STATISTICAL GENETICS

Associating Genotypic Differences with Measurable Outcomes

RONGLING WU

University of Florida

GEORGE CASELLA

University of Florida

During World War I, the French army dispatched ammunition-carrying Briard dogs to the front lines. The Red Cross turned to the shaggy-haired sheepdogs to haul first-aid supplies. And now Briards once again have been called to service in medical research. For the first time ever, animals that were born blind gained the ability to see after undergoing gene therapy, opening the door to the development of treatment for people with a rare, inherited eye disorder.[1]

Gene therapy is a treatment designed to correct defective genes responsible for disease or illness. There are several approaches currently under study, but the most common (and the one used to reverse blindness in the Briard dogs) involves introducing a "normal" gene to replace the abnormal gene that causes the disease. This is a

[1]Based on research from the University of Florida, Cornell University, and the University of Pennsy[illegible] published in the May 2001 issue (vol. 28, no. 1) of *Nature Genetics*.

complicated process, and its success depends on knowing which gene (or combination of genes) is responsible for the disease. Identifying the offending gene can be difficult—one statistician, Mark Yang of the University of Florida, described finding a particular gene as similar to finding a particular one-yard long segment along a 170-mile-long highway. Fortunately, with the aid of statistical analyses, it is now possible to attempt to locate genes (see the essay "How Many Genes? Mapping Mouse Traits" by Melanie Bahlo and Terry Speed, included in this book, for a description of this process).

The success of the gene therapy experiment involving the Briard dogs represented a huge advance for gene therapy research. Briards are susceptible to a genetic disease that causes blindness. Because researchers were able to identify the gene responsible for the disease (a defect in the RPE65 gene), they were able to devise a treatment and test it on Lancelot, a four-month-old Briard puppy who had been blind from birth. Normal RPE65 genes were injected into the subretinal areas of Lancelot's right eye. About three months later, vision was detected in his right eye, and after 10 months, Lancelot could see well enough with his good eye to negotiate a crowded room without bumping into anything. The success of this experimental therapy has medical researchers optimistic about the prospect of developing a treatment for a similar disorder, Leber's congenital anaurosis, which causes blindness in humans, and for which currently there is no cure.

Conditions or diseases that are a consequence of defects in a single gene, as is the case for blindness in Briard dogs, are the best candidates for gene therapy. Unfortunately, some common conditions, such as heart disease, arthritis, obesity, and infection rate, are complex genetic traits. Such traits are controlled by many genes, and each one may have only a minor effect on the trait. Moreover, these traits are often sensitive to both genetic and environmental variation, so it is difficult to isolate the genetic influence. Hence, there is a need for statistical analysis to help make valid conclusions from the data.

This success of gene therapy illustrates the type of applications we see today in the science of genetics. Like many other branches of the biological sciences, genetics has been energized by recent advances in *genomics* (the study of the molecular structure and function of genes). The results of genetic research have started to be felt in every aspect of our daily lives. For example, genetically modified

crops provide us with higher yielding and more nutritional grains, and personalized drugs, designed specifically to correspond to a patient's genetic makeup, are more effective in curbing diseases. Genetics and genomics are even joining in the fight against bioterrorism.

To draw valid inference from genetic data, we strive to understand the genetic architecture of a complex trait, such as the number of genes, gene effects, and interactions between different genes or between genes and the environment. The *phenotypic variation* of any trait (the variation in the observed outcome, such as growth, amount of infection, etc.) can be partitioned into its genetic and environmental components through statistical analysis.

Today, we are most interested in finding the connections between the variability that we see in the phenotype (the observed outcome) and the individual genetic components, referred to as *quantitative trait loci* (QTL). (A *locus,* or point, leads us to a place on the *genome,* the string of genes, and identifies a particular gene that is influencing the trait.) This work has been greatly aided by two different developments. First, advancements in molecular technologies in the mid-1980s led to the generation of a virtually unlimited number of *markers* (loci on the genome that the scientist can identify exactly) that specify the genome structure and organization of any organism. Second, and almost simultaneously, improved statistical and computational techniques were developed, which gave us the computational power needed to tackle statistically complex genetic problems.

