HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2015 June 01.

Published in final edited form as:

Methods Mol Biol. 2012 ; 850: 1–9. doi:10.1007/978-1-61779-555-8_1.

Genetic Terminology

Robert Elston[1], **Jaya Satagopan**[2], and **Shuying Sun**[3]

[1] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

[2] Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

[3] Department of Epidemiology and Biostatistics, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA

Summary

Common terms used in genetics with multiple meanings are explained and the terminology used in subsequent chapters is defined. Statistical Human Genetics has existed as a discipline for over a century, and during that time the meanings of many of the terms used have evolved, largely driven by molecular discoveries, to the point that molecular and statistical geneticists often have difficulty understanding each other. It is therefore imperative, now that so much of molecular genetics is becoming an *in silico* and statistical science, that we have well-defined, common terminology.

Keywords

Gene; allele; locus; site; genotype; phenotype; dominant; recessive; codominant; additive; phenoset; diallelic; multiallelic; polyallelic; monomorphic; monoallelic; polymorphism; mutation; complex trait; multifactorial; polygenic; monogenic; mixed model; transmission probability; transition probability; epistasis; interaction; pleiotropy; quantitative trait locus; probit; logit; penetrance; transformation; scale of measurement; identity by descent; identity in state; Haplotype; phase; multilocus genotype; allelic association; linkage disequilibrium; gametic phase disequilibrium

In this introductory chapter we give the original meanings of various genetic terms (which will be found in the older literature), together with some of the various meanings that are sometimes ascribed to them today, and how the terms will be defined in the following chapters. For simplicity, exceptions are ignored and what is stated is usually, but not invariably, true.

Gene, allele, locus, site

The concept of a gene (the word itself was introduced by Bateson) is due to Mendel, who used the German word “Factor”. Mendel used the word in the same way that we might call

Corresponding Author: Robert C. Elston, PhD, Professor, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Bldg., Rm 1304, 10900 Euclid Ave, Cleveland, OH 44106, robert.elston@cwru.edu, (216) 368-4312.

"hot" and "cold" factors, not in the way that we call "temperature" a factor. In other words, his Factor was the level of what statisticians now call a factor. In the original terminology, still used by some population geneticists, genes occur in pairs on homologous chromosomes. In this terminology the four blood groups A, B, O and AB (defined in terms of agglutination reactions) are determined by three (allelic) genes: A, B and O. Nowadays molecular geneticists do not call these three factors genes, but rather "alleles", defined as "alternative forms" of a gene that can occur at the same locus, or place, in the genome. Whereas Drosophila geneticists used to talk of two loci for a gene, and human geneticists used to talk of two genes at a locus, modern geneticists talk of "two alleles of a gene" or "two alleles at a locus"; this last, which is nowadays so common, is the terminology that will thus be used in this book. It then follows (rather awkwardly) that two alleles at the same locus are allelic to each other, whereas two alleles that are at different loci are non-allelic to each other. A gene is commonly defined as a DNA sequence that has a function, meaning a class of similar DNA sequences all involved in the same particular molecular function, such as the formation of the ABO red cell antigens. (Note the common illogical use of the phrase "cloning genes" by molecular geneticists when, by their own terminology, "cloning *alleles*" is meant). Some restrict the word gene to protein-coding genes, but there are many more sequences of DNA that have function by virtue of being transcribed to RNA without ever being translated to DNA and protein-coding, so this restricted definition of a gene would appear to be unwarranted. See (1) for a more detailed explanation of the evolution of, and a modern definition of, the word "gene".

A locus is the location on the genome of a gene, such as the "ABO gene". By any definition a gene must involve more than one nucleotide base pair. Single nucleotide polymorphisms (SNPs) thus do not occur *at* loci, but rather *in* and *around* loci, and in this book we shall not write SNP markers as being "at" loci. Because of the confusion that occurs when SNPs are described as occurring at loci, some use the term "gene-locus", but we shall always use the term locus for the location of a functional gene. We shall, however, allow SNP markers to have alleles and use the original term for their locations: "sites" within loci or, more generally, sites within the region of a locus or anywhere in the genome.

If in the population only one allele occurs at a site or locus, we shall say that it is monomorphic, or monoallelic, in that population. If two alleles occur, as is common for SNPs, we shall use the original term diallelic which, apart from having precedence, is etymologically sounder than the now commonly used term biallelic. If many alleles occur, we shall describe the polymorphism as polyallelic or multiallelic (the former term is arguably more logical, the latter more common). When there are just two alleles at a locus, the one with the smaller population frequency is called the minor allele. In genetics, the term allele "frequency"--which is strictly speaking a count--is used to mean relative frequency, i.e. the proportion of all such alleles at that locus among the members of a population; thus the term minor allele frequency is often used for diallelic markers.

