A

Review of Useful Elementary Population Genetics

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

Introduction

Population genetics (and evolutionary genetics) deal with groups of organisms and families, usually natural populations.

- Very large ("ideal") idealised groups or populations (deterministic models)
- Small populations, where stochastic models are necessary (*genetic drift*)

Models we are interested in as genetic epidemiologists:

- Genetic equilibrium models for genotype and haplotype frequencies
- Models for persistence or disappearance of mutants in the population (esp the *neutral model*)
- Selection models for maintenance of variation in the population (eg HbS)
- Coalescent and phylogenetic models of haplotypes in the population

Genotype frequencies

In experimental plant and animal models, we often see entire populations that are homozygous at a particular locus. In natural populations, multiple alleles are often segregating at trait and marker loci, more like the F2 generations in experimental line crosses.

For a codominant trait, we genotype a sample from the population, and count the different genotypes.

Race and Sanger (1975) counts for the MN blood group.

Blood Group (genotype)	M (M/M)	MN (M/N)	N (N/N)	Total
Count (percent)	363 (28.4%)	634 (49.6%)	282 (22.0%)	1279 (100.0%)

The percentages are our best estimate of the probability that an individual will carry that genotype in the population of London, Oxford and Cambridge. The *observed heterozygosity* is 49.6%.

Allele frequencies

There is another population described in the above table. It is the population of gametes that gave rise to individuals tested:

Alleles	M	N	Total
Count (percent)	1360 (53.2%)	1198 (46.8%)	2558 (100.0%)

The percentages here are our best estimate of the probability that a sperm or egg taken from that population will carry that particular allele. If the frequency of the commonest allele at a particular locus is less than 99%, we call this a **polymorphic locus** or **polymorphism**.

Hardy-Weinberg Equilibrium (HWE)

Hardy-Weinberg equilibrium describes the relationship between the gametic or allele frequencies, and the resulting genotypic frequencies. It holds if the following properties are true for the given locus,

1. Random mating or panmixia: the choice of a mate is not influenced by his/her genotype at the locus.
2. The locus does not affect the chance of mating at all, either by altering fertility or decreasing survival to reproductive age.

If these properties hold, then the probability that two gametes will meet and give rise to a new genotype is simply the product of the allele frequencies (binomial expansion):

$Pr(MM) = Pr(M) \times Pr(M)$
$Pr(NN) = Pr(N) \times Pr(N)$
$Pr(MN) = 1 - Pr(MM) - Pr(NN) = 2 \times Pr(M) \times Pr(N).$

HWE rederived

The Hardy-Weinberg rule can be also derived by enumerating all the possible mating types in the population, and using the Mendelian laws to derive the probabilities of the different offspring types. For the parental generation, let Pr(M)=p, Pr(N)=q, Pr(MM)=P, Pr(MN)=Q,Pr(NN)=R, p+q=1, P+Q+R=1:

Mating	Proportion of Matings	Proportion of offspring		
		MM	MN	NN
MM x MM	P^2	P^2	–	–
MM x MN	2PQ	PQ	PQ	–
MM x NN	2PR	–	2PR	–
MN x MN	Q^2	$Q^2/4$	$Q^2/2$	$Q^2/4$
MN x NN	2QR	–	QR	QR
NN x NN	R^2	–	–	R^2
Total	$(P+Q+R)^2$	$(P+Q/2)^2$	2(P+Q/2)(Q/2+R)	$(Q/2+R)^2$
	1	p^2	2pq	q^2

HWE rederived – additional conclusions

- Assumption of random mating affects the calculation of the mating probabilities.
- The HWE genotypic frequencies are attained in one generation, regardless of the distribution of genotype frequencies in the first generation
- We will see later that this is not true for intragametic disequilibrium

Testing HWE

We can easily test for deviation from Hardy-Weinberg equilibrium using a chi-square or exact test. Hardy-Weinberg Disequilibrium can arise from,

1. Genotyping Error
2. Population stratification: multiple subgroups are present within the population, each of which mates only within its own group (homogamy), and the allele frequencies are different within each subgroup (Wahlund effect). Mating within each group is random.
3. Admixture: the breakdown of any of the former processes will lead to deviations until equilibrium is reached.
4. Marital assortment: “like marrying like”: genotypic or phenotypic
5. Inbreeding
6. Decreased viability of a particular genotype: individuals carrying a deleterious genotype die early (or in utero).

