

GENETIC MAPPING IN MODEL ORGANISMS

ŚAUNAK SEN*

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

INTRODUCTION

One of the important goals in biology is to find genetic elements responsible for variation in observable traits. Because of our shared evolutionary history, model organisms such as mice, and rats provide valuable clues towards the etiology of human disease. Many of these traits (such as body weight or BMI, Body Mass Index) are inherently quantitative, or have a closely related measurement that is quantitative (such as fasting glucose for diabetes). Regions of the genome responsible for variation in quantitative traits of interest are called QTL (quantitative trait loci).

In this note we will discuss the basic ideas of genetic mapping (QTL mapping) in model organisms using a hypertension mouse cross as an example. QTL identified in model organisms such as mice provide valuable clues for genes in humans. From a statistical perspective, QTL mapping in model organisms may be viewed as a simplified form of linkage mapping, as well as association mapping in humans. Many of the statistical problems and solutions are clearer in this simplified setting.

CROSSES BETWEEN INBRED LINES

After many generations of inbreeding and selection, different strains of mice have been developed. They are maintained in different research labs around the world. These strains show systematic reproducible differences in many observable traits. If two strains of mice raised in the same environment show consistent differences in a phenotype of interest (such as blood pressure), then we can be reasonably confident that there is a genetic basis to the difference. To find the genetic elements contributing to that difference, one may use crosses between the two strains. Crossing strains randomly shuffles the genomes of the two parents. Because this shuffling is random (by the rules governing meiosis), associations between genetic variation and phenotype can be thought of as being causal.

The simplest cross to describe is a backcross (Figure 1). The F_1 generation has no genetic variation and is not useful for studying the association between genetic variation and phenotype. If we cross the F_1 individuals to one of the parental strains, the resulting backcross has genetic variation from meioses in the F_1 mother(s). We can now study the association between genetic variation in the backcross generation and the phenotype of interest (say, blood pressure, if we are studying hypertension).

An idealized version of what the data might look like is shown in Figure 2. Thus the observed data is a matrix of numbers, with a column for the phenotype (or as many columns as the number of phenotypes), and a matrix of marker genotype data, some of which may be missing. The data may be missing “by chance” (if some genotype reactions did not work) or they may be missing “by design” as is the case in the hypertension cross between the normotensive A/J strain and the hypertensive BL/6 strain (Figure 3).

In this cross of 250 backcross animals, a strategy called “selective genotyping” was used (Figure 4). The idea behind selective genotyping is that extreme phenotypic individuals are more “informative” and hence it is more efficient to genotype them more intensely (a similar idea underlies case-control designs). In

* © 2006 Śaunak Sen; Last updated March 17, 2006.

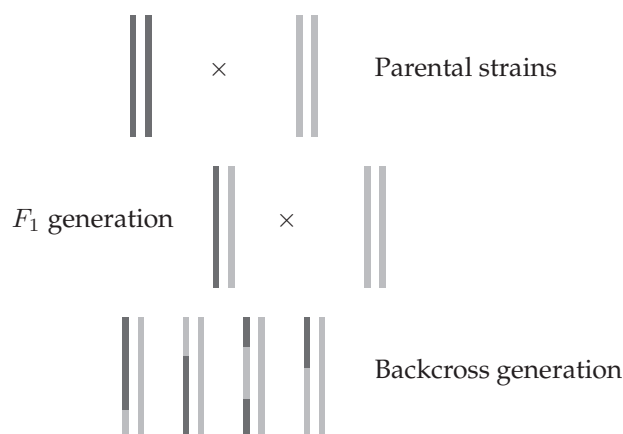


Figure 1: Schematic diagram of a backcross scheme. In the top row we show a single pair of chromosomes of two parental inbred strains (in dark and light grey). When we cross them we get the heterozygous F_1 generation, but everyone in this generation is still genetically identical. When we cross this generation to one of the parental generations, the recombinations in the maternal line create genetic variation in the backcross generation.

