# *In Silico* Study of Transcriptome Genetic Variation in Outbred Populations

**Miguel Pérez-Enciso**[1]

*Universitat Autònoma de Barcelona, Facultat de Veterinária, Departament de Ciència Animal i dels Aliments, 08193 Bellaterra, Spain and Station d'Amélioration Génétique des Animaux, Institut National de la Recherche Agronomique, 31326 Castanet Tolosan, France*

Manuscript received July 9, 2003
Accepted for publication September 24, 2003

ABSTRACT

Dissecting the genetic architecture of regulatory elements on a genome-wide basis is now technically feasible. The potential medical and genetical implications of this kind of experiment being very large, it is paramount to assess the reliability and repeatability of the results. This is especially relevant in outbred populations, such as humans, where the genetic architecture is necessarily more complex than in crosses between inbred lines. Here we simulated a chromosome-wide SNP association study using real human microarray data. Our model predicted, as observed, a highly significant clustering of quantitative trait loci (QTL) for gene expression. Importantly, the estimates of QTL positions were often unstable, and a decrease in the number of individuals of 16% resulted in a loss of power of ~30% and a large shift in the position estimate in ~30–40% of the remaining significant QTL. We also found that the analysis of two repeated measures of the same mRNA can also result in two QTL that are located far apart. The intrinsic difficulties of analyzing outbred populations should not be underestimated. We anticipate that (many) conflicting results may be collected in the future if whole-genome association studies for mRNA levels are carried out in outbred populations.

THE goal of identifying the genetic polymorphisms that affect each of the mRNA levels of an organism, *i.e.,* the merging of genetics and genomics, is currently a most exciting and promising avenue of research (Cheung and Spielman 2002). Thus far, most of the few available studies have been done in crosses between inbred lines, in yeast (Brem *et al.* 2002), Drosophila (Wayne and McIntyre 2002), mice (Schadt *et al.* 2003), or corn (Schadt *et al.* 2003). This strategy has been called a multifactorial perturbation of a biological system (Jansen 2003) because many combinations of the original founder allele are simultaneously produced (in an $F_2$, recombinant inbred line, or similar line). However, assessing the genetic architecture of the transcriptome in outbred populations, like humans, is badly needed if pharmacogenomics is to succeed. Here the challenge is to quantify and position in the genome the most important sources of variation, rather than further "perturbing" the system because this is already very complex.

Certainly, outbred populations pose additional challenges to the genetic dissection of complex traits; *e.g.*, the number of quantitative trait loci (QTL) alleles and their frequencies are unknown and the phases between QTL and marker alleles change from family to family. Many of these drawbacks can be avoided with a very dense genotyping using single-nucleotide polymorphisms (SNPs), thereby increasing the chances of identifying the causal mutation or mutations in an association study. Interestingly, evidence that genetic variation for transcription activity does exist within as well as between populations begins to appear, much the same as for any other quantitative trait (Oleksiak *et al.* 2002; Cheung *et al.* 2003; Whitney *et al.* 2003). Thus, given current technology, it is tempting to believe that it is only a matter of time (and money) before we can identify most, if not all, polymorphisms that affect genetic variation at the transcription regulation level, the so-called eQTL (Schadt *et al.* 2003).

Nevertheless, we are still far away from knowing the genetic architecture of the transcriptome in outbred populations. Moreover, we are almost completely ignorant about the potential power of experiments, the accuracy of, *e.g.*, estimated locations of QTL, how widespread is the presence of "hotspots" of QTL for transcription activity, and so on. Such information would be extremely valuable before tackling very relevant and timely but also expensive studies.

The main objective of this work is to assess the credibility of these transcription activity hotspots and to study the stability of the location estimates for the eQTL. "Ghost" hotspots may result simply from a high correlation between mRNA levels. Computer simulation is, traditionally, an extremely useful tool for exploring the behavior of genetic systems but here we are confronted with new problems; in particular, how do we simulate transcriptome data? Instead of simulating microarray data, a simulated population of SNP haplotypes is applied to published microarray data for two independent

[1]*Address for correspondence:* Universitat Autònoma de Barcelona, Facultat de Veterinária, Departament de Ciència Animal i dels Aliments, 08193 Bellaterra, Spain. E-mail: miguel.perez@uab.es

studies of gene expression in humans. In all simulations, the SNP genotypes are randomly assigned to an individual's microarray, so that there is no underlying functional relationship between SNPs and expression. Results show patterns of clustering of eQTL that resemble those published in actual studies. Moreover, we also retrieve a high instability in the estimated locations for the eQTL.