Heterozygosity at multiallele markers

Rather than quoting the observed heterozygosity for codominant multiallele markers (as we saw earlier for the blood group example), most workers in human genetics calculate the expected heterozygosity or *gene diversity* based on the allele frequencies and assuming HWE. This is given by,

$$H = 1-\Sigma(p_i^2).$$

The gene diversity of a marker locus is, among other things, a measure of the utility of that marker for linkage analysis.

Linkage equilibrium

There are equilibrium for genotype and gametic frequencies at multiple loci. These are complicated if there is linkage between the loci. We will the examine the case of two loci.

In the parental generation,
a locus A has two allelic forms A and a, with frequencies P_A and $1-P_A$.
A marker B has two alleles B and b (frequency P_B). The recombination fraction between A and B is *c*. A parent can produce a gamete:
AB, *Ab*, *aB*, or *ab*.

The frequency of the different haplotypes in the gametes that gave rise to the parental generation are:
$Pr(AB)=x_1$,
$Pr(Ab)=x_2$,
$Pr(aB)=x_3$ and
$Pr(ab)=x_4$.

At equilibrium, the haplotype frequencies will be the product of the allele frequencies.

Linkage disequilibrium

Linkage disequilibrium is expressed as the difference between this equilibrium value and that observed for the parental generation $D=x_1-P_AP_B$. Another name for linkage disequilibrium is (intragametic) allelic association, where *D* is a measure of the strength of association between the alleles at the two loci eg the A and B alleles.

The gametic distribution emitted by all the parents in a population can be calculated by enumerating all the genotypes and then allowing for recombination events. For example, an AB gamete will be produced by a parent with the AB/AB genotype (population frequency x_{1^2}) with probability 1, and by AB/ab genotype (coupling, population frequency $2x_1x_4$) with probability *(1-c)/2*, and so on. Multiplying and summing probabilities we obtain,

Gamete:	AB	Ab	aB	Ab
Frequency:	x_1-Dc	x_2+Dc	x_3+Dc	x_4-Dc

Linkage disequilibrium recurrence relation

D decreases each subsequent generation according to the recurrence relation [Jennings et al, 1917; Bennett 1954],

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

If the two loci are unlinked, linkage disequilibrium will decrease by 50% in each generation.

For loci separated by a recombination distance of 1%, a 50% decrease would take 69 generations.

This is unlike the case for HWE, where equilibrium is reached after one generation.

Measures of Linkage disequilibrium

By definition, D can take values from $-P_A P_B$ to $min[P_A,P_B]-P_A P_B$. When comparing disequilibrium coefficients for different loci (or even for different alleles at the same multiallele locus), D is often rescaled, either by standardizing it to a binary correlation coefficient (dividing by its variance),

$$r = \frac{D}{\sqrt{P_A(1-P_A)P_B(1-P_B)}},$$

or expressing it as a proportion of its maximal value for the given allele frequencies (D').

$$D' = \frac{D}{\min(P_A, P_B) - P_A P_B},$$

Neither measure is not completely satisfactory. The r^2 measure is best for power calculations, while D' is better for population genetic inference.

Extent of linkage disequilibrium in humans

Many studies have attempted to survey the extent of linkage disequilibrium between loci in humans. Reich et al [2001] found *D'* in admixed-type European and US populations to average 0.95 between loci separated by 5 kbp, 0.50 at 80 kbp, and 0.35 at 160 kbp (the average *D'* value for unlinked loci was 0.15).

The extent of LD is greater in African populations, and fairly comparable in European and Asian outbred populations.

It will be greater in isolated populations, where the number of founders is small: Ashkenazi Jews in Eastern Europe, Northern Finland.

Mutation and linkage disequilibrium

If a new allele appears in a particular individual, and subsequently spreads through the population, alleles at loci closely linked to the mutated locus will be in linkage disequilibrium (associated) with the new allele.

These alleles present in that first individual, make up an *ancestral haplotype* associated with the new trait.

