Genetical genomics with Affymetrix gene expression arrays

Alberts, Rudi

CHAPTER 7

General discussion, Dutch summary, Dankwoord, CV, Publications, Poster Award

7.1 General discussion

In 2001 Jansen and Nap introduced the concept of genetical genomics. This involves the combination of molecular marker data with gene expression profiles derived from genetically related individuals. It reveals regulatory links between genes and helps in uncovering the determinants of complex traits (Cheung et al. 2005). Since gene expression microarrays have become more and more affordable, multiple large scale genetical genomics studies have been carried out in the past years. This thesis focused on genetical genomics performed using the popular Affymetrix GeneChip® arrays. Although the Affymetrix technology is now widely used, we show that there are still some problems when applying it to genetical genomics experiments and we provide solutions to these problems.

Batch effects

A first problem, which is not specific for Affymetrix arrays, is that the preprocessing of microarrays in multiple batches possibly introduces batch effects in the resulting data. In the dataset of Bystrykh et al. (Bystrykh et al. 2005) for example, the effect of the three processing batches was clearly visible, even after array normalization. Spurious correlation between batches and alleles for molecular marker D4Mit17 on chromosome 4 caused hundreds of false ('ghost') QTLs on this chromosome (see Table 7.1). Application of our ANOVA model including batch effect successfully eliminates the batch effect from the data. The hundreds 'ghost' QTLs are not detected anymore. The application of similar models to existing or new data sets is expected to avoid these problems in the future. A drawback of correcting for batch effect in this way is that genes that have a real QTL effect for this marker on chromosome 4 won't be detected. The model can not discriminate between the real QTL effect and the batch effect in this case. Moreover, the real QTL effect may be confounded with batch effect. No QTL effect will be detected. This can be solved by adding more individuals to the experiments, so that the spurious correlation between batches and alleles for the marker on chromosome 4 decreases.

	batch 1	batch 2	batch 3
allele B6	10	10	2
allele D2	0	2	6

Table 7.1: Spurious correlation between alleles and batches for marker D4Mit17 on chromosome 4. Numbers indicate amounts of mice.

Probe accuracy

Since the time of probe design by Affymetrix, knowledge about messenger and genomic sequences has been increased. For example, UniGene clusters could have been merged, because what was thought to be two genes appeared to be one gene, or clusters can be depricated, e.g. because they were based on only one or a few erroneous EST sequences. This led us to verify Affymetrix probe sequences against the latest messenger and genomic databases and update the probe set definitions. For most recent Affymetrix arrays this revealed that around 85% of the probe sequences were sequence verified. Although the percentage of unverified probes seems low, the improvement in the detection of differentially expressed genes using updated probe set definitions shows that it is worthwile to reannotate Affymetrix probe sets based on the latest sequence information. Moreover, the unverified probes can be very influential, even if there is only one unverified probe per probe set. An SNP in only one probe can cause a large QTL effect for the whole probe set.

Our probe verification has indicated the probes for genes GSTM1 and GSTM2, reported as *cis*-acting in Cheung et al. (2005), as cross-hybridizing. The genes are located in duplicated regions and the signals should indeed be interpreted with care since it is unclear of which genes the mRNA is measured by the probes.

Probes telling different stories

Application of our statistical model to the probe level Affymetrix data has revealed that many probe sets show probe by QTL interaction. When comparing *cis* and *trans* genes, we have seen that especially *cis* genes show strong probe by QTL interaction frequently. We have hypothesized that this is due to sequence polymorphisms, alternative splicing, insertions and deletions etc. leading to a difference in hybridization and we have shown a few examples where SNPs caused the interaction effect. It is advisable to use our statistical model in genetical genomics experiments to filter out probe sets without consistent QTL effect over probes, i.e. having QTL by probe interaction, and study those with more care since other factors may be underlying the differences in signal.