## METHODS

The data sets used were those by Rosenwald *et al.* (2002) and Whitney *et al.* (2003), corresponding to large B-diffuse cells from lymphoma patients [Rosenwald (R) set] and whole blood from healthy individuals [Whitney (W) set], respectively. These data sets were chosen because, among those publicly available, they were of moderate to large size: $n = 240$ (R) and $n = 76$ (W) and from outbred populations. The R data set consisted of 240 patients from several countries that suffered from diffuse B-cell lymphoma. The so-called "lymphochip" was used, which is a cDNA microarray constructed from a germinal B-cell library containing in total ~12,000 clones and is expected to contain most of the genes expressed in cancerous lymphoma tissues. Full details are given elsewhere (Alizadeh *et al.* 2000; Rosenwald *et al.* 2002); here we used data from 7399 cDNA measurements that were taken directly from their web site (http://llmpp.nih.gov/DLBCL/). The W set was extracted using the SMD system (http://genome-www.stanford.edu/normalblood). For a given cDNA (sample) to be selected, it should have at least 80% valid measurements across samples (cDNAs), which are the default values suggested by the authors; 3441 cDNA levels from 76 individuals were retained. The W set was composed of 69 healthy individuals from the United States plus 7 Nepal individuals. The log 2 of the mean ratios between the test sample and a control sample was analyzed in both data sets. In all cases, a single microarray per individual was used. Both R and W sets provided repeated measurements of some clones. There were 7399 (3441) spots measured but only 4503 (2925) distinct mRNAs in the R (W) sets, according to the annotation provided. Some cDNAs were not annotated so it might well be that the degree of redundancy is higher than reported.

For a comparison, we also reanalyzed yeast data (Brem *et al.* 2002) that consist of 40 haploid segregants of a cross between a lab (BY) and a wild (RM) strain. We preselected those mRNA levels for which at least 30 segregants had data (75%), resulting in 5714 mRNA levels analyzed with 2% remaining missing data; a total of 3312 SNPs distributed across the whole genome were available and ~2% of genotypes were missing; 17 of the significant clones reported (Brem *et al.* 2002) were not used here because they did not meet our criterion of <25% missing data.

In theory, the usual coalescence approaches (Hudson 2002) can be used to simulate long stretches of DNA sequence with many opportunities for recombination, but the memory and CPU requirements soon increase exponentially and this approach was not practical. Thus we simulated $2n$ chromosomes using a mixture of coalescence and gene-dropping methods. First we simulated 3000 chromosomes with exactly 100,000 polymorphic sites uniformly distributed and parameter $\rho = 4N_e r = 1000$ (Hudson 1993, 2002), where $N_e$ is the population effective size and $r$ is the recombination fraction between the ends of the segment simulated. Hudson's (2002) programs with a fixed number, 100,000, of segregating sites and $\rho = 1000$ were used.

This initial sample of 3000 chromosomes contained an extreme degree of disequilibrium and thus was not realistic at all. Then, we generated offspring from these chromosomes using a simple gene-dropping strategy for $t = 1000$ discrete generations with random mating and without selection; in each generation, 3000 new chromosomes were generated, which were produced from meioses between two of the former generation chromosomes. A total chromosome length of 100 cM was assumed. We also considered $t = 200$, 400 but the results relevant to this study (*e.g.*, change in power, stability of estimates and so on) were very similar and thus are not presented. As a result of drift in the gene-dropping procedure and of the absence of new mutations, the final number of segregating SNPs was ~25,000 in all replicates. For the sake of QTL analyses (described below), we used only those SNPs with minor allele frequency >0.10, resulting in ~20,000 SNPs finally used.

Two distinct chromosomes from the final generation were assigned at random to each individual set of cDNA measurements. Five replicates were run, and the same set of chromosomes was used with either the R or the W sets. The results were very similar across replicates so the strategy was considered stable. The search for QTL was done using maximum likelihood (ML) as follows. The likelihood that the $k$th locus (SNP) has an effect on the expression level of the $j$th cDNA clone is given by, assuming normality of the mRNA abundance log ratio,

$$L_k(\mathbf{y}) = \prod_{i=1}^{n} \phi(y_{ij} - \mu_j - a\alpha_{ik} - d\delta_{ik}, \sigma_j^2), \qquad (1)$$

where $y_{ij}$ is the log 2 measure of mRNA abundance of the $i$th individual for the $j$th clone; $\mu_j$ is the general mean for the $j$th clone; $a$ and $d$ are the additive and dominant effects, respectively; $\alpha_{ik}$ is an indicator variable taking values −1, 0, 1 if the $ik$th SNP genotype is 00, 01, and 11, respectively; whereas $\delta_{ik}$ is 1 if the individual is heterozygous, 0 otherwise; $\sigma_j^2$ is the residual variance; and $\phi(x, \sigma^2)$ stands for the normal density function of mean $x$ and variance $\sigma^2$. ML estimates of $\mu$, $a$, $d$, and $\sigma^2$ were obtained at each locus, $k = 1, 2, \ldots, n$ loci, and the position showing the maximum-likelihood ratio