MENDELIAN GENETICS

Arbitrarily pick out two individuals from a population. You will find considerable differences between these two individuals in various traits such as body size and form, eye color, and behavior. Some of the things causing these differences are *genes* (the physical entities transmitted from parent to offspring that determine the inheritance of a trait). The set of genes present in an individual constitute its *genotype,* and the physical expression of the genotype is called the *phenotype.* Genes can exist in different forms or states, called *alleles.*

Genes are lined up in a *chromosome,* and the location of the gene along a chromosome is the *locus* of the gene. (You may recall the famous work of James Watson and Francis Crick, who showed that the

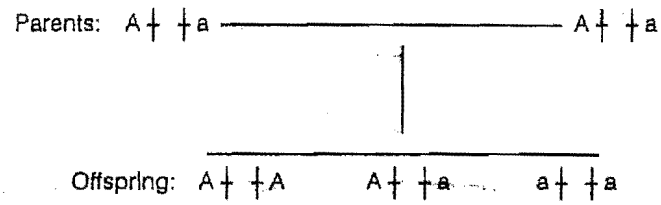

FIGURE 1 Possible offspring from heterozygous parents

chromosomes in an organism are arranged in a double helix.) If we have two alleles, symbolized by *A* and *a*, at a gene locus, then different combinations of these alleles in a *diploid* (pair) form three possible genotypes, including two homozygotes (*AA*, *aa*) and one heterozygote (*Aa*), each corresponding to a particular phenotypic value for a trait.

Gregor Mendel was the first man to postulate that genotypes could influence phenotypes. He observed an almost perfect 3:1 segregation ratio for all seven characteristics that he measured in hybrid plants derived from a series of crosses for the garden pea. For example, he found that 5474 seeds were round and 1850 seeds were wrinkled, and $5474/1850 = 2.96$. These data allowed Mendel to hypothesize that an offspring obtains one allele randomly from each parent.

Mendel's hypothesis became *Mendel's first law*—a gene will segregate during the formation of the reproductive cells (*meiosis*), thus passing into different *gametes* (reproductive cells). Figure 1 illustrates Mendel's first law, where crossing two heterozygous parents, *Aa*, generates $\frac{1}{4}$ homozygote *AA*, $\frac{1}{2}$ heterozygote *Aa*, and $\frac{1}{4}$ homozygote *aa*. If allele *A* dominates allele *a*, we will have $\frac{3}{4}$ dominant phenotype (a mix of *AA* and *Aa*) and $\frac{1}{4}$ recessive phenotype, the 3:1 ratio (see Question 1).

Based on his pea hybrid study, Mendel developed a second law, which says that if there are two or more pairs of genes on different chromosomes, they segregate independently (combine independently in the offspring). However, in his study of fruit flies in the 1930s, Thomas Morgan found that genes that were close together on the same chromosomal region (two linked genes) will not segregate independently. Morgan's finding laid a solid foundation for modern linkage analysis and the construction of genetic maps from molecular markers.

LINKAGE ANALYSIS

The ultimate goal of complex trait dissection is to identify the actual genes involved in the trait and to understand the cellular roles and functions of these genes. As a first step, a linkage analysis seeks to associate marker genotypes with measurable traits.

Molecular markers are segments of the genome that exhibit heritable variation, that is, the alleles that can be determined. We try to detect the location of a QTL (recall that a QTL is a locus on the genome that identifies a particular gene that is influencing a trait) by finding markers that are close to the QTL and are associated with the trait of interest. To do this we measure the phenotypic trait of members of a randomly segregating population and also determine their molecular genotype. We then try to determine if an association exists between any of the markers and the quantitative trait. Markers that are close in location to the QLT should have traits that are associated with the trait being studied.

A common method of determining the association is by analyzing phenotypic and genotypic data with standard statistical methods such as a one-way analysis of variance or regression analysis. If the genotype class is found to be significant, then the molecular marker used to define the genotype class is considered to be associated with a QTL. This idea is later illustrated in Figure 2, which displays loci that are associated with differential growth in poplar trees.

As a simple example, suppose that we have data on the heights of 12 tomato plants (in cm)—79, 82, 85, 87, 100, 101, 102, 103, 124, 125, 126, 127—and we want to see if we can associate height with a particular gene. At each marker we can group the above data according to the genotype of the marker for each particular plant. A grouping at one marker might be giving some evidence that genotype *AA* is associated with increased height.

	GENOTYPE		
	aa	Aa	AA
Data	79, 82, 85, 87	100, 101, 102, 103	124, 125, 126, 127
Mean	83.25	101.5	125.5

On the other hand, at another marker the data might group as follows:

	GENOTYPE		
	bb	Bb	BB
Data	79, 87, 102, 125	82, 100, 103, 126	85, 101, 124, 127
Mean	98.25	102.75	109.25

And here we have less evidence that tomato plant height is associated with this gene.

Remember that the genotypes that we use are those of the marker, for we do not know the location of the QTL. The accuracy and precision of locating QTL depends, in part, on how close the markers are to one another, and of course it is better to have the markers close together. In the next section we will look more closely at how we can link the markers to the QTL.

