

$$\mathrm{Freq}(A_iB_j) = \sum_{\mathrm{k}=1}^{n} w_k \star pA_{ik} \star pB_{jk}$$

The population-wide disequilibrium becomes

$$\begin{aligned} D_{ij} &= \mathrm{Freq}(A_iB_j) - \mathrm{Freq}(A_i) \star \mathrm{Freq}(B_j) \\ &= \sum_{k]=1}^{n} w_k \star p_{A_{ik}} \star p_{B_{jk}} - \left(\sum_{k=1}^{n} w_k \star p_{Aik} \right) \left(\sum_{k=1}^{n} w_k \star p_{Bik} \right) \end{aligned}$$

Consider the simplest case of k=2 populations

Let p_i be the frequency of A_i in population 1, $p_i + \delta_i$ in population 2.

Likewise, let q_j be the frequency of B_j in population 1, $q_j + \delta_j$ in population 2.

The expected disequilibrium becomes

$$D_{ij} = \delta_i * \delta_j * [\, w_1(1 - w_1)\,]$$

Here, w_1 is the frequency of population 1

Genetic Drift

Random sampling of 2N gametes to form the N individuals making up the next generation results in changes in allele frequencies.

This process, originally explored by Wright and Fisher, is called Genetic Drift.

Suppose there are currently i copies of allele *A*, so that freq(*A*) = p = i/(2N)

That probability that, following a generation of random sampling, the freq of *A* is j/(2N) is

This probability follows binominal sampling,

$$\Pr(i \text{ copies} \to j \text{ copies}) = \frac{N!}{(N-j)!j!} \left(\underbrace{\frac{i}{N}}_{p\,=\,i/N}\right)^{j} \left(\underbrace{\frac{N-i}{N}}_{1-\,p}\right)^{N-j}$$

Hence, if the current allele frequency is p, the expected allele frequency in the next generation is also p, but with sampling variance p(1-p)/(2N)

Thus, with N is large, the changes in allele frequency over any generate are expected to be rather small

However, the cumulative effects of generations of such sampling are very considerable.

Eventually, any random allele will either be lost from the population or fixed (frequency one).

If the allele has initial frequency p, then

$$\Pr(\text{Fixation}) = p$$

$$\Pr(\text{loss}) = 1- p$$

The expected time to fixation is on order of 4N generations.

