Genetical genomics with Affymetrix gene expression arrays

Alberts, Rudi

CHAPTER 5

Sequence polymorphisms cause many false *cis* eQTLs

Submitted as: Alberts, R., Terpstra, P., Li, Y., Breitling, R., Matthijssen, D.I., Nap, J.P., Jansen, R.C. – *"Sequence polymorphisms cause many false cis eQTLs"*

Abstract

Motivation Many papers have recently reported the successful mapping of quantitative trait loci for gene expression phenotypes (eQTLs). Especially local (putatively *cis*-acting) eQTLs, where gene and controlling locus coincide, are of great interest, since they are direct candidates for previously mapped physiological quantitative trait loci. How often do sequence polymorphisms in the mRNA regions that are interrogated by microarrays lead to false detection of *cis* eQTLs and can these false *cis* eQTLs be eliminated?

Results We show that polymorphisms falsely suggest *cis*-regulation on a large scale. We also show that many such polymorphisms can be detected by a sensitive statistical approach that takes the individual probe signals into account. We conclude on the basis of application of this approach to recent mouse and human eQTL data that many false eQTLs can be successfully eliminated.

Availability `http://gbic.biol.rug.nl/supplementary/2006/probetools`

5.1 Introduction

Genetical genomics – linkage and association analyses of gene expression phenotypes with the help of microarray data – is a promising strategy to identify regulatory determinants of complex traits or diseases (Jansen and Nap 2001, Rockman and Kruglyak 2006). Notably, *cis*-acting genes, which regulate their own expression, may establish direct targets for diagnosis and treatment. However, mRNA sequence diversity in probe regions is known to influence hybridization on microarrays considerably. In this case, differential hybridization is incorrectly interpreted as *cis*-regulated differential gene expression. Here we show examples that demonstrate how polymorphisms in the mRNA sequence have led to many false positive *cis* eQTLs in recent studies. Furthermore we provide evidence that this is a systematic problem. Until now, the popular tools for processing GeneChip data (MAS 5.0, dChip, RMA; `http://www.bioconductor.org`) calculate average expression levels per probe set, and there are no tools with which probe-level signals can be easily explored. We here present a user-friendly software with which probe-level analysis can be easily performed to discriminate genes that are differentially expressed from genes that show differential hybridization caused by polymorphisms or other mechanisms such as alternative splicing.

5.2 Methods

eQTL analyses

In this paper we have analysed a *C. elegans*, a human, and a mouse dataset. The *C. elegans* data were analyzed as in (Li et al. 2006). Average eQTL effects for both temperatures were used. The human immortalized lymphoblastoid cell data (Cheung et al. 2005) and the mouse hematopoietic stem cell data (Bystrykh et al. 2005) were reanalyzed using a sensitive statistical analysis of variance (ANOVA) model reported in Alberts et al. (2005) and extended here for filtering out possible false positive eQTLs. The model decomposes the probe signals for a given probe set into $log(y_{ij}) = m + P_j + A_i + PA_{ij} + e_i + e_{ij}$, where y_{ij} is the signal of the j^{th} probe of the i^{th} sample, m is the average signal, P_j is the average effect of the j^{th} probe, A_i is the average effect of the allele carried by the i^{th} sample at a given genome position, PA_{ij} is the interaction effect between probe and allele type, e_i is an error term per sample, and, finally, e_{ij} is a probe-specific error term per sample. For the mouse data additional parameters for batch effect were added (see Alberts et al. 2005).

Probe backwards elimination procedure

In this procedure probes are eliminated from the probe set one by one until the interaction effect PA_{ij} between probe and allele in the aforementioned model has been disappeared. To determine whether the interaction has disappeared, we use a threshold based on the *trans*-regulated mouse genes, because those genes are expected not to be affected by the hybridization artifacts and to give a realistic picture of the (limited) amount of interaction that is present in any probe set not affected by hybridization artifacts. We computed the P-values for statistical significance of the interaction terms PA_{ij} for each of the *trans*-regulated mouse genes. We used the 99^{th} percentile of these P-values as a threshold for *cis*-regulated genes in the human and mouse data to backwards eliminate the (most) deviating probes one by one. The procedure starts by temporarily removing each probe one by one and calculating the interaction effect among the non-dropped probes in each case. Then the probe whose removal caused the largest decrease in interaction effect is permanently eliminated. This procedure is repeated with the remaining probes until the interaction effect has crossed the threshold. Only the remaining probes are used for the final eQTL analysis.