QTL: KEYS TO AGRICULTURAL PRODUCTIVITY

Most of the traits important to agriculture are quantitatively inherited. In other words, they are shaped by a number of different genes that interact with each other and the environment. Yield, flavor, and stress adaptation, including drought and salinity tolerance, are examples of economically important traits that can be called quantitative.

In agriculture, a breeder will cross two parents and practice selection until advanced-generation lines with the best phenotype for the quantitative trait under selection are identified. This type of program, though, requires a large input of labor, land, and money. Therefore plant breeders are interested in identifying the most promising lines as early as possible in the selection process. Another way to state this point is that the breeder would like to identify as early as possible those lines that contain those QTL alleles that contribute to a high value of the trait under selection.

The molecular dissection of complex traits into individual QTLs needs two steps: (1) the construction of a genetic linkage map (a map that tells us the ordering of the genes) using *polymorphic markers* (which have more than one allele in a population), and (2) the identification of QTLs for the traits on the genetic map.

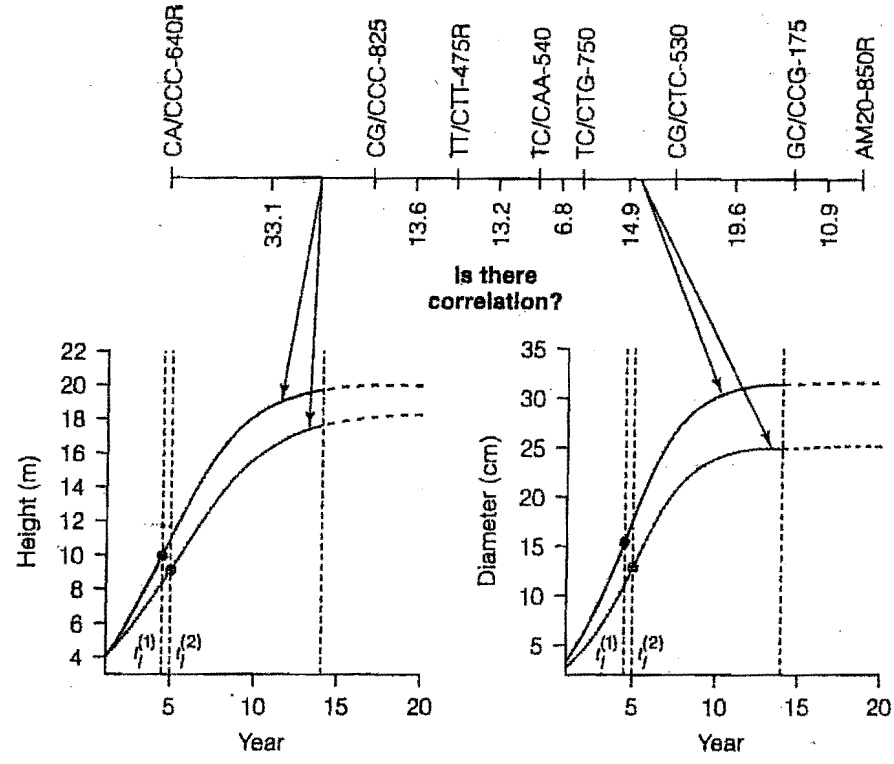

FIGURE 2 Possible locations of QTL for height and diameter in poplar trees

In the first step, all individuals in a controlled population are genotyped for different polymorphic markers and measured for a complex trait of interest. The distances between each pair of the markers are calculated, and a genetic linkage map composed of these markers is constructed using statistical algorithms. Of course, the higher the density of the markers on the map, the more precise the location of the potential QTL.

The top panel in Figure 2 depicts a linkage map. There we see the markers arranged along a line (the linkage group). The distance on the line between two markers is representative of the distance between the two markers on the chromosome.

In the second step, the markers are systematically associated with the phenotypic values of the trait, the QTL underlying the trait are identified on the map, and the number of the QTL and their genetic effects are estimated.

The bottom panel of Figure 2 shows two quantitative traits (growth trajectories for height and diameter) in poplar trees. The two

QTLs exert increased effects on stemwood growth when trees develop, but the QTL effects will be constant after age 10 years or so. The goal of the analysis is to find loci on the linkage map (top panel) for which different genotypes results in differential growth trajectories, therefore allowing the breeder to ultimately control the growth of the poplar tree by specifying the QTL genotype. (Ma, Casella, and Wu [2002] describe such an analysis in detail.)