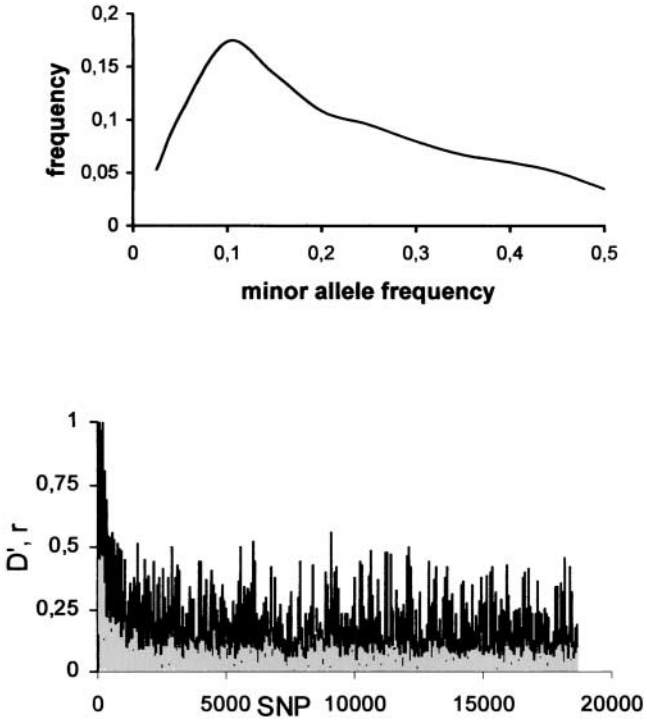


FIGURE 1.—(Top) Spectrum of the SNP allele frequencies in the simulated chromosomes. (Bottom) Disequilibrium profile between the first SNP and the rest, $D'$ (solid lines) and $r$ (shaded lines), in one of the replicates.

(LR) over the null model was taken to be the ML position estimate. The null model included only the general mean ($\mu_j$).We deemed that the eQTL was significant if the $P$ value was $<0.5 \times 10^{-6}$.

Clustering was assessed by dividing the chromosome in bins, each containing 50 consecutive SNPs, and by counting the number of significant eQTL in each bin. The significance of clustering was assessed by simulation: 10,000 replicates were run where each of the significant eQTL had the same chance of being assigned to any of the bins.

## RESULTS AND DISCUSSION

**Simulation strategy:** So far, there is no computationally feasible strategy to simulate polymorphism and linkage disequilibrium patterns over a whole chromosome. Here we have used a mixed strategy that combines coalescence techniques and gene dropping. Coalescence methods allow us to produce a set of chromosomes with the desired degree of polymorphism (HUDSON 1993), albeit with a very limited recombination rate and thus with extreme linkage disequilibrium. Subsequently, gene dropping allows for recombination and generates a realistic linkage disequilibrium decay, although it is computationally far more expensive than coalescence; we also assumed no new mutations during the gene-dropping process.

Figure 1 shows the spectrum of allele frequencies found in the last generation. The mean minor allele frequency was 0.18 with an SD = 0.14; compared to some real data there is an underrepresentation of very low frequencies $<0.05$: compare Figure 1 here to Figure 1 in PHILLIPS *et al.* (2003), but the rest of the histogram is very similar. Standard measures of disequilibrium, $D'$ and $r = \sqrt{r^2}$, were computed (WEIR 1996). The mean $D'$ was 0.91 within the 10 closest SNPs, 0.60 between those 100 SNPs away, and 0.49 at 1000-SNP distance, but, as observed in real data (REICH *et al.* 2001), there was also a wide variability in linkage disequilibrium (LD) decay. As an example, Figure 1 shows the $D'$ and $r$ profiles between the first SNP and the rest; as is known, $D'$ was much more variable and with a higher mean value than $r$. Overall, and given that very low-frequency polymorphisms are unlikely to contribute much to genetic variance, we deem that the simulation strategy resulted in a rather realistic polymorphism and linkage-disequilibrium picture. To avoid instability of the estimates as much as possible we further restricted the SNPs analyzed to those with frequency $>0.10$. The reader, of course, should be aware that some of the conclusions drawn from this work could be modified according to the true polymorphism pattern. Nevertheless, and importantly, main results (clustering, variability in QTL positions, *cf.* below) were robust to several values of the number of generations $t$ in the gene-dropping procedure and to the fact of using all or a restricted subset of SNPs.

**QTL *hotspots*:** A prediction of our model, verified empirically in several studies (CARON *et al.* 2001; BREM *et al.* 2002; SCHADT *et al.* 2003), was that gene-expression QTL are clustered, *i.e.*, show the presence of *hotspots* containing many more significant eQTL than are expected by chance in nearby positions (Figure 2). Note that individual plots in the same row were obtained with identical chromosomes but applied to different data sets (R or W). Averaged over replicates, the first 5 most populated bins contained 20% of all significant QTL, and 30% of all QTL were contained within the 10 largest bins in the R set. The respective values in the W data were 18 and 28% of all QTL. We showed by simulation that these figures depart significantly from the null hypothesis of random QTL locations ($P \ll 10^{-4}$). It should be noted that the degree of clustering found here is larger than that reported with real data in mice (SCHADT *et al.* 2003) and comparable to that reported in yeast if we exclude the largest bin in BREM *et al.* (2002).

