Gene Frequencies

By William Beavis, Kendall Lamkey (ISU)

Objectives

- Demonstrate the relevance of population genetics concepts to plant breeding populations.
- Demonstrate the relevance of a purely theoretical Ideal Population to plant breeding populations.
- Demonstrate understanding for purpose of populations in Harey-Weinberg Equilibrium
- Distinguish populations in Hardy-Weinberg Equilibrium from the Ideal Population.
- Describe impact of mutation, selection and drift on breeding populations.

Fig. 1 A red Darwin hybrid tulip "Appeldoorn" with a mutation resulting in half of one petal being yellow. Photo by LepoRello; Licensed under CC BY-SA 3.0 via Wikimedia Commons.

Introduction

The challenge of Quantitative Genetics is to connect traits measured on quantitative scales with genes that are inherited and evaluated as discrete units. This challenge was addressed through development of theory between 1918 and 1947. The theory is now referred to as the **modern synthesis**, and required another 50 years for technological innovations and experimental biologists to validate. Luminaries such as RA Fisher, Sewell Wright, JBS Haldane and John Maynard Smith were able to develop theory that is still widely applied without the benefit of high throughput 'omics' technologies. Indeed the modern synthesis was developed before the knowledge of the structure of DNA.

Population genetics characterizes how discrete units, i.e., alleles, change in breeding populations. Such characterization is the basis for understanding structure of genomes and breeding populations. The forces of mutation, migration, selection and drift will alter the structure of breeding populations. Herein we will learn how to characterize population structure at one or two loci in diploid crop species. This will set the foundation for characterizing structure based on any number of loci and for polyploid crops that you may encounter in more advanced courses.

Allelic and Genotypic Variation

Ideal Population

In order to understand the genetic structure of a population it is necessary to establish a standard reference population so that the breeding population can be characterized relative to the standard. For this purpose an 'ideal' conceptual base population can be defined as infinitely large with potential to extract finite sub-populations through sampling such as depicted in the following figure and described in Falconer and Mackay (1996):

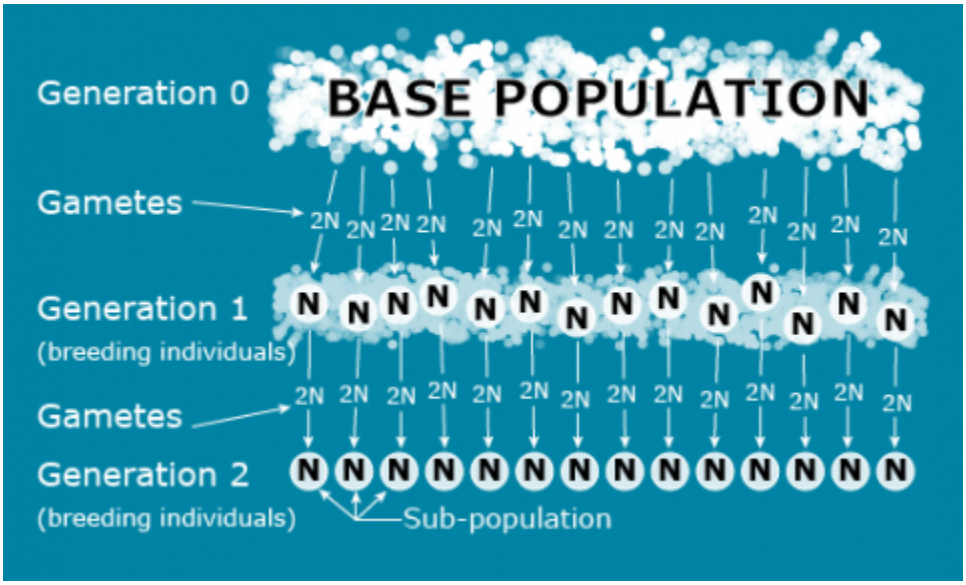


Fig. 2 Reference population. Adapted from Falconer and Mackay, 1996.

Note, that the sub-populations depicted in **Fig. 1** are based on a genetic sampling process that is affected by the reproductive biology of the species. Unlike animal species, crop species can reproduce in a variety of ways:

- Sexual
    - Cross Pollination
    - Self Pollination
    - Mixtures of Self and Cross pollination

- Asexual
    - Clonal
    - Doubled haploids
    - Apomixis

Assumptions

In the ideal population depicted in **Fig. 2**, the following assumptions are true:

1. The base population is infinite, or at least too large to count.
2. There is no migration between sub-populations.
3. There is no breeding between overlapping generations.
4. The number of breeding individuals is the same in each sub-population.
5. There is random mating within a sub-population.
6. There is no Selection.
7. There is no Mutation.