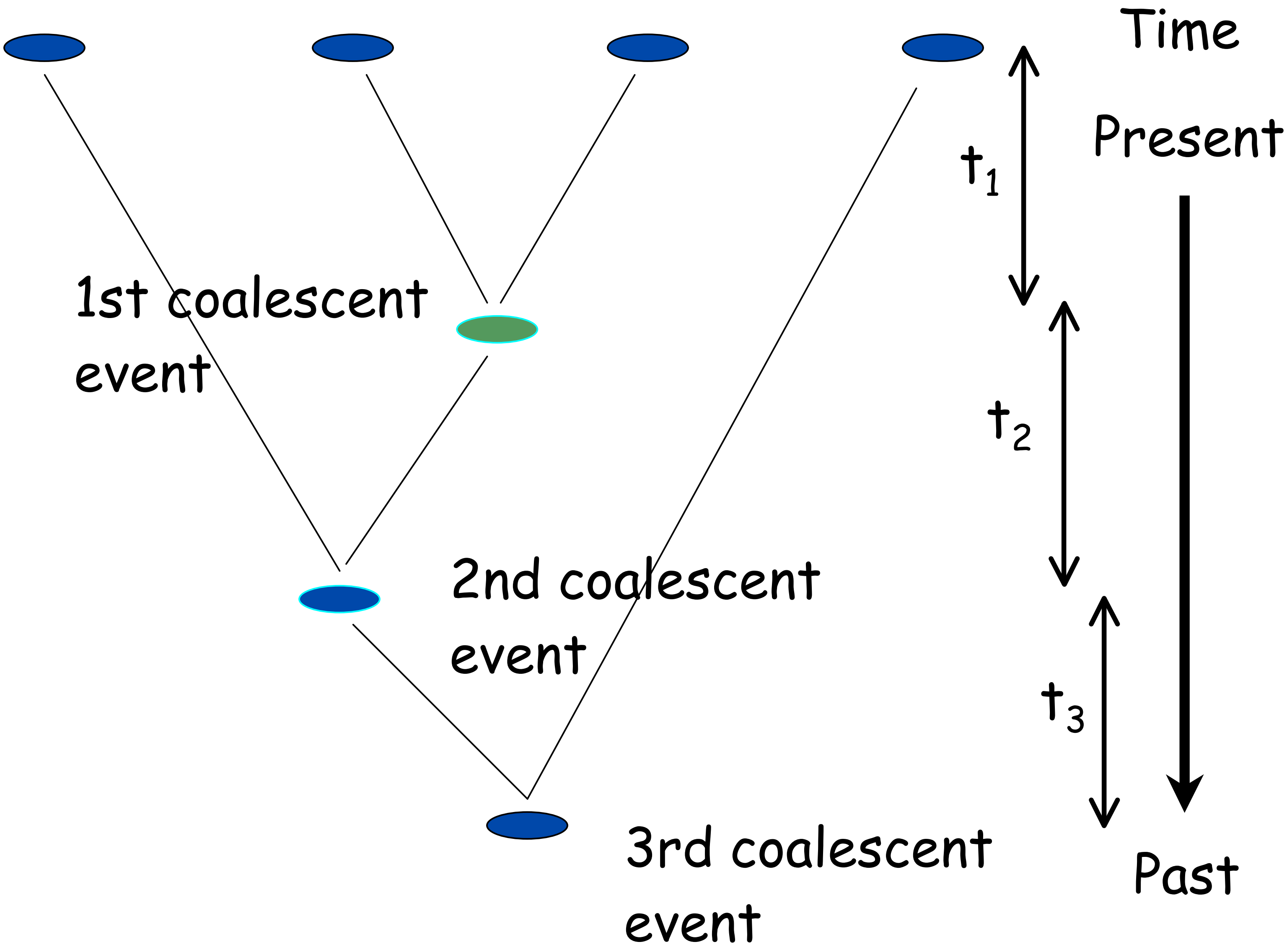
Coalescence Theory

One very powerful way to due statistics under genetic drift is the method of coalescent theory

A consequence of drift is that all of the alleles in a population (or in a sample) all eventually will have descended from a single common ancestral sequence

In this sense, the alleles coalescent in the past, hence the name.

Consider a sample of four alleles:



Time

Present

t_1

t_2

t_3

1st coalescent event

2nd coalescent event

3rd coalescent event

Past

For a sample of k alleles, their genealogical topology (age distribution back to MRCA) is described by the vector $t_1, \ldots, t_k$ of coalescent times

Under pure drift, the distribution of this vector is completely determined by N, the population size

For a sample of two random alleles, time to MRCA follows a geometric distribution with success parameter 1/(2N)

$$\Pr(\text{Coalescence in generation}\, t) = \left(1 - \frac{1}{2N}\right)^{t-1}\left(\frac{1}{2N}\right)$$

The probability that two randomly-chosen alleles have a common ancestor within the last t generations is

$$1 - (1 - 1/[2N])^t$$

Mutation

The second force that can change allele frequencies is mutation.

Mutation is a key fundamental force in that it introduces new variation into the population.

The simplest model is that a gene mutates back and forth between two states, *A* and a.

Let Pr(*A* -> a) = μ and Pr(a -> *A*) = ν

The new frequency, p′, of allele A following a single generation of mutation is

$$p' = p(1-\mu) + (1-p)\nu$$

Randomly-picked allele is A
and this allele DOES not mutate

Random allele is a, but this
allele mutates to A

Allele frequencies change on the order of the mutation rate (typically VERY slow), on order of 10^{-3} to 10^{-9}

At equilibrium, $p' = p = \mu/(\mu+\nu)$

This 2-state model is very unrealistic given what we know about DNA and gene structure

A gene is a DNA sequence of thousands of nucleotides, so back mutation very unlikely

Variations:

Infinite alleles

Infinite sites

Stepwise (for STRs)

Crow and Kimura's (1964) Infinite Alleles Model

Each new mutation leads to a allele

Crow & Kimura looked at the balance between mutation introducing new variation and genetic drift (finite population size) removing it.

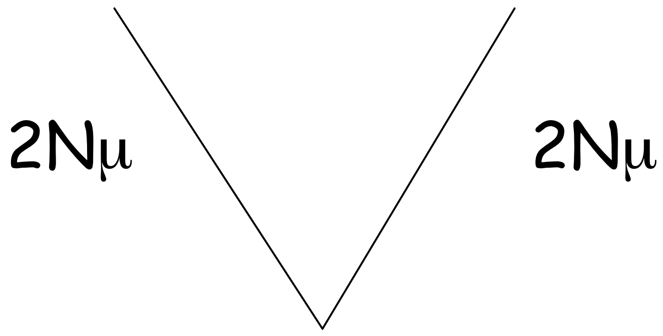
At equilibrium, the expected heterozygosity H of the population is

$$H = \frac{4N\mu}{1 + 4N\mu}$$

We can use coalescent theory to see where this result comes from.

Consider two randomly-chosen alleles. They are different if they have experienced a mutation since their most recent common ancestor (i.e., their coalescence)

For a population of size N, this time has expected value 2N. This gives the expected number of mutations as $4N\mu$.