Certainly, clustering in our model is not random but it is also not caused by a putative mutation with a regulatory effect, as the assignment of genotypes to phenotypes was done at random. It rather stems from the complex interplay of correlation between different mRNA levels together with linkage disequilibrium patterns. The correlation between different mRNA levels is caused in all likelihood by an underlying common regulation of gene expression. It is striking that hotspots

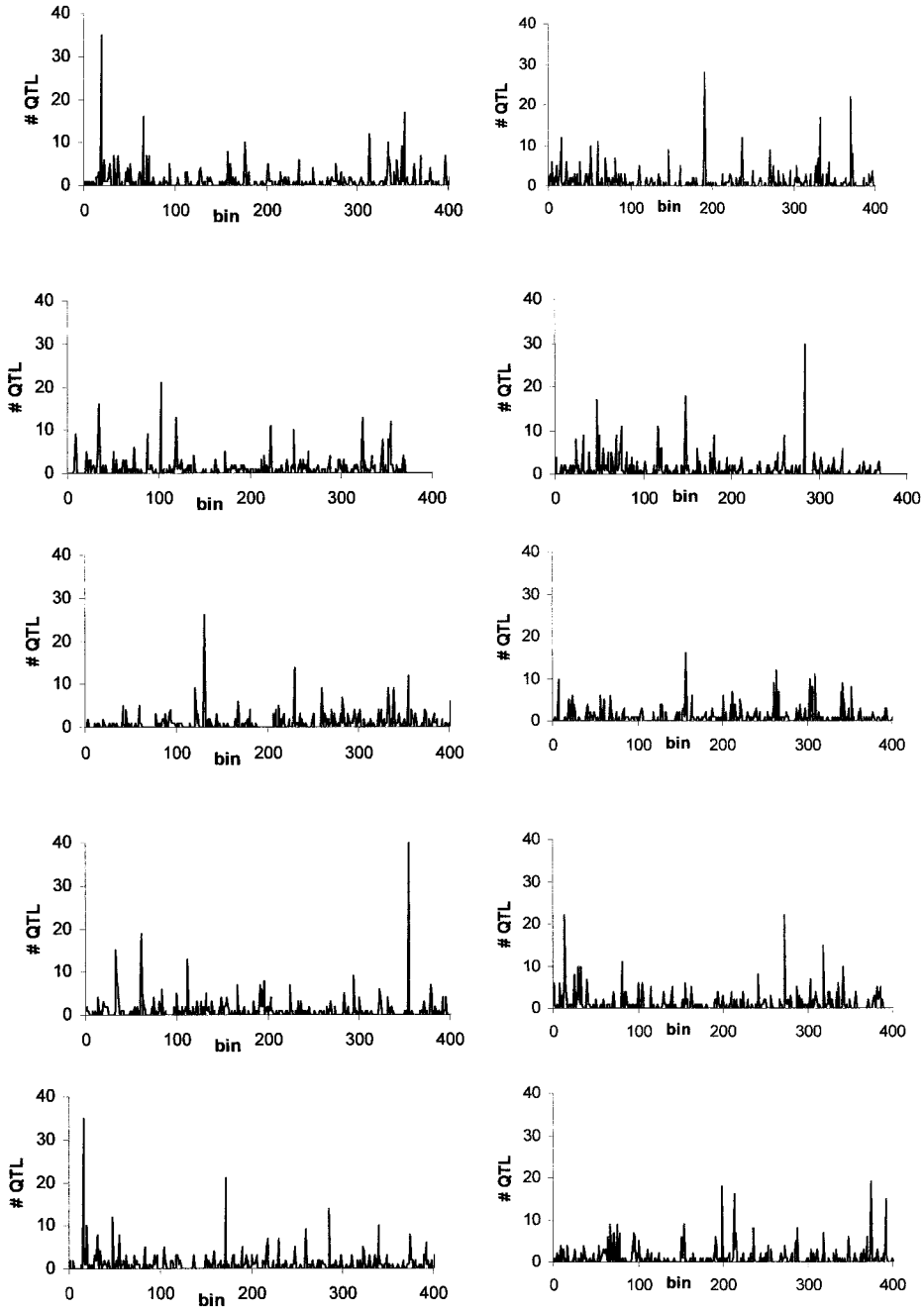


FIGURE 2.—Plots showing that significant QTL are clustered in the model used here. The five different replicates are shown in different plots; left plots correspond to the R set and right plots to the W set. The position is in abscissas and the number of eQTL is in ordinates. Within each row, the same chromosome population was used. The chromosome was split into bins containing 50 SNPs each. If there were repeated measurements for the same gene, only one was chosen to avoid spurious clustering. About 370 clones were significant in each replicate.

were detected with two completely independent data sets across different genotypes samples, suggesting that common regulation may be an important source of correlation between many different mRNA levels. It has been pointed out (DARVASI 2003) that the presence of such eQTL hostpots is a very interesting finding that may explain why expression levels of genes in a given pathway are correlated. Yet, experimental results should be interpreted with caution, as no "functional" relation between genes is required to retrieve a highly structured distribution of eQTL along the genome: compare Figure 2 here with Figure 1a in SCHADT *et al.* (2003) or with Figure 3 in BREM *et al.* (2002).