Of course in real populations these assumptions are violated.

Allelic and Genotypic Frequencies

We first model a single locus with only two alleles in an ideal breeding population of diploid individuals. Define the following:

N = number of breeding individuals in a sub population (population size)
t = time usually measured in terms of generations
q = frequency of one of two alleles at a locus within a sub population
p = 1 – q = frequency of a second allele at a locus within a sub population
$\overline{p}$ = frequency of a second allele across the sub populations (the mean of p)
p_0 = frequency of a second allele in the base population

Due to the assumptions associated with an ideal reference population, $\overline{q}$ = q_0 at any stage or generation of the sampling process, so q_0 can be used interchangeably with $\overline{q}$.

The alleles, allele frequencies, genotypes and genotypic frequencies can be represented as follows:

Table 1

	Alleles		**Genotypes**		
	A	a	AA	Aa	aa
Frequency	p	q	P_{AA}	P_{Aa}	P_{aa}

Where, p + q = 1 and P_{AA} + P_{Aa} + P_{aa} = 1.

Variance of Allele Frequency

The relationship between allele frequencies and genotype frequencies can be expressed as:

Invalid Equation

A particular sub-population is a random sample of N individuals or 2N gametes (for a diploid) from the base population. Therefore, the expected gene frequency of a particular allele in the sub-populations is q_0 and the variance of q is

$$\sigma_q^2 = \frac{p_0 q_0}{2N}$$

Since q_0 is a constant, the variance of the change in allele frequency (q_1 - q_0) is also:

$$\sigma_{\Delta q}^2 = \frac{p_0 q_0}{2N}$$

Frequency Estimators

In addition to the genetic sampling process depicted in **Fig. 2**, a statistical sampling process can be used to estimate frequencies, variances and covariances of alleles and genotypes in a sub-population. If we sample n individuals from a population of size N, and notationally let

$$n = n_{AA} + n_{Aa} + n_{aa}), then (n_A = 2n_{AA} + n_{Aa}) and (\widehat{p}_A = \frac{1}{2n} n_A$$

is the sample frequency of the A allele and

$$\widehat{p}_{AA} = \frac{1}{n} n_{AA}$$

is the sample frequency of genotype AA.

Expected Number of Alleles

$$Var\left(\hat{p}_A\right) = \frac{1}{2n}\left(p_A + P_{AA} - 2p_A^2\right)$$

Recognizing that statistical sampling at a locus with two alleles in a diploid population is represented as a binomial random process, the expected number of A alleles in a sample is

$$E(n_A) = 2nP_{AA} + nP_{Aa} and (E(\widehat{p}_A) = P_A$$

Thus $\widehat{p}_A$ is an unbiased estimator of the population parameter p_A.

Using the definition of variance we can likewise find

$$Var(n_A) = 2n(p_A + P_{AA} - 2p_A^2)$$

and the

$$Var\left(\hat{p}_A\right) = \frac{1}{2n}\left(p_A + P_{AA} - 2p_A^2\right)$$

Note that $p_A + P_{AA}$ are usually unknown, so we often substitute $\hat{p}_A + \hat{P}_{AA}$ in the calculation of the Var ($\hat{p}_A$). Note that Var($\hat{p}_A$) is not the variance of a Binomial distribution. If the population sampled is in Hardy-Weinberg Equilibrium (see below) the genetic sampling of alleles will be random so that

$$P_{AA} = p_A^2 \text{ and } P_{Aa} = 2p_A p_a$$

and

$$Var\left(\hat{p}A\right) = \frac{1}{2n}p_A\left(1 - p_A\right)$$

which has the form of the variance from a binomial distribution.