Genotype, phenotype, dominant, recessive, codominant, additive

An individual's genotype is the totality of that individual's hereditary material, whereas an individual's phenotype is the individual's appearance. However, the terms genotype and

phenotype are usually used in reference to a particular locus or set of loci, and to a particular trait or set of traits. Genotypes are not observed directly, but rather inferred from particular phenotypes. Thus, with respect to the ABO locus, the four blood types A, B, O and AB are (discrete) phenotypes; and the possible genotypes, formed by pairs of alleles, are AA, AO, BB, BO, BB and OO. With respect to the ABO blood group phenotypes, the A allele is dominant to the O allele and the O allele is recessive to the A allele. Similarly, the B allele is dominant to the O allele and the O allele is recessive to the B allele. The A and B alleles are codominant. Note that for the words "dominant" and "recessive" to have any meaning, at least two alleles and two phenotypes must be specified. If a particular allele at a locus is dominant with respect to the presence of a disease, there must be at least one other allele at that locus that is recessive with respect to absence of that disease.

Geneticists loosely talk about a disease being dominant, meaning that, with respect to the phenotype "disease", the underlying disease allele is dominant, i.e. the disease is present when either a single or two copies of the allele is present. Similarly, they talk of a disease being recessive, meaning that, with respect to the same phenotype, the underlying disease allele is recessive, i.e. the disease is present only when two disease alleles are present. Alternatively, they may talk of an allele being dominant or recessive, the particular phenotype (often disease) being understood. The important thing to realize is that "dominance" and "recessivity" describe a relationship between one or more genotypes and a particular phenotype. This leads to the concept of phenosets: the genotypes AA and AO form the phenoset corresponding to the A blood type, and the genotypes BB and BO form the phenoset corresponding to the B blood type. In the case of the ABO blood group, a person who has one A allele and one B allele has the blood type AB, which is a different phenotype from that of either of the corresponding homozygotes, AA and BB; this relationship is called codominance. In general, a locus is codominant with respect to the set of phenotypes it controls if the phenotypes of each heterozygote at that locus differ from that of each of the corresponding homozygotes. We make a distinction between codominant and additive; the latter implies that the phenotype (or phenotypic distribution, see below under quantitative phenotypes) corresponding to the heterozygote is in some sense half-way between those of the two corresponding homozygotes. Whereas the term additive is only meaningful when a scale of measurement has been defined, codominance is a more general concept that does not require the definition of a scale of measurement.

Polymorphism, mutation

The A, B, O and AB blood types comprise a polymorphism, in the sense that they are alternative phenotypes that commonly occur in the population. A polymorphic locus was originally defined operationally as a polymorphism-determining locus at which the least common allele occurs with a "frequency" of at least 1% (2); but a more appropriate definition would be a locus at which the most common allele occurs with a "frequency" of at most 99%. Different alleles arise at a locus as a result of mutation, or sudden change in the genetic material. Mutation is a relatively rare event, caused for example by an error in replication. Thus all alleles are by origin mutant alleles, and a genetic polymorphism was conceived of as a locus at which the frequency of the least common allele has a frequency too large to be maintained in the population solely by recurrent mutation. However, what is

important at a locus is the degree of polymorphism, and a locus in which there are 1,000 equifrequent alleles would be considered much more polymorphic than a locus at which there are two alleles with frequencies 0.01 and 0.99. Many authors now use the term mutation for any rare allele, and the term polymorphism for any common allele. We shall avoid this usage here.

Complex trait, multifactorial, polygenic, monogenic

The term “complex trait” was introduced about two decades ago without a clear definition. It appears to be used for traits that do not exhibit clear one-locus (“Mendelian”) segregation, usually because segregation at more than one locus is involved. Whereas multifactorial and complex are ill-defined and often used interchangeably, a clear distinction should be made between multifactorial and polygenic.

Multifactorial implies that more than one factor is involved in the etiology of the phenotype, whether genetic, environmental, or both. Polygenic, on the other hand, implies that only genetic factors are involved, usually in an additive fashion, with the original definition that the number of factors (loci) is so large that they cannot be individually characterized. Thus, strictly speaking, the term polygenic should not be used to include any environmental factors--though in practice it often is used that way.

Monogenic inheritance implies segregation at a single locus, and the term “mixed model” is used by geneticists to denote an additive combination of monogenic and polygenic inheritance. When both components are present in a segregation model in which both components are *latent* variables (the former discrete and the latter continuous), the underlying *statistical* model is *random*, not mixed, because there are two random components other than any error term. Statistical geneticists often use the term “transmission probability” in two quite different senses. In this book we carefully distinguish *transmission probabilities*, probabilities that a parent having a particular genotype transmits particular alleles to offspring, from *transition probabilities*, probabilities that offspring receive particular genotypes from their parents. This distinction was introduced in (3).