**Stability of estimates:** A traditional matter of concern in functional genomic studies has been the repeatability of the results. Classical phenotypic measurements like growth or disease resistance are supposed to be measured with little error, although this can be debated. In contrast, some mRNA levels can vary dramatically within seconds in the cell. There are also several steps in mRNA quantification, like retro-transcription into cDNA, which can change the original mRNA abundance ratios; thus, mRNA quantification is potentially exposed to numerous sources of experimental error and it has been advocated strongly (CHURCHILL 2002) that mRNA measures should be replicated. The effect of random noise on the estimates can be assessed by comparing the results obtained with each of the repeated measures of the

**TABLE 1**

**Relation between correlation and position estimates with repeated measurements**

| | Whitney *et al.* | | Rosenwald *et al.* | |
|---|---|---|---|---|
| δ[a] | %[b] | $\bar{\rho}$[c] | % | $\bar{\rho}$ |
| 0 | 40.7 | 0.84 | 40.2 | 0.88 |
| 1–100 | 10.5 | 0.83 | 5.4 | 0.74 |
| 101–1,000 | 2.3 | 0.70 | 3.4 | 0.64 |
| 1,001–10,000 | 23.2 | 0.54 | 33.2 | 0.67 |
| 10,001–20,000 | 23.2 | 0.72 | 17.8 | 0.68 |

All pairs were considered when more than two clones per expressed sequence tag were available, and all replicates were analyzed jointly.

[a] Difference (SNP positions) between the two QTL position estimates.

[b] Percentage of pairs falling at the given δ distance.

[c] Average correlation between all repeated measurements.

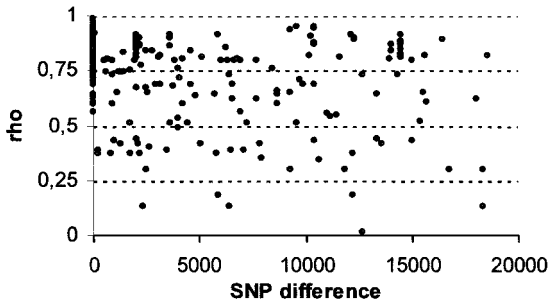


FIGURE 3.—Distance between the eQTL for replicates of the same mRNA as a function of their correlation (rho). R data set and average of all replicates are shown.

same mRNA. We compared the estimates of repeated measures for the same gene: only ~40% of the positions for pairs of repeated measurements were identical in both data sets (Table 1). Logically, it can be expected that the higher the correlation between repeated measurements, the closer the two position estimates will be. However, the observed pattern was not so straightforward (Table 1). The correlation was maximum between mRNA pairs that had coincident positions, but there was not a clear trend showing less correlation for more distant position estimates; some pairs had distant position estimates, yet their correlation was high (Figure 3).

The reliability of potential eQTL studies in outbred populations can be studied if we assume that the estimates obtained with the original data set are the "true" values of the parameters; next we introduce some stochasticity in the system by deleting information and compare the results between the true values and the "estimated" values (those obtained with partial data). We did this in two steps: first, we deleted 16% of individuals at random and second, we additionally removed 90% of the SNPs such that the total number of polymorphisms available was ~2000, uniformly distributed along the chromosome. Provided that marker density is very high, the increase in power by adding more individuals (−Ind sets) is modest: ~70% of true QTL would have been detected (Table 2). However, if marker density is "only" ~2000 SNPs (−Ind and −SNP sets), only ~52% (W set) and 40% (R set) of true eQTL will be identified. But the most dramatic effect was on the estimates of the QTL positions, especially in the smallest data set (W). For instance, when we had 2000 SNPs and 64 individuals, 26% of position estimates were within the 100 closest markers to the true position. If we increased the number of SNPs nine times, this percentage increased to 56%, and 52% of all significant QTL coincided with the true position. For larger populations (R set) the effect of increasing the number of SNPs was not so important in position accuracy, but rather in increasing power. Increasing the number of SNPs from 2000 to 20,000 in the R set increased the percentage of significant genes whose position was estimated accurately ($\delta < 100$), from 60 to 77 (a 28% increase); yet, it did increase power from 40 to 72% (67% increase).