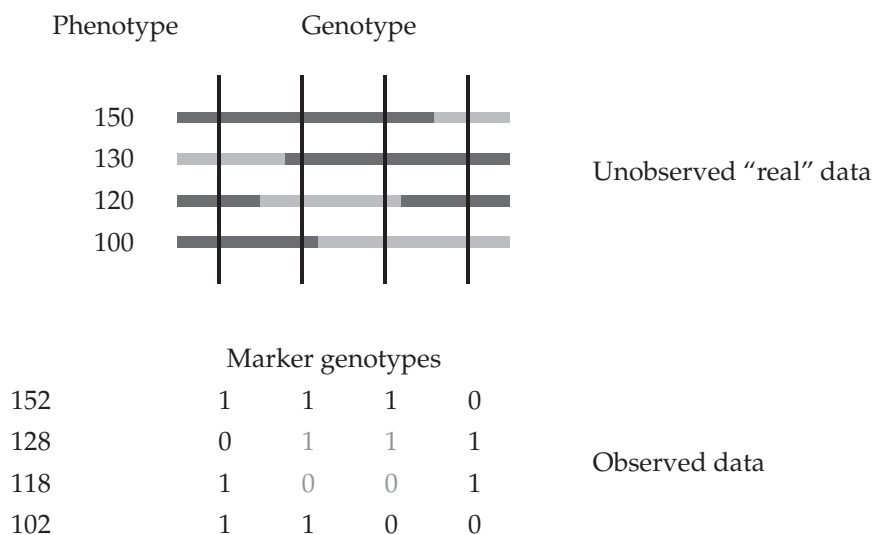


Figure 2: Figure showing the observed data, and the underlying unobserved data from a backcross experiment. The top panel shows the underlying genotype and phenotype data. Each individual has genetic contributions from both parental strains (shown in light and dark grey). The bottom panel shows the actual observed data, which is an incomplete version of the original data. The observed phenotypes may be noisier versions of the underlying "true" phenotype. The genotype of each individual is not known at every genomic location. Instead it is known at specific markers (microsatellites, or SNPs, etc). We code the markers 0 or 1 according as the genetic contribution is from the light or dark grey strain. Some of the genetic marker data may be missing. The missing genotype data may be due to random technical mishaps or by design, as in selective genotyping.