5.3 Implementation

ProbeTools is a computer program written in Java and R with which probe level analysis can be easily performed. It reveals important information about individual probes that remains unnoticed when only probe set averages are calculated. The input of ProbeTools is a set of Affymetrix .CEL files and optionally a table with genotype data. The user can select MAS 5.0, dChip or RMA for background correction and normalization. After this preprocessing, one can select probe sets (genes) to study in more detail and apply the backwards elimination method. For each probe set a figure with relative probe locations on the mRNA, the probe signals from the .CEL files and results of the elimination method are displayed (Figure 5.1). This visualization provides direct insight in the behavior of different probes within a probe set. After clicking the link beneath the figure, the individual probe positions are displayed directly in the UCSC Genome Browser (Figure 5.1) providing useful genome information about, e.g., known polymorphisms, deletions or insertions within probes and/or alternative splicing. ProbeTools is freely available for non-commercial use via `http://gbic.biol.rug.nl/supplementary/2006/probetools`.

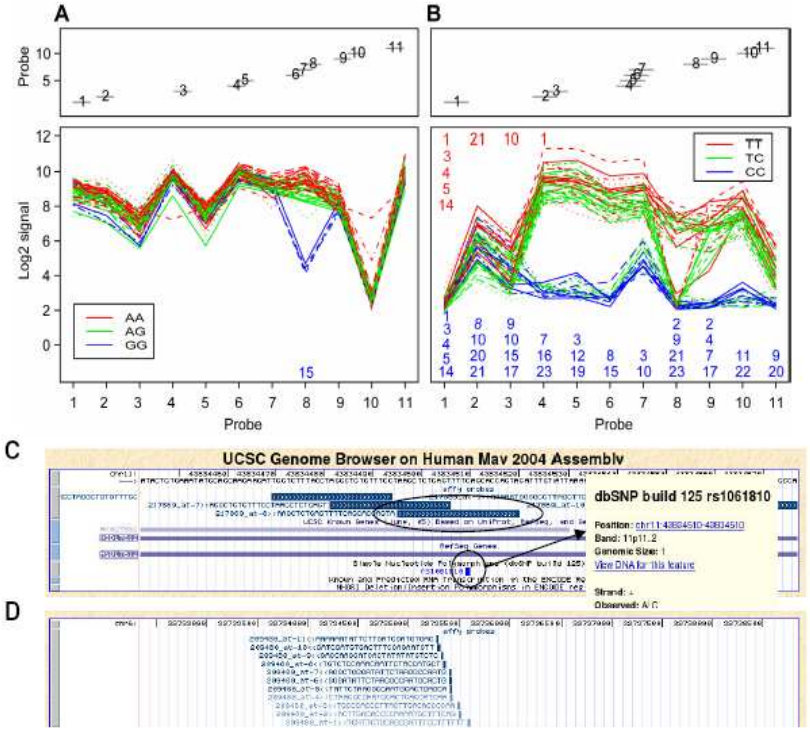

Figure 5.1: Two false *cis* eQTLs in human association analysis. (A) Relative probe positions on the mRNA (top) and hybridization signals (bottom) for gene HSD17B12 reported as *cis*-regulated in (Cheung et al. 2005). Each individual (line) is colored according to the allele he carries for the associating marker rs4755741. This marker is located in an intron of the HSD17B12 gene and it is strongly linked with SNP rs1061810 located in probe 8. By omitting the data for probe 8, the presumed significance of the marker effect disappears. (B) Similar plot for gene HLA-DQB1 also reported as *cis*-regulated in (Cheung et al. 2005). Lines were colored according to the allele the individual carries for associating marker rs6928482. One red-line and one blue-line individual have been sequenced for the probe region; the numbers in red and blue indicate the positions of SNPs within the 25-mer probes. The number in italic indicates a single nucleotide insertion in probe 2. We observed 23 SNPs (15 new ones) between these mRNA sequences and the 11 probes. (C) Visualization of the probes for gene HSD17B12 in the UCSC Genome Browser. The probes are displayed in blocks, with labels indicating probe set name, probe number, orientation on the genome and probe sequence. Probe 8 and SNP rs1061810 are encircled. The inset shows information about this SNP. Further browsing to dbSNP shows the diversity of the SNP for the CEPH population. The Genome Browser shows one SNP in probe 8, as was expected from the probe signals. (D) Similar plot for gene HLA-DQB1. Probes 1-4, that do not perfectly match the genome, are displayed in light blue. The Genome Browser currently shows fewer SNPs than found in our own sequencing effort.