We suspected that there could be an effect of SNP allele frequency on the variability of estimates: perhaps SNPs at extreme allele frequencies would be more "unstable" than SNPs at intermediate frequencies because discarding even a small number of individuals might produce large changes in allele frequencies and thus in the respective LRs. However, allele frequency was not related in a clear manner with the difference between the *P* values obtained in the complete data set and the *P* value in the reduced data set (results not presented).

From these analyses, it seems that some eQTL estimates are potentially very unstable. At least in part, this instability is due to random errors in the measurement of mRNA levels (Table 1), but we also wanted to know whether the SNP linkage disequilibrium pattern had an effect. To this end, we reanalyzed the data from BREM *et al.* (2002) [the Brem (B) set], which consists of 40 haploid $F_2$ segregants from a cross of two yeast lines. We repeated the analysis, deleting 6 segregants (16%) and deleting additionally 90% of the 3312 available SNPs. We compared the two sets of estimates and the results are in Table 2 (bottom). Loss in power was comparable to previous results; however, the position estimates were quite stable despite the relatively small data set. Next, we examined the effect of analyzing B data with a simulated set of chromosomes used in the R data (the BR set), and we found almost identical results as with the R and W sets (Table 2). All in all, it seems that the complex pattern of linkage disequilibrium in outbred populations, together with the diversity of allele frequencies, makes eQTL estimates rather unstable and harder to interpret than those in crosses between inbred lines.

Of course, it should be evident that all these results are "false" positives because there is no connection what-

**TABLE 2**

**Adding noise to the system**

| | Data set[a] | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | W−Ind | | W−Ind−SNP | | R−Ind | | R−Ind−SNP | |
| | All | $P < 10^{-6}$ | All | $P < 10^{-6}$ | All | $P < 10^{-6}$ | All | $P < 10^{-6}$ |
| %[b] | — | 69 | — | 52 | — | 67 | — | 40 |
| $\delta = 0$[c] | 46 | 52 | 8 | 12 | 60 | 70 | 8 | 23 |
| $\delta < 10$[d] | 48 | 53 | 20 | 26 | 63 | 73 | 22 | 50 |
| $\delta < 100$[d] | 50 | 56 | 27 | 34 | 66 | 77 | 27 | 60 |
| $\delta > 10^4$[e] | 13 | 13 | 21 | 20 | 8 | 5 | 17 | 6 |

| | Data set | | | | | |
|---|---|---|---|---|---|---|
| | B−Ind | | B−Ind−SNP | | BR−Ind | |
| | All | $P < 10^{-6}$ | All | $P < 10^{-6}$ | All | $P < 10^{-7}$ |
| % | — | 72 | — | 42 | — | 73 |
| $\delta = 0$ | 79 | 76 | 9 | 17 | 60 | 58 |
| $\delta < 10$ | 94 | 94 | 55 | 70 | 61 | 59 |
| $\delta < 100$ | 99 | 100 | 80 | 99 | 65 | 64 |
| $\delta > 10^4$ | 0.4 | 0 | 18 | 0.1 | 7 | 6 |

[a] The data sets are: W, Whitney *et al.*, simulated chromosomes; R, Rosenwald *et al.*, simulated chromosomes; B, Brem *et al.* data and chromosomes; BR, data from Brem *et al.* and simulated chromosomes as in R. Suffixes are: −Ind, 16% of observations randomly removed; −SNP, 90% of markers removed (1 every 10 successive markers was saved).

[b] Percentage of QTL that were significant in the complete and reduced data sets.

[c] Percentage of QTL whose positions were the same in the complete and reduced data sets.

[d] Percentage of QTL whose position differed by $<X$ SNPs in the complete and reduced data sets.

[e] Percentage of QTL whose position differed by $>10^4$ SNPs in the complete and reduced data sets.

soever between the phenotypes and the genotypes. But the important point here is that the system behaves similarly whether there is a correlation or a causal force; *i.e.*, the statistical analyses carried out here cannot discriminate between correlation and causation. Thus, if we accept that the results obtained with all data can be assimilated to the true values, there is no reason to question that results with a partial data set cannot be considered as "estimates" and the behavior of the system will be that of a typical statistical inference scenario. Similarly, the issue of setting the relevant $P$ value has been avoided because what matters is the relative changes of the $P$ values in two analyses rather than the absolute number; *e.g.*, we are interested in the comparison between the full-data $P$ values *vs.* the partial data set $P$ values.