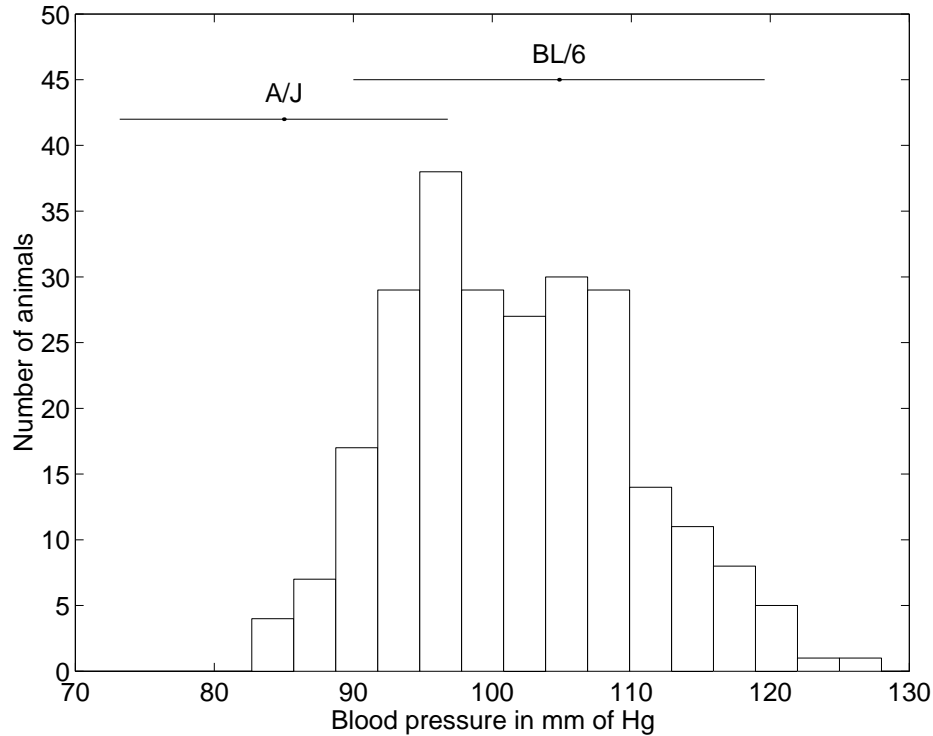


Figure 3: Histogram of systolic blood pressure from 250 backcross mice. The range of blood pressure of the parental strains (mean $\pm 2 \times$ sd) is shown above the histogram. The blood pressure ranges of the normotensive (A/J) strain and the hypertensive (BL/6) strains overlap a bit. It is noticeable that the blood pressure range of the backcross mice is not intermediate between the two parental strains. This is typical of complex traits, and may be due to epistatic effects of co-adapted gene complexes. Notice also that the systolic blood pressure of mice is not very different from humans.

the hypertension cross, the 100 mice with the most extreme blood pressure were genotyped at regularly-spaced markers throughout the genome. The intermediate mice were not genotyped, except for select chromosomal regions that were thought to be “promising” from a preliminary genome scan (see below).

Examining the genotype patterns by eye in the hypertension cross it is possible to guess where some of the QTL might be. However, such a subjective approach is impractical for routine analysis. For this reason we need formal statistical methods.

STATISTICAL ANALYSIS

The goal of the statistical analysis is to help us find the location of the QTL, and how they act. There are two major challenges. First, there may be a lot of missing data (as with the hypertension data). The missing data is primarily incomplete genotype information (but phenotypes may be missing also). The statistical analysis of the data has to explicitly take into account missing data. The most common missing data methods used are the EM (Expectation Maximization) algorithm, and multiple imputation. Second, even if the genotype data is complete, i.e. each individual’s genetic composition is known without ambiguity, we have to determine how many QTL there are. This is a problem of model selection (similar to selecting the number of variables to adjust for in a multivariable regression analysis) for which no fully satisfactory solution has been found. Below, we will illustrate how we currently analyze such data.