5.4 Results/Discussion

Affymetrix GeneChip arrays (Lockhart et al. 1996) are frequently used to measure mRNA abundances. On this platform, multiple 25-mer probes (probe set) are used for each gene. To estimate how many probe sets are expected to contain SNPs and hence to be 'sensitive' for differential hybridization, we simulated randomly distributed SNPs in Affymetrix probes using a conservative estimate of one SNP per 1000 basepairs (Consortium 2005). This showed that many probe sets are expected to contain one or more SNPs. For example, on the human genome HG-U133 Plus 2.0 arrays, at least 200 probe sets will carry SNP variation in three or more probes.

The expected hybridization effects, resulting from such variation, can seriously mislead the interpretation of individual genes, even if only a single probe is affected. In a recent extensive genetical genomics experiment using Affymetrix arrays (Cheung et al. 2005), the expression of thirteen human genes was found to be significantly *cis*-regulated in immortalized lymphoblastoid cells. One of these genes was HSD17B12. Analysis of variance of the probe data (see Methods and Alberts et al. 2005) of HSD17B12 reveals strong evidence for *cis*-regulation when all probe data are used (Figure 5.1). The evidence reduces to a non-significant level when the data of probe 8 is omitted. Probe 8 is the only from eleven probes that shows a clear differential signal. The standard MAS 5.0 algorithm used to summarize probe level data is not able to single out this outlier probe, nor are the alternative methods dChip or RMA. In the CEPH individuals used, HapMap (Consortium 2005) only reports an SNP in probe 8 (rs1061810), located at position 15 in the probe and showing A/C variation for the CEPH individuals. This SNP is strongly linked with the SNP found in the association study (rs4755741) showing A/G variation. The data indicate that the mRNA from individuals that are homozygous A for SNP rs4755741 hybridizes better than mRNA from individuals that are homozygous G for SNP rs4755741. This is as expected, because the mRNA from individuals that are homozygous A is identical to the probe sequence, whereas the mRNA from individuals that are homozygous G has a mismatch.

A similar problem is seen for HLA-DQB1, another of the putative *cis*-acting genes in the same study (reported as HLA-DRB2 in Cheung et al. 2005). The differential hybridization that suggested *cis*-regulation in fact only reflects SNP variation in the individual probes. We sequenced the HLA-DQB1 alleles of two individuals. Differences in hybridization signal between individuals could be attributed to differences between probe sequences and mRNA sequences from different individuals (Figure 5.1).

Alberts et al. (2005) show an example of mis-identification of a *cis*-acting gene in a genetical genomics experiment, using a panel of 30 recombinant inbred mice derived

from parental strains C57BL/6 (B6) and DBA/2 (D2). Here we assess how many of the *cis* genes in the original study (Bystrykh et al. 2005), from which this example was taken, were caused by differential hybridization. Without a hybridization artifact we expect that 50% of the *cis* eQTLs should show a higher signal for the mice carrying the B6 allele (called *cis*+), and 50% show a higher signal for the mice carrying the D2 allele (called *cis*–). Due to the fact that the microarrays were mainly designed based on B6 sequence, the hybridization artifact will cause an excess of *cis*+ eQTLs. In this study there were indeed many more *cis*+ eQTLs: 254 *cis*+ vs. 145 *cis*–. We now expect to have 145 'real' *cis* eQTLs in either direction, and the excess of 'artifactual' *cis*+ eQTLs is calculated as the amount of *cis*+ minus the amount of *cis*–, here 254-145=109. Among the 100 most significant *cis* genes, the excess is 40. This means that almost half of the reported most significant *cis*-linked genes are probably due to a hybridization artifact. Doss et al. (2005), Manly et al. (2005) and Pierce et al. (2006) observed similar excesses of *cis*+ eQTLs in mouse tissues. It should be noted that such a systematic difference in *cis*+ and *cis*– eQTLs is only expected when the arrays are designed using sequences from one of the parental lines. They are not seen, e.g., in a recent rat eQTL study (Petretto et al. 2006), when probes are mainly based on ESTs and cDNA sequences from outbred animals. However, the absence of a *cis*+ excess does not mean that no hybridization artifacts are present.