Hardy-Weinberg Equilibrium

Assumptions

Proof of Hardy Weinberg Equilibrium (HWE)

The proof of HWE requires the following assumptions (Falconer and Mackay, 1996):

1. Allele frequency in the parents is equal to the allele frequency in the gametes
    1. Assumes normal gene segregation
    2. Assumes equal fertility of parents
2. Allele frequency in gametes is equal to the allele frequency in gametes forming zygotes
    1. Assumes equal fertilizing capacity of gametes
    2. Assumes large population
3. Allele frequency in gametes forming zygotes is equal to allele frequencies in zygotes
4. Genotype frequency in zygotes is equal to genotype frequency in progeny
    1. Assumes random mating
    2. Assumes equal gene frequencies in male and female parents
5. Genotype frequencies in progeny does not alter gene frequencies in progeny.
    1. Assumes equal viability

For a two allele locus in a population in HWE:

$$P_{AA} = p^2), (P_{Aa} = 2pq), (P_{aa} = q^2$$

HWE at a given genetic locus is achieved in one generation of random mating. Genotype frequencies in the progeny depend only on the allele frequencies in the parents and not on the genotype frequencies of the parents.

Disequilibrium

As discussed there are several processes that can force allelic and genotypic frequencies to deviate from HWE. Deviations from equilibrium are referred to as Disequilibrium, and are often denoted with a disequilibrium coefficient, D. In the two allele case the genotypic frequencies can be represented as

$$P_{AA} = p_A^2 + D_A), (P_{Aa} = 2p_Ap_a - 2D_A), (P_{aa} = p_a^2 + D_A) Thus, (\widehat{D}_A = \widehat{P}_{AA} - p_A^2) Notethat (E(\widehat{D}_A) = D_A - \frac{1}{2n}[p_A(1 - p_A) + D_A]$$

Thus $\widehat{D}_A$ is biased. Although the estimate of D_A is biased, as the sample size, n, becomes large the bias becomes small. Thus, emphasizing the need for large sample sizes in drawing inferences about Disequilibrium from Hardy-Weinberg.

Variance

The Var($\widehat{D}_A$) can likewise be derived as

$$\cong \frac{1}{n}[p_A^2(1-p_A)^2 + (1-2p_A)^2 D_A - D_A^2]$$

If n is large, $E(\widehat{D}_A) \cong D_A$ and

$$\widehat{D}_A \sim N[E(\widehat{D}_A), Var(\widehat{D}_A)]$$

So that a standard normal variate, Z can be constructed as:

$$Z = \frac{\widehat{D}_A - E(\widehat{D}_A)}{\sqrt{Var(\widehat{D}_A)}}$$

Goodness of Fit

Alternatively, because $Z^2 = X^2 \sim X^2$

$$X_A^2 = \frac{nD_A^2}{p_A^2(1-p_A)}$$

which enables the direct use of genotypic 'counts' n_{AA}, n_{Aa}, n_{aa}; i.e.

Table 2

	Genotypes		
	AA	**Aa**	**aa**
Observed (O)	n_{AA}	n_{Aa}	n_{aa}
Expected (E)	np^2_A	$2np_A(1-p_A)$	$n(1-p_A)^2$
O-E	nD_A	$-2n\hat{D}_A$	nD_A

The Goodness of Fit Statistic

$$\chi_A^2 = \frac{(0-E)}{E}$$

Non-Random Mating

Two methods of non-random mating that are important in plant breeding are assortative mating and disassortative mating.

Assortative mating occurs when similar phenotypes mate more frequently than they would by chance. One example would be the tendency to mate early x early maturing plants and late x late maturing plants. The effect of assortative mating is to increase the frequency of homozygotes and decrease the frequency of heterozygotes in a population relative to what would be expected in a randomly mating population. Assortative mating effectively divides the population into two or more groups where matings are more frequent within groups than between groups.

Disassortative mating occurs when unlike or dissimilar phenotypes mate more frequently than would be expected under random mating. Its consequences are in general opposite those of assortative mating in that disassortative mating leads to an excess of heterozygotes and a deficiency of homozygotes relative to random mating. Disassortative mating can also lead to the maintenance of rare alleles in a population.

Factors Affecting Allele Frequency

The factors affecting changes in allele frequency can be divided into two categories: **systematic processes**, which are predictable in both magnitude and direction and **dispersive processes**, which are predictable in magnitude but not direction. The three systematic processes are migration, mutation, and selection. Dispersive processes are a result of sampling in small populations.

Migration

Assume a population has a frequency of m new immigrants each generation, with 1-m being the frequency of natives. Let q_m be the frequency of a gene in the immigrant population and q_0 the frequency of the same gene in the native population. Then the frequency in the mixed population will be:

$$q_1 = mq_m + (1-m)q_0)(= m(q_m - q_0) + q_0$$

The change in gene frequency brought about by migration is the difference between the allele frequency before and after migration.

$$\Delta q = q_1 - q_0)(= m(q_m - q_0)$$

Thus the change in gene frequency from migration is dependent on the rate of migration and the difference in allele frequency between the native and immigrant population.