But why are QTL position estimates sometimes so unstable? The response lies in the shape of the LR test profiles. We took, as a manner of example, two repeated measures of the same mRNA from the R set. The correlation between measures was 0.75, $P$ values were $\sim 10^{-10}$ in both replicates, yet the maxima were located in distinct positions (Figure 4). The minimum $P$ values of the local maxima in the first replicate differ only slightly so it is foreseeable that even small changes in the data set may result in dramatic changes in the maximum-likelihood estimate of the position. There was no significant linkage disequilibrium between the SNPs that held the maxima ($D' = 0.02$ and $\sqrt{r^2} = 0.01$). The SNPs had allele frequencies of 0.16 and 0.14. Interestingly, the second profile showed only one of the two maxima that were present in the first profile. Thus, despite rather high correlations between both measurements and the sharing of one LR maxima, local discrepancies may also occur. We found that the presence of local distinct maxima was frequent in both R and W data sets, and this is the main reason for the lack of stability in the eQTL position estimates. In contrast, we found only rarely this pattern in the B set. This, again, is the result of a much simpler LD pattern and allele frequencies in crosses than in outbred populations. In light of these results (Figure 4), the marker density should be a matter of concern. If marker spacing is not sufficiently dense, the LD decaying rapidly in outbred populations (Kruglyak 1999), we will miss completely the complex pattern observed in Figure 4. This is precisely what we can see in Table 2. It is important to note that the conclusions drawn apply to the two independent data sets studied (R and W); results varied only quantitatively.

**Robustness of the method:** Gene expression levels being known for their "unpleasant" distributions, it can be argued that our results may be due to using a para-

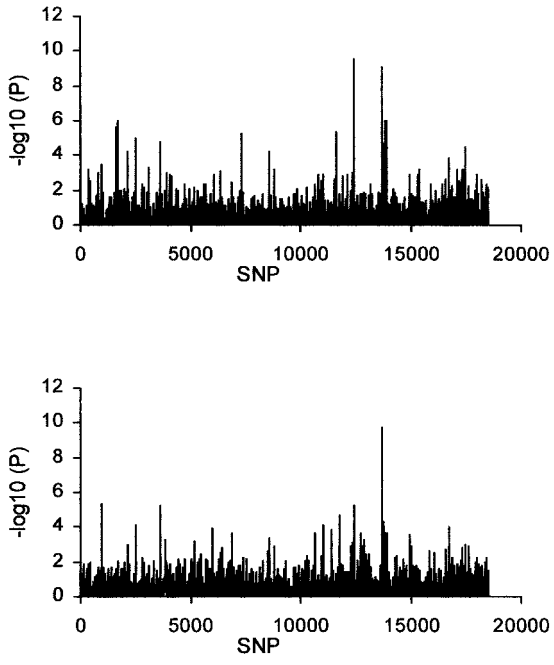


FIGURE 4.—Log 10 *P*-value profile of two repeated mRNAs (correlation = 0.75) in the R set. (Top) Profile of first replicate, two local maxima are at positions 12,443 (global) and 13,703. (Bottom) Profile of second replicate, maximum is at position 13,703.

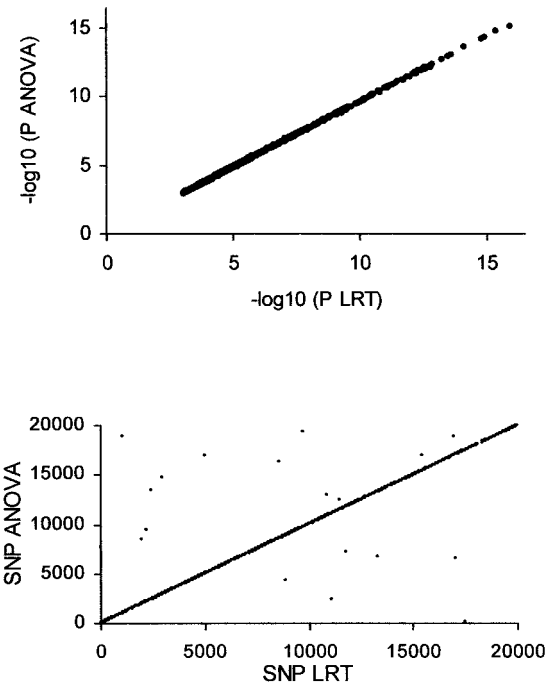


FIGURE 5.—Comparison between the analysis of variance and the maximum-likelihood methods, in terms of *P* values (−log 10) and of position estimates. Results correspond to one replicate of the W set.

metric model based on the normal distribution. We tested the effect of the method in two ways. First, we reran the analysis of one of the replicates in the W set, using an analysis of variance. The classificatory variable was the SNP genotype and *P* values were computed from the corresponding *F* distribution. The correlation between the likelihood-ratio test (LRT) and ANOVA *P* values was virtually 1 and the coincidence between position estimates was almost perfect, even for nonsignificant mRNA levels (Figure 5). At equal *P* values, the LRT resulted in 345 significant clones, very similar to 354 obtained with ANOVA. For those significant repeated clones, the position estimates coincided exactly in all cases.