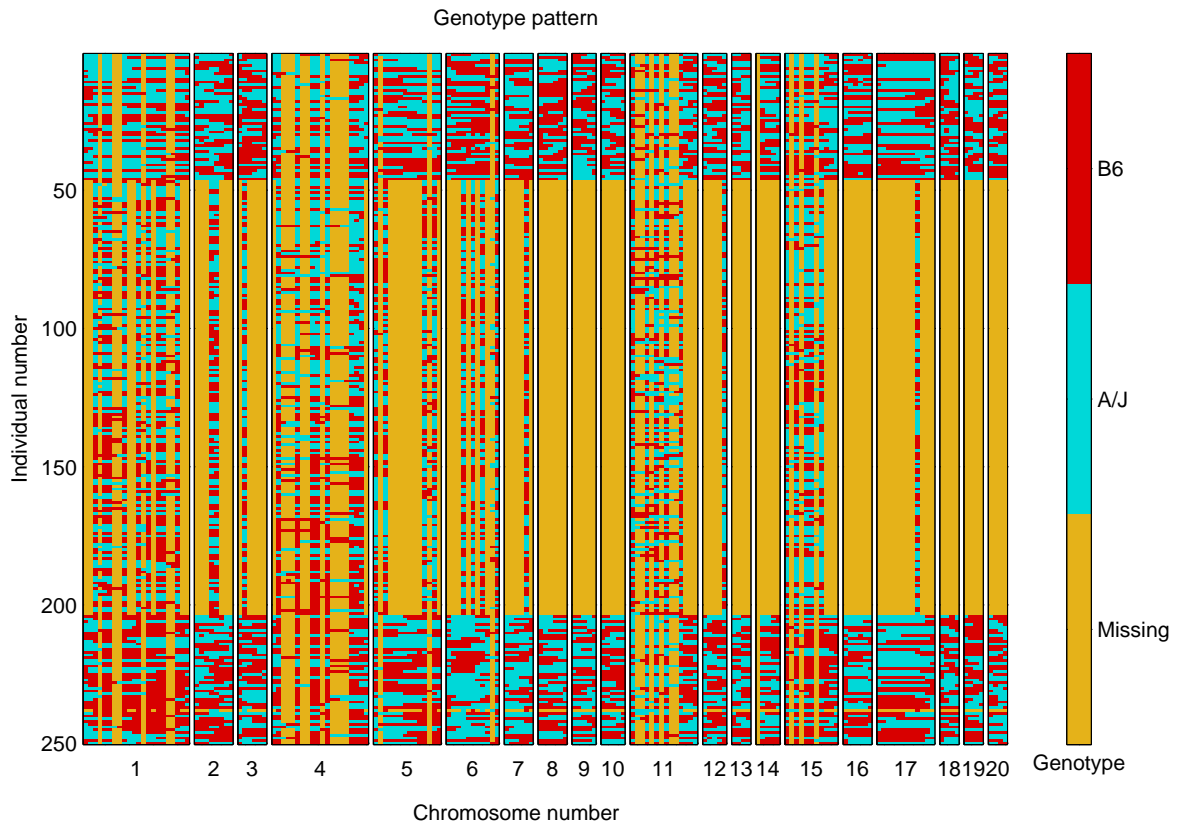


Figure 4: Genotype pattern of hypertension backcross. Each row in this figure corresponds to a mouse. The mice (rows) are ordered by blood pressure, with the mouse with the lowest blood pressure in the first row, and the mouse with the highest blood pressure at the bottom. Each column corresponds to a typed marker. The markers have been ordered by their position on the genome. Genotypes are coded red if they come from the hypertensive (BL/6) strain, blue if they are from the normotensive (A/J) strain, and yellow ochre if they are missing. A few patterns are immediately apparent. Half of the genotype data is missing (but half of the information is not). The missing data pattern is systematic. The extreme phenotypic individuals are more heavily genotyped than the intermediate ones. Some chromosomes are more heavily genotyped (these are the ones that were deemed promising from an initial genome scan). Some markers are more heavily genotyped than others (more on this in the next figure). Looking at the genotype data, we can also see a preponderance of red genotypes on chromosome 1 for the most hypertensive mice, and the opposite pattern for the least hypertensive mice. This suggests that there may be a locus contributing to blood pressure variation on chromosome 1. A similar pattern is also seen for chromosome 4. A reverse pattern is seen on chromosome 6 (more blue, or normotensive strain) genotypes for the hypertensive mice. This also suggests a locus contributing to blood pressure on chromosome 6, but one that acts opposite to the expected direction – genetic contributions from the normotensive strain appear to raise blood pressure.