Mutations

Mutations are the source of all genetic variation. Loci with only one allelic variant in a breeding population have no effect on phenotypic variability. While all allelic variants originated from a mutational event, we tend to group mutational events in two classes: rare mutations and recurrent mutations where the mutation occurs repeatedly.

Rare Mutations

By definition a rare mutation only occurs very infrequently in a population. Therefore, the mutant allele is carried only in a heterozygous condition and since mutations are usually recessive, will not have an observable phenotype. Rare mutations will usually be lost, although theory indicates rare mutations can increase in frequency if they have a selective advantage.

Fate of a Single Mutation

Consider a population of only AA individuals. Suppose that one A allele in the population mutates to a. Then there would only be one Aa individual in a population of AA individuals. So the Aa individual must mate with a AA individual.

AA x Aa → 1AA:1Aa

From Li (1976; pp 388), this mating has the following outcomes:

1. **No offspring are produced** in which case the mutation is lost.
2. **One offspring is produced:** the probability of that offspring being AA is 1/2 so the probability of losing the mutation is 1/2.
3. **Two offspring are produced:** Aa can mate with more than one of the AA individuals in the population, thus if Aa mates with two AA individuals, the probability of both offspring being AA is 1/4, so the probability of losing the mutation is 1/4.

If k is the number of offspring from the above mating then the probability of losing the mutation among the first generation of progeny is $(1/2)^k$.

Probability of Loss

The probability of losing the gene in the second generation can be calculated by making the following assumptions:

- Number of offspring per mating is distributed as a Poisson process (which means that they follow a stochastic distribution in which events occur continuously and independently of one another).
- With the average number of offspring per mating = 2.
- New mutations are selectively neutral.

With these assumptions, the probabilities of extinction are:

Table 3 Probability of extinction in different generations.

Generation	Probability of Loss
1	0.37
7	0.79
15	0.89
31	0.94
63	0.97
120	0.98

Recurrent Mutations

Let the mutation frequencies be:

$$\begin{array}{lccc} \text{Mutation Rate} & A & \underset{v}{\overset{u}{\rightleftharpoons}} & a \\ \text{Frequency} & p_0 & & q_0 \end{array}$$

Then the change in gene frequency in one generation is:

$$\Delta q = up_0 - vq_0) at equilibrium (pu = qv) (q = \frac{u}{v+u}$$

Conclusions:

- Mutations alone produce very slow changes in allele frequency
- Since reverse mutations are generally rare, the general absence of mutations in a population is due to selection

Selection

Selection is one of the primary forces that will alter allele frequencies in populations. **Selection** is essentially the differential reproduction of genotypes. In population genetics this concept is referred to as **fitness** and is measured by the reproductive contribution of an individual (or genotype) to the next generation. Individuals that have more progeny are more fit than those who have less progeny because they contribute more of their genes to the population.

The change in allele frequency following selection is more complicated than for mutation and migration, because selection is based on phenotype. Thus, calculating the change in allele frequency from selection requires knowledge of genotypes and the degree of dominance with respect to fitness. Selection affects only the gene loci that affect the phenotype under selection—rather than all loci in the entire genome—but it also would affect any genes that are linked to the genes under selection.

Effects of Selection

Change in allele frequency

The strength of selection is expressed as a coefficient of selection, s, which is the proportionate reduction in gametic output of a genotype compared to a standard genotype, usually the most favored. Fitness (relative fitness) is the proportionate contribution of offspring.

Partial selection against a completely recessive allele

To see how the change in allele frequency following selection is calculated consider the case of selection against a recessive allele:

Table 4

	Genotypes			
	AA	**Aa**	**aa**	**Total**
Initial Frequencies	p^2	2pq	q^2	1
Coefficient of Selection	0	0	s	
Fitness	1	1	1 - s	
Gametic Contribution	p^2	2pq	$q^2(1 - s)$	$1 - sq^2$

Frequency Equations

The frequency of allele *a* after selection is:

Invalid Equation

The change in allele frequency is then:

Invalid Equation

In general, you can show that the number of generations, t, required to reduce a recessive from a frequency of q_0 to a frequency of q_t, assuming complete elimination of the recessive (s = 1) is:

$$t = \frac{1}{q_t} - \frac{1}{q_0}$$

Small Population Size

Unlike the three systematic forces that are predictable in both amount and direction, changes due to small population size are predictable only in amount and are random in direction.