Expected number = $2N\mu + 2N\mu = 4N\mu$

Number is Poisson-distributed, mean = $4N\mu$

Pr(k mutations) = $(4N\mu)^k \exp(-4N\mu)/k!$

Pr(0 mutations) = $\exp(-4N\mu)$

If $4N\mu \gg 1$, most randomly-drawn alleles will be different. Lots of heterozygosity, H near one.

If $4N\mu \ll 1$, most randomly-drawn alleles will be identical. Almost no heterozygosity, H near zero.

Infinite sites model

- Infinite allele model typically used for the analysis of haplotypes
- Infinite sites model assumes each new mutation occurs at a new site (i.e., nucleotide)
- Used in the analysis of (unphased) DNA sequence data

Stepwise mutational models

A second popular class of mutational models that is motivated from DNA structure are stepwise mutation models

Consider a microsatellite locus (an STR).

ACCACCACCACCACC = $(ACC)_5$

ACCACCACCACCACCACCACC = $(ACC)_7$

Such loci are scored by their repeat number (5 and 7 above). The mutation process on such repeats is that they typically change repeat size by +/- one

Hence, if two alleles have the same repeat size (say 5), this could result from:

The alleles sharing a common ancestor recently enough so that no mutations have occurred.

Or the alleles could have experienced several mutations since their most recent common ancestor. For example, one allele could have been a 6 that mutated to a 5, and the other a 4 that mutated to a 5.

Under stepwise mutation, identity in state (alleles have the same sequence) does not imply identity by descent (no mutation since MRCA)

The symmetric single-step mutation model is the most widely used.

Given allele is in state (repeat number) i,

Pr(stays at i after one generation) = $1-\mu$

Pr(i -> i+1) = $\mu /2$

Pr(i -> i-1) = $\mu /2$

The analysis of even this simple model can be rather involved (requiring the use of Type II Bessel functions)

Selection

One locus with two alleles

Genotype	AA	Aa	aa
Frequency (before selection)	p^2	$2p(1-p)$	$(1-p)^2$
Fitness	W_{AA}	W_{Aa}	W_{aa}
Frequency (after selection)	$\frac{p^2 W_{AA}}{\overline{W}}$	$\frac{2p(1-p) W_{Aa}}{\overline{W}}$	$\frac{(1-p)^2 W_{aa}}{\overline{W}}$

Where $\overline{W} = p^2 W_{AA} + 2p(1-p) W_{Aa} + (1-p)^2 W_{aa}$

is the mean population fitness, the fitness of an random individual, e.g. $\overline{W} = E[W]$

The new frequency p′ of *A* is just freq(*AA* after selection) + (1/2) freq(*Aa* after selection)

$$p' = \frac{p^2 W_{AA} + p(1-p) W_{Aa}}{\overline{W}} = p\,\frac{pW_{AA} + (1-p) W_{Aa}}{\overline{W}}$$

The fitness rankings determine the ultimate fate of an allele

If $W_{XX} \geq W_{Xx} > W_{xx}$, allele X is fixed, x lost

If $W_{Xx} > W_{XX}$, W_{xx}, selection maintains both X & x

Overdominant selection

General expression for n allelles

Let p_i = freq(A_i), W_{ij} = fitness A_iA_j

$$p_i' = p_i \frac{W_i}{\overline{W}}, \qquad W_i = \sum_{j=1}^{n} p_j W_{ij}, \quad \overline{W} = \sum_{i=1}^{n} p_i W_i$$

W_i = marginal fitness of allele A_i

$\overline{W}$ = mean population fitness = $E[W_i]$ = $E[W_{ij}]$

If $W_i > \overline{W}$, allele A_i increases in frequency

If a selective equilibrium exists, then $W_i = \overline{W}$ for all segregating alleles.

Selection and Drift

If the strength of selection is weak relative to the effects of drift, drift will overcome the directional effects of selection.

Suppose genotypes AA : Aa : aa have fitnesses
1 + 2s : 1 + s : 1

Kimura (1957) showed that the probability U(p) that A is fixed given it starts with frequency p is

$$U(p) = \frac{1 - \exp(-4Nsp)}{1 - \exp(-4Ns)}$$

Note if 4Ns >> 1, allele *A* has a very high probability of fixation

If 4Ns << -1 (i.e. allele is selected against), *A* has essentially a zero probability of becoming fixed.

If 4N| s | << 1, the U(p) is essentially p, and hence The allele behaves as if it selective neutral

An interesting case is when p = 1/(2N), i.e., the allele is introduced as a single copy

Even if 4Ns >> 1, U is 2s. Hence, even a strongly favored allele introduced as a single copy is usually lost by drift.