GENETICS AND OBESITY

Obesity is a major health problem in the United States and most industrialized civilizations. It is estimated that 50% to 60% of Americans over age 30 are overweight and 25% to 30% are clinically obese. More worrisome is obesity in children, where rates in developed nations worldwide have risen two- to fourfold in the last 10 to 25 years. Obesity is a significant medical concern because numerous studies link obesity with increased risk of cardiovascular disease, metabolic disorders (such as type 2 diabetes mellitus and lipid abnormalities), and some forms of cancer. Obesity is thought to be a complex disease, influenced by numerous genes, along with other demographic and environmental factors like diet, age, physical activity, sex, and ethnicity.

It is now estimated that 40% to 70% of the variability in body weight is related to genetic factors, and obesity represents a classic example of the negative interaction between genetics and modern lifestyles. The ability to store energy when food was available represented a selective advantage for times when food was scarce. Only in recent times, when there is an overabundance of food and when many people are not engaged in physical labor, has obesity begun to rise and represent a significant health problem. It is also now clear that obesity is *polygenic* (influenced by multiple genes) in nature, and it is likely that populations in different geographic areas were under different environmental pressures, resulting in different sets of genes that contribute to fat/energy storing efficiency. Given the considerable rise in obesity rates worldwide, there is substantial interest in attempting to unravel its genetic basis.

Current approaches for genetic studies of obesity in humans broadly include (1) twin studies used to estimate the heritabilities of

fat mass and other obesity-related traits, (2) animal models characterizing spontaneous single-gene mutations or candidate genes to understand obesity in humans, (3) gene association studies with obese cases and non-obese controls, and (4) a genomewide scan for obesity traits using linkage analysis, aimed at characterizing candidate QTL carrying human obesity genes.

Some polymorphic sites (QTL) predisposing to human obesity were identified using data from 643 women drawn from the Women's Ischemia Syndrome Evaluation (WISE) study (see Wu et al., forthcoming b). This study found that different QTLs are responsible for two different aspects of obesity—the amount and the distribution of body fat. It appears that the obesity QTLs are population dependent. For example, more QTLs for obesity were detected in the black population than white population, and there is no common QTL between these two populations. Such a population difference suggests using different strategies to prevent or reduce obesity in different populations.

GENETIC DIFFERENCES IN HIV DYNAMICS

The way in which the human immunodeficiency virus type 1 (HIV-1) particles that cause AIDS change over time varies considerably from patient to patient, and it is possible that such variation in HIV-1 behavior is related to an individual's genetic makeup. Traditional genetic analysis of HIV-1 infection has been based on various biochemical approaches, but it has not yet had great success. This is because HIV-1 dynamics, as a complex trait, is under polygenic control (influenced by more than one gene) and, moreover, seems sensitive to environmental changes. Moreover, at this time, there is a lack of data measuring both the amount of virus, and the genetic markers, on a population large enough to detect any such correlation.

However, there are statistical *models* that can help us. Using such a model, we have found statistical evidence for the possible existence of a host major gene responsible for HIV dynamics. The three genotypes at the detected major gene displayed marked differentiation in their viral load trajectories (Figure 3). The heterozygote (*Aa*) and one