When inbred populations are used and when the microarrays are designed based on the sequence of one of the parental lines, one can assess the occurrence of the hybridization artifact by looking at the directions of the *cis* genes. When an SNP caused differential hybridization, the signal of the individuals carrying the allele on which the microarray was designed is expected to be higher (cis+ genes), since they perfectly match the probe on the arrays. For multiple studies we have shown significant excesses of cis+ genes compared to cis- genes, indicating the prevalence of the problem. For Affymetrix arrays we have introduced a statistical method that

automatically backwards eliminates the deviating probes (SNP containing probes) from *cis* probe sets. This method appeared to work very well on the Bystrykh et al. dataset (Bystrykh et al. 2005): among the 24 probe sets with 30 SNPs between B6 and D2, in 14 of the probe sets the SNPs caused ad difference in hybridization, while in the other 10 probe sets the SNPs had no effect. 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. Table 7.2 indicates for the genetical genomics studies mentioned in the Introduction of this thesis, whether the authors were aware of the hybridization artifacts, which appears to be the case in 3 out of the 5 studies.

paper	problem realized?	problem solved / circumvented?
Hubner et al. 2005	Yes	No. They sequenced 15 *cis* genes and did qRT-PCR. Only one showed sequence difference.
Bystrykh et al. 2005	Yes	No. They say that only 0.3% of the probes overlap one of 1.2 million Celera B6/D2 SNPs, most of which are not *cis*.
Chesler et al. 2005	Yes	No. They say that only 0.6% of the probes overlap one of 1.2 million Celera B6/D2 SNPs and that only 1 out of 10 cause a difference in signal
Morley et al. 2004	No	-
Cheung et al. 2005	No	-

Table 7.2: Were the authors of other genetical genomics studies aware of the hybridization artifact?

In the upper three papers in the table, the authors think it is a low-percentage problem. However, in Chapter 5 we have shown that SNPs can have a strong effect on hybridization, also if the SNP is only in one of the probes, and that this occurs in many of the strongest *cis*-acting genes.

Alternative splicing

Suppose a gene with 5 exons is interrogated by a probe set with 16 probes, 8 of them covering exon 4 and 8 of them covering exon 5. Further suppose that the gene has two splice variants, one variant with all exons and one variant that lacks exon 4. In this case, the probe signals may look like the ones depicted in Figure 7.1.

The individuals with a black color in the figure are the ones with the variant containing all exons, and show a high signal for all probes. The individuals with the grey color are the ones with the variant without exon 4. Since probes 1 till 8 are located on this exon, their signal is lower. Probe sets in which this happens are

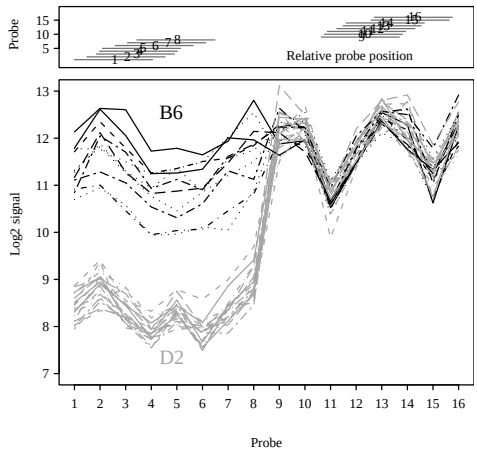

Figure 7.1: Probe signals for a gene that shows alternative splicing.

directly detected by our ANOVA model, since they will have a large QTL by probe interaction effect. When averages of probe signals per individual would have been used, this probe set would have been identified as *cis*-acting, while in reality it is not.

Comparing microarray platforms

In this thesis we have seen that the Affymetrix technology uses multiple short-oligonucleotides per gene to measure its expression. Furthermore, we have seen that in genetical genomics experiments, sequence polymorphisms lead to the detection of false *cis* eQTLs. This even appeared to occur in a *C. elegans* experiment where one 60-mer probe per gene was used (Illumina arrays). In this situation, one does not have the possibility to compare the signals of multiple probes per gene and confirm that only SNP covering probes show a difference in signal, while this is possible when multiple probes per gene are used. In this case the SNP covering probes can be filtered out and the QTL analysis can be proceeded with the remaining probes. Because of this I advice to use microarray platforms with multiple probes per gene in genetical genomics experiments, preferably covering all exons.

General conclusion and outlook

In general, this thesis shows that past eQTL studies should be carefully interpretated and some conclusions from published studies probably need to be modified.

"Ghost" eQTLs caused by batch effects or polymorphisms have been shown to occur frequently, especially in the lists of most significant *cis*-acting genes. However, our increasing knowledge of messenger sequences, progression in microarray technologies and improvements in bioinformatic analysis techniques will ultimately strengthen the conclusions that are drawn from eQTL studies.