The length of this ancestral haplotype (in cM) is proportional to the age of the initial trait mutation, approximately:

$t \cong \frac{1}{r}$, where r is the haplotype length, and $t > 20$.

(eg Piccolo et al 1993).

Linkage disequilibrium based estimation of allele age

The simplest model applies to two locus haplotypes. A marker locus **B** (alleles *B* and *b* with frequencies P_B and 1-P_B) is close to the trait locus **A** (recombination fraction *c*) and the trait mutation (*A* allele) occurred on a haplotype carrying the *B* allele marker.

We assume there is no recurrent mutation.
At the time of the mutation, linkage disequilibrium is at its maximum value D_{max}. After *t* generations it decays to its present value of *D*. Remembering our definition of *D'*, we can reorganize the formula $D^{(t)}=(1-c)^t D^{(0)}$ as:

$$t = log(D')/log(1-c)$$

Using a different approach, Kaplan and Weir [1995] derive a very similar equation:

$$t = log(D')/(-c).$$

Age of the torsion dystonia gene

An example of such a calculation is for idiopathic torsion dystonia (locus *DYT1*) among the Ashkenazi Jews. The closest genetic marker to *DYT1* in the study of Risch et al (1995) was *ASS*. This was approximately 0.018 cM distant from *DYT1*. The estimate of D' was (0.806-0.086)/(1-0.086)=0.788, giving the age of the disease allele as t^*=13.1 generations.

Age of an allele based on frequency

Slatkin and Rannala [1997] present one simple model for assessing the age of a mutant allele based solely on its population frequency.

$$t^*_1 = 4\,N\,p$$

where N is the population size, and p is the present allele frequency. In the case of an exponentially expanding population or one where the mutant allele is undergoing selection,

$$t^*_1 = log(4Np(r+s) + 1)/(r+s)$$

where r is the exponential growth rate parameter (approximately the proportional increase per generation), and s is the selection coefficient for heterozygotes. The equivalence between selection and exponential population growth is very approximate.

Values of r of 0.004 to 0.016 are plausible for older large populations (the world, or Western Europeans).

Effects of population size growth on allele frequency

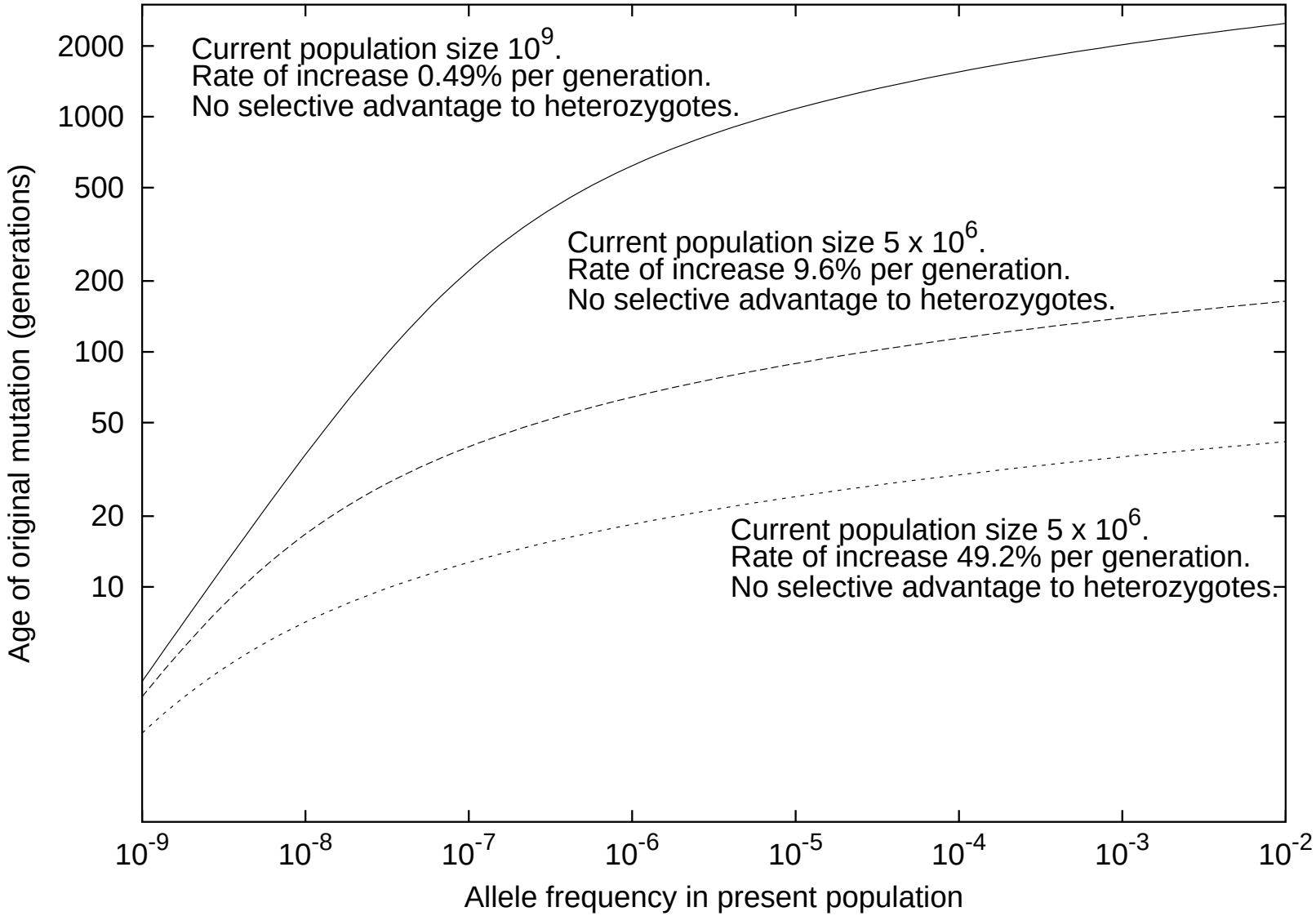
For the world population, assuming N_0=5000 founders t=2500 generations ago gave rise to to N_t=10^9 current descendents, substituting into:

$$N_t = N_0 \, exp(rt)$$

gives r=0.0049.

Hastabacka et al [1992] modelled r for the Finnish population as 0.09.

This gives rise to the following relationship in such populations (assuming s=0):


Current population size 10^9.
Rate of increase 0.49% per generation.
No selective advantage to heterozygotes.
Current population size 5 x 10^6.
Rate of increase 9.6% per generation.
No selective advantage to heterozygotes.
Current population size 5 x 10^6.
Rate of increase 49.2% per generation.
No selective advantage to heterozygotes.
Age of original mutation (generations)
2000
1000
500
200
100
50
20
10
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
Allele frequency in present population

QIMR

Age of the gene for idiopathic torsion dystonia 2

In Eastern Europe, the Askenazi Jewish population increased rapidly in size from approximately 10^5 in 1650 to 5×10^6 in 1900, giving us an estimated r of 0.40.

Risch et al [1995] estimate the *DYT1* allele frequency in the current population at 1/6000 to 1/2000. This gives an estimate of the age of the first mutation at 18-20 generations ago (about the year 1500).

Research questions

- Are common human diseases due to common variants or multiple rare variants?
- Will rare or common SNPs be better candidates for a particular disease?
- If a disease suceptibility allele is common in one population (eg *ApoE*4)*, does this represent the effects of selection (eg heterozyogte advantage)?
- If so, will treatment targetting that gene product be a net harm?
- Can large differences between populations in the frequency of an allele be merely due to chance?

Population genetics of the SLC6A4 promoter polymorphism

Caspi et al [2003] found that the “short” allele at the (VNTR) promoter polymorphism in the serotonin transporter predicted a greater risk of depression in the face of adverse life events.

Genotype		Odds Ratio (per life event)
L/L	(31%)	1.13 (0.83-1.56)
S/L	(51%)	1.47 (1.08-2.02)
S/S	(18%)	1.68 (1.22-2.30)

Population genetics of the SLC6A4 promoter polymorphism

The frequency of the short allele varies markedly across populations.

Group	N	Frequency of S Allele
East Asia	551	0.77
Askenazi Israelis	224	0.52
UK	461	0.43
European Americans	221	0.43
Italy	552	0.41
African Americans	1210	0.23