Haplotype, phase, multilocus genotype

Let A,B be two alleles at one locus, and D,d be two alleles at another locus. If one parent transmits A and D to an offspring, while the other transmits B and d, the offspring genotype is denoted AD/Bd (or Bd/AD), in which the parental origins are separated by ‘/’.

The two alleles transmitted by one parent constitute a two-locus haplotype; with respect to two alleles at each of two loci the are four possible haplotypes--AD, Ad, BD and Bd in this case, with AD/Bd and Bd/AD being the two possible phases. If n_1 alleles can occur at one of the loci and n_2 at the other, n_1n_2 two-locus haplotypes are possible. At the first locus $n_1(n_1 + 1)/2$ genotypes are possible (n_1 homozygotes and $n_1(n_1 – 1)/2$ heterozygotes), while at the second locus $n_2(n_2 + 1)/2$ genotypes are possible. If we pair these genotypes, one from each locus, the total number of pairs possible is

$$[n_1(n_1+1)/2][n_2(n_2+1)/2] = n_1n_2(n_1n_2+1)/2 = [n_1(n_1-1)/2][n_2(n_2-1)/2].$$

On the other hand, at the two loci together, there are n_1n_2 haplotypes; and pairing these we have $n_1n_2(n_1n_2 + 1)/2$ possible pairs of two-locus haplotypes, or diplotypes. In this book we shall define “two-locus genotypes” this way, i.e. without differentiating the two phases, so that for the same number of alleles at each locus there is a smaller number of two-locus genotypes than there are of two-locus diplotypes. Thus we shall consider the two phases of the double heterozygote, Ad/BD and BD/Ad, as being the same two-locus genotype. Usually, the term “multilocus genotype” refers to genotypes when the phases are not distinguished, and the term diplotype is useful for the case when they are distinguished (though this term is not yet in common usage).

More generally, a haplotype is the multilocus analogue of an allele at a single locus. It consists of one allele from each of multiple loci that are transmitted together from a parent to an offspring. When haplotypes made up of multiple alleles (one from each locus) are paired, a pair in which the genotype at each of n loci is heterozygous corresponds to 2n–1 different diplotypes, or phases. It is usual nowadays to restrict the word haplotype to the case where all the loci involved are on the same chromosome pair, so that all the alleles involved are on the same chromosome. Typically, but not always, it is assumed that all the different phases of a particular multiple heterozygote have the same phenotype.

Epistasis, interaction, pleiotropy

When two loci are segregating, each typically influences a separate phenotype. For example, A and B may be alleles at the ABO locus, determining the ABO blood types, while D and d are alleles at a disease locus, determining disease status. But if alleles at a single locus influence two different phenotypes, we say there is pleiotropy. It is known that a person's ABO genotype influences the risk of gastric cancer as well as determining blood type. Thus the ABO locus is pleiotropic. Alternatively, alleles at two different loci may determine the same phenotype, such as presence or absence of a disease; and if the phenotype associated with the genotypes at one locus depends on the genotypes at another locus, we say there is epistasis. Thus gastric cancer may perhaps be caused by the epistatic effect of alleles at two (or more) loci. Epistasis and pleiotropy are sometimes confused in statistical genetics.

Allelic association, linkage disequilibrium, gametic phase disequilibrium

If the alleles at one locus are not distributed in the population independently of the alleles at another locus, the two loci exhibit allelic association. If this association is a result of a mixture of subpopulations (such as ethnicities or religious groups) within each of which there is random mating, the association is often denoted as “spurious”. In such a case there is true association, but the cause is not of primary genetic interest. If the association is not due to this kind of population structure, it is either due to linkage disequilibrium (LD) or gametic phase disequilibrium (GPD); in the former case the loci are linked, i.e. they co-segregate in families, in the latter case they are not linked, i.e. they segregate independently in families.

Owing to an unintended original definition, loci that are not linked have often been mistakenly described as being in LD (4, 5).

Identity

The concept of allelic identity is an important one. Alleles are identical by descent (IBD) if they are copies of the same ancestral allele, and must be differentiated from alleles that are physically identical but not (at least within the previous dozen or so generations) ancestrally identical. Such alleles, when not IBD, are identical in state (IIS). It is well understood that molecules, atoms, etc., can be in different states (not “by” different states), and the same is true of alleles, though here the states are ancestrally, not physically, different. Whereas in the animal and plant genetics literature the phrases “identity in state” and “identical in state” are commonly used, for no good reason the phrases “identity by state” and “identical by state” are now commonly used in the human genetics literature. In this book, to stress the difference and to be consistent with both the earlier common usage and the usage in the animal and plant genetics literature, we shall use the terminology IIS, not IBS.