The effects of small population size can be understood from two different perspectives. It can be considered a sampling process and it can be considered from the point of view of inbreeding. The inbreeding perspective is more interesting, but looking at it from a sampling perspective lets us understand how the process works.

Consequences of small population size

1. Random genetic drift: random changes in allele frequency within a subpopulation
2. Differentiation between subpopulations
3. Uniformity within subpopulations
4. Increased homozygosity

Example 1: Let q = 0.5 and N = 50, then

Invalid Equation

Example 2: Let q = 0.5 and N = 4, then

Invalid Equation

Inbreeding and Small Populations

Inbreeding is the mating together of individuals that are related by ancestry. The degree of relationship among individuals in a population is determined by the size of the population. This can be seen by examining the number of ancestors that a single individual has:

Table 5

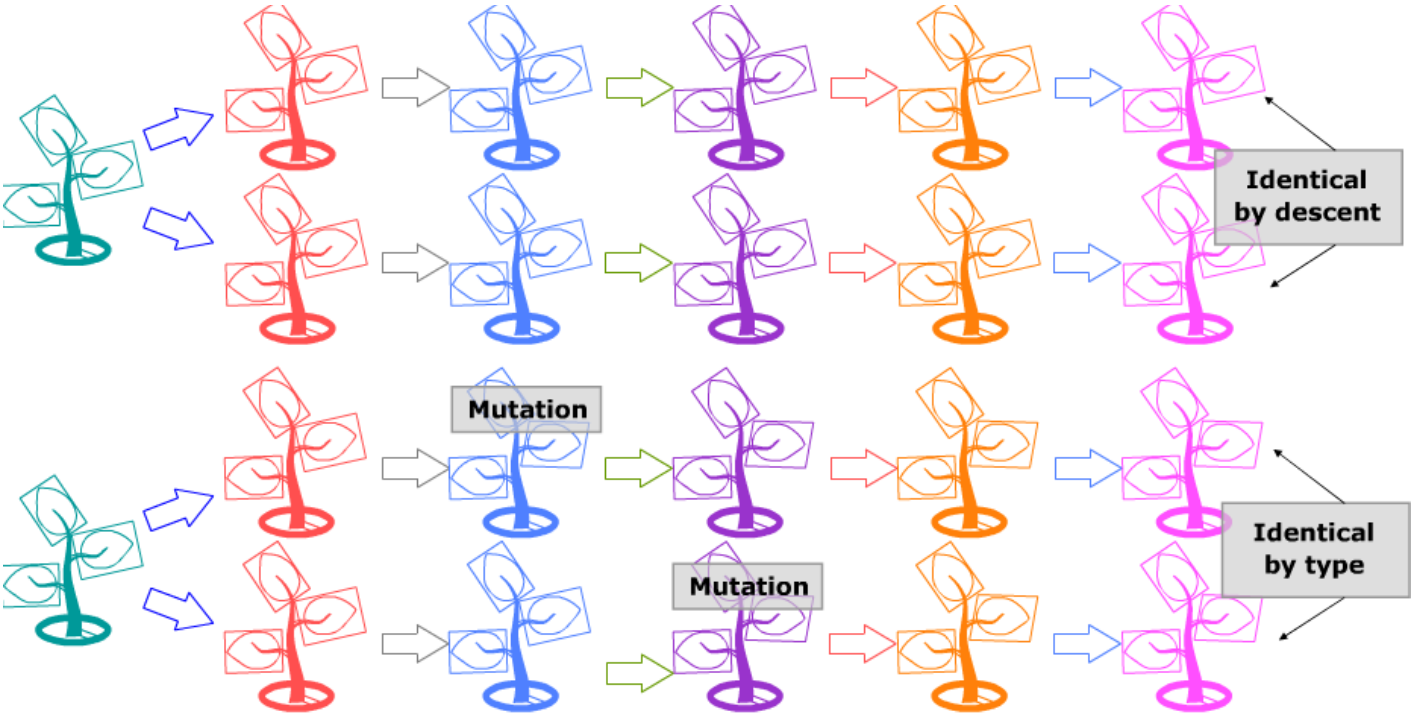
Generation	Ancestors
0	1
1	2
2	4
3	8
4	16
5	32
6	64
10	1,024
50	1,125,899,906,842,620
100	1,267,650,600,228,230,000,000,000,000,000
t	2^t

Just 50 generations ago note that a single individual would have more ancestors than the number of people that have existed or could exist on earth.