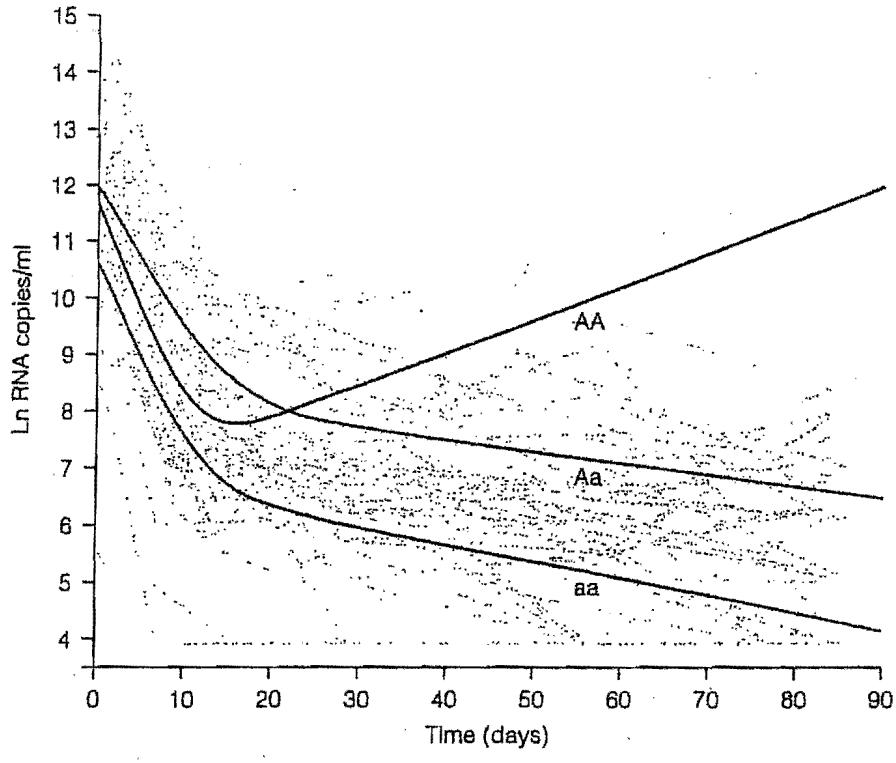

FIGURE 3 Load trajectories of HIV-1 virions (measured in viral RNA copies) for 53 patients (with shadow curves). The *x*-axis is time (in days) and the *y*-axis is a measure of the viral load of the patient. A high viral load implies that the drug therapy is not effective. The three thick curves each represent a different genotype (denoted by *AA*, *Aa*, or *aa*) at a possible major gene detected by the statistical model.

homozygote (*aa*) that together account for an overwhelming majority of patients (94%) were found to decline consistently with time in viral load after initiation of antiviral drugs. For the second homozygote (*AA*) in the frequency of 6%, viral load turns out to increase from day 20 following a short period of decline after the treatment of antiviral drugs.

For the staggering number of people infected with HIV (40 million), it is imperative to develop the most potent drugs to quickly eliminate all HIV from the blood and from the body. The possibility of a major gene for HIV dynamics suggests that HIV infection may be more efficiently prevented and curbed using personalized strategies of gene therapy targeted for the gene a particular patient carries.

CONCLUSIONS

The past 20 years have witnessed great progress and promise in genetics and genomics, which shapes our daily lives. The ability to gather data at the genome level, combined with statistical algorithms and computing power, has opened new doors in the analysis of biological systems. The use of this information, both as a diagnostic tool and a prescriptive tool, has the promise of greatly increasing the power of modern medicine. It can lead us to not only more accurate diagnoses of disease but also to the use of personalized drug therapies that will provide ever more effective treatments, helping us to cure diseases that, as of now, remain without cure.

REFERENCES

Fisher, R. A. (1918). "The Correlation between Relatives on the Supposition of Mendelian Inheritance." *Transaction of the Royal Society of Edinburgh* 52:399–433.

Ma, C. X., G. Casella, and R. L. Wu. (2002). "Functional Mapping of Quantitative Trait Loci Underlying the Character Process: A Theoretical Framework." *Genetics* 161, no. 4:1751–62.

Wu, R. L., et al. (Forthcoming a). "A Major Gene Detected for HIV-1 Dynamics." *Journal of Theoretical Biology.*

———. (Forthcoming b). "Genetic Variants for Human Obesity as Revealed by Linkage Disequilibrium Mapping." *Pharmacogenetics.*

ADDITIONAL READING

The Story of DNA and the Double Helix

Watson, J. D., and L. Bragg. (1991). *The Double Helix: A Personal Account of the Discovery of the Structure of DNA,* reissue ed. New York: New American Library.

Statistical Genetics

Doerge, R. W., Z. B. Zeng, and B. S. Weir. (1997). "Statistical Issues in the Search for Genes Affecting Quantitative Traits in Experimental Populations." *Statistical Science* 12, no. 3:195–219.

Lynch, M., and B. Walsh. (1997). *Genetics and Analysis of Quantitative Traits*. Sunderland, Mass.: Sinauer Associates.

Mendelian Genetics

Weir, B. S. (1996). *Genetic Data Analysis*. Sunderland, Mass.: Sinauer Associates.

Gregor Mendel

Henig, R. M. (2001). *The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics*. New York: Mariner Books.

The educational website http://www.mendelweb.org covers the material in Mendel's 1865 paper "Experiments in Plant Hybridization."

QUESTIONS

1. Explain why the reproduction scheme portrayed in Figure 1 leads to a 3:1 ratio of dominant phenotype to recessive phenotype. Hint: To see all possible combinations of offspring, fill in the table.

PATERNAL GENOTYPE	MATERNAL GENOTYPE	
	A	a
A		
a		

2. How does the association of a marker trait with the trait of interest help to detect the location of the gene that controls the trait of interest?
3. What are the implications of Figure 3 for the effectiveness of the drug? That is, if a patient's genotype is known, how does Figure 3 advise the doctor?