Quantitative traits

A locus at which alleles determine the level of a quantitative phenotype is called a QTL (quantitative trait locus). Typically, the word “quantitative” is used interchangeably with “continuous” when describing a phenotype. However, quantitative traits can be discrete. Care should be taken to distinguish between those methods of analysis of quantitative traits for which distributional assumptions, such as conditional normality, are critical, and those for which they are not. Transforming the phenotype of a QTL corresponds to changing its units if the transformation is linear, or more generally to changing the scale of measurement (e.g. square root or logarithmic) if the transformation is non-linear. On the scale of measurement used, alleles at a QTL have an additive effect if the phenotypic distribution of the heterozygote is the average of the corresponding two homozygote phenotypic distributions. With respect to that phenotype, allele A is dominant to the allele B, and allele B is recessive to allele A, if the whole phenotypic distribution of the heterozygote AB is the same as that of the homozygote AA. Any variance among the phenotypic means of the genotypes at a locus over and above that due to additive allele action is called dominance, or dominant genetic, variance. Thus dominance variance can arise as a result of one allele being dominant to another, but such simple allele action is not necessarily implied by dominance variance. The presence of dominance variance depends on the scale of measurement; dominant allele action (complete dominance, as decribed above for discrete traits such as the ABO blood group) does not. If the phenotypic distribution of a heterozygote is not the average of the corresponding two homozygote phenotypic distributions, we shall say there is codominance. Thus in this book we shall not restrict the word codominance to the case of additivity (with the result that codominance is scale independent).

Just as dominance has a different meaning when applied to quantitative traits, so does epistasis. From a statistical point of view, dominance can be considered as *intralocus* interaction, or non-additivty of the allelic contributions to the phenotype. Epistasis is a

genetic term, now generalized when applied to quantitative traits to indicate non-additivity of the effects on the phenotype of the genotypes at two (or more) loci in a population. It is thus from a statistical viewpoint *interlocus* interaction, and so dependent on how the phenotype is measured. Statistical interaction is a term with a similar limitation, but is not restricted to genetic factors. Statistical interaction should be carefully distinguished from biological interaction (5,6). Whereas biological interaction does not require the presence of statistical interaction, the presence of the latter implies the existence of the former. Indeed, statistical interaction is removable if a monotonic transformation can make the effects of the two factors involved (e.g. segregation at two loci, or segregation at one locus and levels of an environmental factor) additive. Furthermore, the magnitude of any interaction effects can depend critically on how the individual factor effects (single locus genotypes in the case of genetic factors) are defined (5).

There is usually no loss of generality in assuming that disease status, unaffected or affected, is a quantitative trait that takes on the values 0 or 1, respectively, so that its mean value is the population prevalence of the disease. Then everything that has been written here with regard to dominant allele action, dominance variance and epistasis also holds in the case of a binary disease phenotype, except that now the scale of measurement (in the sense of a non-linear monotonic transformation) is irrelevant in the absence of a quantitative measure. However, if there is a quantitative measure, such as a relative risk or odds ratio, then the scale of measurement will determine whether or not there is interaction. Also, in the case of a binary disease phenotype the penetrance, or probability of being affected, is often transformed to a probit (or logit), giving rise to what is called the “liability” to disease, and this liability is treated as a continuous phenotype. Dominance and epistatic variance can be quite different on this liability scale from that measured on the original “penetrance” scale.

For a QTL, dominance variance is present when there is intralocus non-additivity. By the same token, epistatic variance is present when there is interlocus non-additivity. Each locus gives rise to its own components of additive genetic and dominant genetic variance. If multiple loci affect a QTL, there are multiple components of epistatic variance. Except in the case of a binary phenotype with no associated quantitative measure, the relative magnitudes of all such components are scale (i.e. transformation) dependent, just as corresponding components of genotype (or allele) x environment interaction are scale dependent.

Finally, for those who wish to have a better theoretical understanding of statistical human genetics, reference (7) provides an exceptionally good introduction.

References

1. Gerstein MB, et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 2007; 17:669–681. [PubMed: 17567988]
2. Ford, EB. Polymorphism and taxonomy.. In: Huxley, J., editor. The new systematics. Oxford: 1940.
3. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Hum Hered. 1971; 21:523–542. [PubMed: 5149961]
4. Lewontin RC. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics. 1964; 49:49–67. [PubMed: 17248194]

5. Wang X, Elston RC, Zhu X. The meaning of interaction. Hum Hered. 2010; 70:269–277. [PubMed: 21150212]
6. Wang X, Elston R, Zhu X. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? Nat Rev Genet. 2010 doi:10.1038/nrg2579-c2.
7. Ziegler, A.; König, IR. A statistical spproach to senetic spidemiology: Concepts and applications. 2nd edn.. Wiley-VCH; Weinheim: 2010.