Therefore, in small populations individuals are necessarily related to one another. Pairs mating at random in a small population are more closely related than pairs mating together in a large population. Small population size has the effect of forcing relatives to mate even under random mating, thus with small population sizes inbreeding is inevitable.

Identical by Descent and Identical by State

In finite populations there are two sorts of homozygotes: Those that arose as a consequence of the replication of a single ancestral gene — these genes are said to be identical **by descent** (Bernardo, 1996). If the two genes have the same function, but did not arise from replication of a single ancestral gene, they are said to be alike **in state**. It is the production of homozygotes that are identical by descent that gives rise to inbreeding in a small population.



Coefficient of Inbreeding

The probability that two genes are identical by descent is called the **coefficient of inbreeding** and will be the measure of relationship between mating pairs.

The coefficient of inbreeding (**F**) refers to the individual and expresses the degree of relationship between an individual's parents. The coefficient of inbreeding is always expressed relative to a specified base population. The reference population is assumed to be non-inbred (F=0).

Consider a base population consisting of N individuals each shedding equal numbers of gametes uniting at random. Because the base is non-inbred, each individual in this population carries genes that are non-identical. The only way a homozygote that carries genes that are identical by descent can arise is by the mating of a male and female gamete from the same individual that carry a replication of the same gene. Because there are 2N gametes the probability that two mating gametes are identical by descent is $\frac{1}{2N}$.

Equation of Coefficient of Inbreeding

In the second generation there are two ways genes are identical by descent can be joined:

1. by a new replication of the same ancestral gene; and
2. by the previous replication that occurred in generation 1.

The probability of a new replication event is $\frac{1}{2N}$. The remaining proportion of zygotes, $1-\frac{1}{2N}$, carry genes that are independent in origin from generation 1, but may have been identical in their origin in generation 0. The probability the genes are identical by descent from generation 1 is the inbreeding coefficient of generation 1 ($F_1 = \frac{1}{2N}$).

Therefore the probability of identical homozygotes in generation 2 is:

$$F_2 = \frac{1}{2N} + (1 - \frac{1}{2N})F_1$$

Where F_1 and F_2 are the inbreeding coefficients of generations 1 and PG. The same arguments apply to future generations, so we can write the recurrence equation:

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}$$

Inbreeding Coefficient

The inbreeding of any generation is composed of two components: New inbreeding, which arises from self-fertilization and the “old” that was already there.

Note that inbreeding is cumulative and that the absence of inbreeding in generation t does not change the fact that a population has inbreeding from prior generations.

Through a series of algebraic steps, we can write the inbreeding coefficient as a function of the number of generations removed from the reference populations:

$$F_1 = 1 - (1 - \Delta F)^t) where, (\Delta F = \frac{1}{2N}$$

Dispersion

To relate inbreeding back to population size, we can rewrite the variance of the change in allele frequency:

Invalid Equation

Δ expresses the rate of dispersion and F expresses the amount of dispersion.

Changes in Frequencies

The genotype frequencies in a population can then be expressed as:

Table 6

			Origin	
	Original Frequencies	Change due to inbreeding	Independent	Identical
AA	p_0^2	$+ P_0q_0F$	$= p_O^2(1-F)$	$+ p_0F$
Aa	$2p_0q_0$	$- 2p_0q_0F$	$- 2p_0q_0(1-F)$	
Aa	q_0^2	$+ p_0q_0F$	$= q_O^2(1-F)$	$+ q_0F$

The algebra summarizes what is expected to happen “asymptotically”. In any given breeding population the results will vary due to sampling.

References

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Pearson, Burnt Mill, England.

Acknowledgements

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Quantitative Genetics Gene Frequencies Author: William Beavis, and Kendall Lamkey (ISU)

Multimedia Developers: Gretchen Anderson, Todd Hartnell, and Andy Rohrback (ISU)

How to cite this module: Beavis, W. and K. Lamkey. 2016. Gene Frequencies. *In* Quantitative Genetics, interactive e-learning courseware. Plant Breeding E-Learning in Africa. Retrieved from **https://pbea.agron.iastate.edu**.