Affymetrix experiments, which use very short individual probes, are expected to be most sensitive to hybridization artifacts, as single polymorphisms will strongly affect the binding behavior. However, (Hughes et al. 2001) show that SNPs also cause hybridization differences in 60-mer microarrays. (Li et al. 2006) used 60- mer cDNA microarrays in a genetical genomics experiment using 80 C. elegans recombinant inbred strains derived from parental strains N2 (Bristol) and CB4856 (Hawaii). The arrays were designed based on N2 sequence. There was an excess of 223 out of 399 *cis* genes having a higher signal for individuals carrying the N2 allele (*cis*+), and an excess of 74 among the 100 most significant *cis* genes. This shows that hybridization differences lead to pervasive artifactual *cis* eQTLs even for long-oligomer microarrays.

Filtering out the false eQTLs in the *C. elegans* study by statistical approaches is impossible, because only one probe per gene was used. However, when multiple probes per gene are used, differential signal can change dramatically from one probe to another due to SNPs as exemplified here by human HSD17B12 and HLADQB1 (Figure 5.1) and in Alberts et al. (2005) by mouse ALDH9A1 (*cis*+, probe set 96243_f_at, harbouring two SNPs). In this case a sensitive statistical approach, that takes the data of individual probes into account, can eliminate deviating probes and test *cis*-regulation using the remaining probes only (see Methods). Screening the 100 most significant mouse *cis* eQTLs, this approach tagged 23 *cis*+ eQTLs as false positives (22 lost significance, one became *cis*–). In contrast, only two *cis*– eQTLs were

filtered out (for reasons that we explain below). The fact that 23 *cis*+ and only two *cis*– eQTLs were filtered out is a strong indication that the method works well, given the fact that the hybridization artifact causes an excess of *cis*+ genes. To further assess the utility of the method, we focus on probe sets with known SNPs below. In cases where many probes contain SNPs, simply eliminating probes might not work. E.g., twelve of the 16 probes of *cis*+ ALDH9A1 were affected by two known SNPs, but the statistical approach actually eliminated 4 probes not hit by the SNPs and 6 probes hit by the less influential SNP. To solve this problem, we tagged all genes for which $\geq$50% of the probes are eliminated. This uncovered ALDH9A1 and one other mouse *cis*+ gene as false positives.

Not every SNP causes a difference in hybridization. Eg. when the SNP is located at the very beginning or end of the probe, it can have little or even no effect (Hughes et al. 2001). To assess the utility of our method, we focus on thirty known SNPs between B6 and D2 that were found to affect 24 probe sets. By investigating the probe signals and knowing which probes contain SNPs, we saw that in 14 of the probe sets the SNPs caused a difference in hybridization, while in the other 10 probe sets the SNPs had no effect. Strikingly, our method correctly tagged the 14 SNP containing probe sets and successfully eliminated only the SNP containing probes in those probe sets. The 10 probe sets in which the SNPs had no effect, and hence no probes needed to be eliminated, correctly remained untagged. Most probes are based on B6 sequence, but not all. It is therefore possible that a deletion or insertion in B6, or an SNP between B6 and the probes can cause false *cis*– eQTLs. Indeed, out of the two *cis*– genes that were filtered out, at least one was false (caused by a combination of a known deletion and an SNP in the B6 sequence; gene H2-D1).

Because hybridization data can incorrectly suggest *cis*-regulation in genetical genomics experiments, we recommend using multiple (tiling) probes per gene that allow statistical filtering as defined in this paper. However, critics would be right to argue that statistics alone can not solve what is ultimately a biological problem; e.g. human HLA-DQB1 (a highly polymorphic locus) in the CEPH population contains SNPs in all probes and, therefore, this probe set can not be corrected by our probe elimination method. For this reason we strongly advocate that additional genome-wide methods to characterize polymorphisms (Gresham et al. 2006), re-sequencing of probe regions, or alternative ways of gene expression profiling are employed whenever strong claims about *cis*-regulation are to be made. Only then large-scale mapping of the determinants of gene expression will become truly informative.

5.5 Acknowledgements

This work was funded by The Netherlands Organisation for Scientific Research (to RA, RB, and YL) and by Senter Novem (The Netherlands Ministry of Economic Affairs, to DIM).