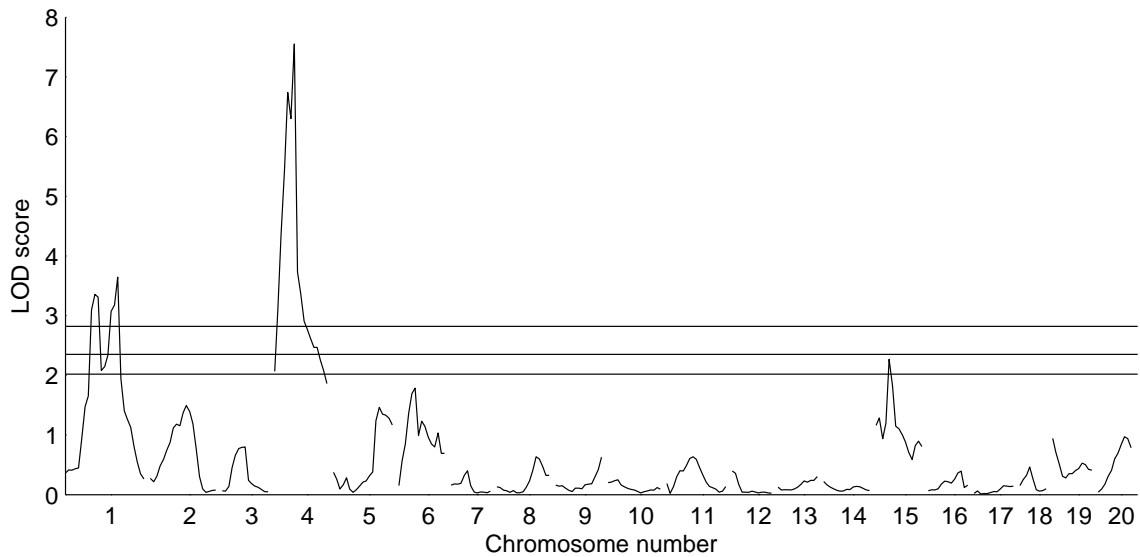


Figure 5: One-dimensional genome scan of hypertension data. The figure plots the LOD score as a function of genome position. The three horizontal lines are genome-wide LOD thresholds obtained by permutation testing (1000 permutations). The top line is the 1% threshold, followed by the 5% threshold, and the 10% threshold at the bottom. This figure indicates QTL on chromosomes 1, 4, and possibly, 15.

Missing data methods

The primary goal of missing data methods in QTL statistical analysis is to properly account for the missing genotype data. Note that even if all markers are completely genotyped, we have missing data. This is because we will still not know the genotype in between typed markers unambiguously unless the markers are very dense (say 1cM apart).

Ignoring the missing data can lead to biased, and inefficient results. For the hypertension data, almost every mouse has some missing genotype data. Thus deleting mice with incomplete genotype data is not an option. The less extreme option, throwing away data from the intermediate blood pressure mice, will lead to exaggerated estimates of the blood pressure effects at typed markers.

Likelihood-based methods, such as the EM algorithm, and multiple imputation, are able to provide unbiased results if an important assumption holds: the missing data pattern must depend on the observed data (and not on any missing data). For the hypertension data, where selective genotyping was used, we must include the phenotypes of all mice and all observed genotyped in constructing the likelihood. This is because the missing data pattern depended on the phenotypes, and observed genotypes. This assumption will not hold if we delete the phenotypes of intermediate individuals. In that case the missing data pattern would depend on the unobserved phenotypes of the intermediate mice.

One-dimensional genome scan

A one-dimensional genome scan (also known as genome scan) is a survey of all single-locus models. We walk through each location on the genome and ask: is this locus associated with the phenotype of interest? At each locus, we perform a likelihood ratio test for association between genetic variation at that locus, and the phenotype. The likelihood ratio tests are expressed in base ten logarithms (by historical tradition in genetics). Figure 5 shows the one-dimensional genome scan for the hypertension data.

Two-dimensional genome scan

A two-dimensional genome scan is a survey of all two-locus models explaining the phenotype. The advantage of this method is that it enables us to examine possibly linked and epistatic loci. Such loci are difficult (or impossible) to detect using one-dimensional genome scans. However, two-dimensional scans are time-consuming, and have a higher multiple comparisons burden. Developing reasonable model selection methods is an active research area.