Second, we compared the original results of BREM *et al.* (2002), who used a nonparametric test (Mann-Whitney), with our results. They used a significance *P* value of $5 \times 10^{-5}$ and obtained 570 significantly different mRNAs. The *P* value that corresponds to the 570 most significant clones with our method was slightly more strict (*P* value = $10^{-6}$). Nevertheless, it should be noted that many mRNAs had very similar *P* values so a slight change in the cutoff threshold made a large difference in the number of significant mRNA levels. Of the 570 mRNAs that were deemed significant in BREM *et al.* (2002), 17 were not analyzed here because they had >75% missing data, 74 had a *P* value $>10^{-6}$, but most were in the bordering region, with only 13 mRNAs having *P* values $>10^{-5}$. The remaining mRNA levels were deemed significant in both analyses. In terms of positions, 260 position estimates coincided exactly with the same position, 75 were in the next contiguous SNP, 463 in total were within the closest 10 SNPs, 17 linkages resulted in different chromosomes, and the remaining ones were within 100 SNPs away. Thus, globally, both methods gave similar results, with a large part of the conflicting results in the bordering on significance region. All in all, it does not seem that the results presented here have been largely affected by the choice of the maximum-likelihood method.

**Conclusion:** This work faced two difficult simulation problems, as it is not evident either how to simulate a complete chromosome polymorphism or how to simulate microarray data. We avoided the latter by using real data and we proposed an approximation for the former. Certainly, the importance of the potential risks identified in this work depends on how realistic our simulation is, as with any other simulation study.

All in all, it seems that the intrinsic difficulties of dissecting the transcriptome genetic architecture in outbred populations should not be underestimated. We anticipate that (many) conflicting results may be collected in the future if whole-genome association studies for mRNA levels are carried out in outbred populations. The evidence of eQTL hotspots should be carefully evaluated before concluding that they are not mere artifacts due to a high correlation between mRNA levels and/or markers (as a result of linkage disequilibrium). It also seems that much variation of expression levels will remain uncovered or unconfirmed.

Association studies, like the one presented here, are known to be risky because of the difficulty of controlling false positives, and this study reflects in part these risks. Alternative options include more robust strategies like transmission-disequilibrium tests or sib-pair analysis (Lynch and Walsh 1998). We should be aware, nevertheless, that in the end causal mutations will be found using association studies.

In the meantime, in terms of experimental design, we recommend that related individuals should be sampled. By sampling related individuals (*e.g.*, sibs) two objectives are achieved: first, robust although imprecise linkage methods can be applied; second, more importantly, the number of distinct haplotypes present in the sample is reduced and amplified in a manner that resembles crosses between inbred lines. The usual tradeoff applies: for a constant $n$, the larger the families the larger the power but also an increased risk of missing important variability in the whole population. Large families result also in a smaller linkage disequilibrium signal and, thus, in an increased confidence interval for the QTL position. Finally, we also recommend comparing the profiles of the individual replicates of the same mRNA level to look for possible inconsistencies.

I am grateful to Richard Hudson for his coalescence programs, to Julio Rozas for comments on the coalescence process, to Adeline R. Whitney and Gaël Yvert for help with their data and discussion, and to all people involved in making their data public. Significant ($P < 10^{-12}$) improvements to this work resulted from the reviewers' comments. This work was funded by Action en Bioinformatique of Ministère de la Recherche, France. The author holds a professorship from Institut Català de Recerca i Estudis Avancats (ICREA).

## LITERATURE CITED

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos *et al.*, 2000 Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403:** 503–511.

Brem, R. B., G. Yvert, R. Clinton and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science **296:** 752–755.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins *et al.*, 2001 The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291:** 1289–1292.

Cheung, V. G., and R. S. Spielman, 2002 The genetics of variation in gene expression. Nat. Genet. **32** (Suppl): 522–525.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K. Y. Jen *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. Nat. Genet. **33:** 422–425.

Churchill, G. A., 2002 Fundamentals of experimental design for cDNA microarrays. Nat. Genet. **32** (Suppl): 490–495.

Darvasi, A., 2003 Genomics: gene expression meets genetics. Nature **422:** 269–270.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanics of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer, Sunderland, MA.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

Jansen, R. C., 2003 Studying complex biological systems using multifactorial perturbation. Nat. Rev. Genet. **4:** 145–151.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

Oleksiak, M. F., G. A. Churchill and D. L. Crawford, 2002 Variation in gene expression within and among natural populations. Nat. Genet. **32:** 261–266.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding *et al.*, 2003 Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat. Genet. **33:** 382–387.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Rosenwald, A., G. Wright, W. C. Chan, J. M. Connors, E. Campo *et al.*, 2002 The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N. Engl. J. Med. **346:** 1937–1947.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature **422:** 297–302.

Wayne, M. L., and L. M. McIntyre, 2002 Combining mapping and arraying: an approach to candidate gene identification. Proc. Natl. Acad. Sci. USA **99:** 14903–14906.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

Whitney, A. R., M. Diehn, S. J. Popper, A. A. Alizadeh, J. C. Boldrick *et al.*, 2003 Individuality and variation in gene expression patterns in human blood. Proc. Natl. Acad. Sci. USA **100:** 1896–1901.

Communicating editor: C. Haley