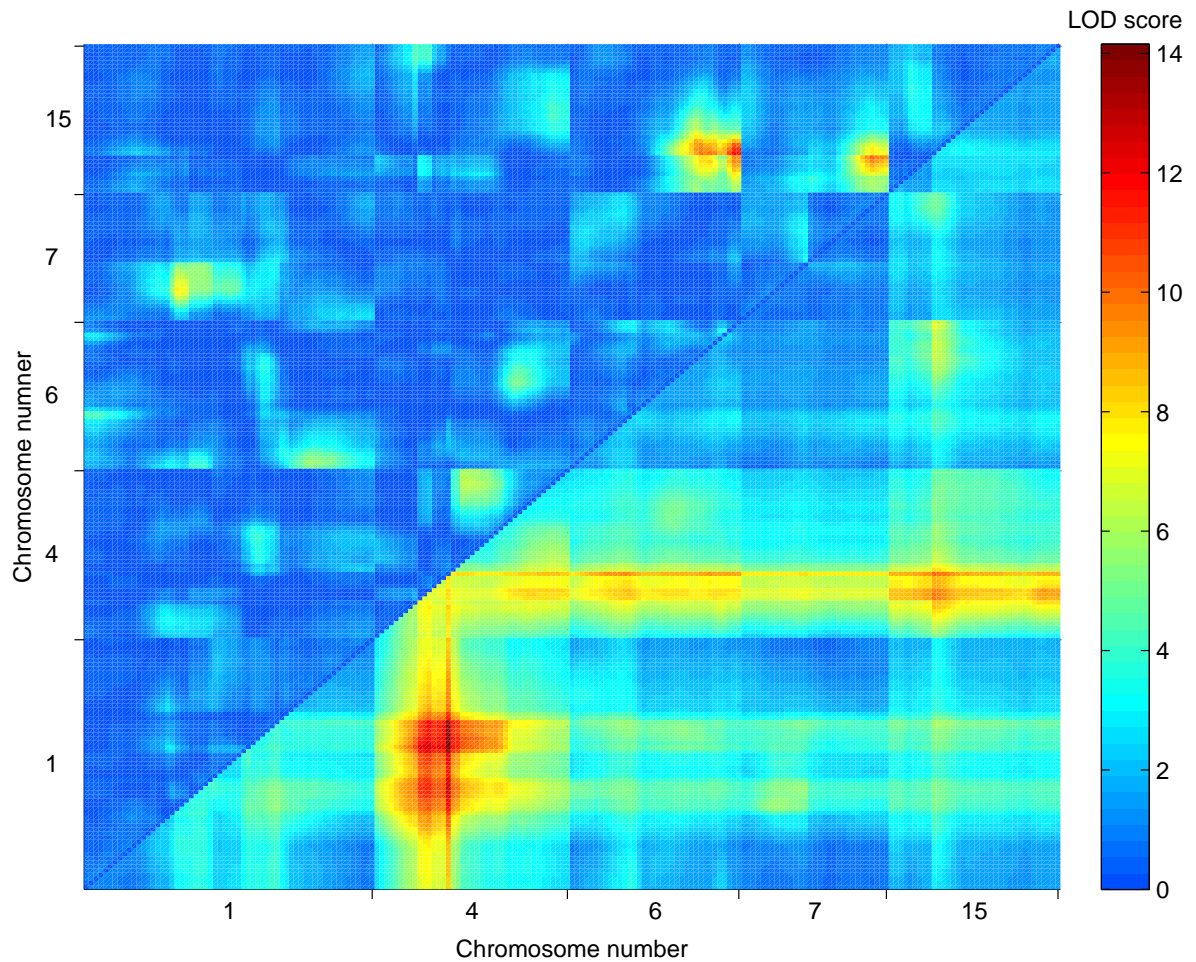


Figure 6: Two-dimensional genome scan of hypertension data for selected chromosomes (1, 4, 6, 7, and 15). The lower triangle shows the LOD score for a two-locus model including an interaction term for each pair of loci in the selected chromosomes. The upper triangle shows the LOD score for the interaction term alone (inflated by a factor of 3 for visual purposes) for each pair of loci in the selected chromosomes. The two-dimensional scan shows that chromosomes 1 and 4 have major additive loci. There is evidence that loci on chromosomes 6 and 15 are acting epistatically (two-locus interactions). There is a hint of two linked loci on chromosome 1, and epistasis between loci on chromosomes 7 and 15.