Validity studies on alternative interpretation of microarray data

Leo Lahti*

15th April 2005

Abstract

Although any given microarray is designed on the most up-to-date sequence information available, rapidly evolving genomic sequence data can reveal inconsistencies and inaccuracy in the annotations of microarray data as the body of knowledge grows. Moreover, exploitation of the most current knowledge on genomic sequences makes more accurate analyses possible. In this paper, evidence on the validity of such approach is provided based on preliminary studies on Affymetrix HG-U133A and HG-U133B chips.

1 Introduction

Microarrays enable researchers to monitor the activity of thousands of genes or other transcripts simultaneously in a given sample, often with an objective of finding a small set of relevant genes for further study. One type of microarrays are high-density short oligonucleotide chips [6] that contain typically 8-16 probes for each transcript target that the chip is intended to measure. Each 25-mer probe is expected to uniquely hybridize with its respective target, and the detected hybridization level is used as a measure of the activity of the transcript. The use of several probes for each mRNA target leads to more robust estimates on transcript activity but it is clear that probe quality may significantly affect the results of a microarray study.

In recent studies, a remarkable number of probes on various oligonucleotide arrays have been detected not to uniquely match their intended mRNA target [3, 5, 8, 9]. The rapidly evolving nature of genomic sequence data contributes to this finding as our knowledge of genomic mRNA i.e. the *transcriptome*, is greater today than at the time of array design. Matching all probes on a given array to updated genomic mRNA and using this information for building an alternative interpretation of the array potentially provides researchers with more accurate lists of significant genes and can make justified comparisons between different array types possible. These are the two major goals in our studies.

In this paper, we briefly review earlier studies and present our own results from validity studies on updated probe matching approach for human genome arrays HG-U133A (chip A) and HG-U133B (chip B) of Affymetrix. Both chips contain probe sets for >22000 human transcripts, with probe sets for known genes on chip A, and for expressed sequence tags (ESTs) on chip B.

*Turku Centre for Biotechnology, P.O.Box 123, BioCity, Turku, FIN-20521, FINLAND and Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 HUT, FINLAND

2 Data

Gene expression data from nine human embryonic stem cell lines and a fibroplast sample were used for the expression analyses of this study, with two replicates for each line (data not public). Additionally, an embryoid body control mix sample was available. Summing these up, we have data from 21 samples on both chip types A and B. In this study, we concentrate on the general properties of short oligonucleotide array data and further knowledge of the biological properties of the samples is not needed. The 22715 human RefSeq sequences used throughout this study were extracted from the vertebrate mRNA data downloaded from NCBI ftp site [1].

3 Experiments

3.1 Matching probes to genomic mRNA

Matching probe candidates to genomic mRNA is part of a typical probe design process (see e.g. [1]). Sequence information evolves, however, and it is possible to assess probe quality [5, 8, 9] or even to construct alternative probe sets [3] by using the most up-to-date genomic data. This has been shown to increase the accuracy and cross-platform consistency of microarray analyses [7, 9].

In principle, we could match our probes against a collection of the latest releases of all available genomic mRNA sequence data. Not surprisingly, the use of a more comprehensive data seems to lead to more comprehensive results [9] but more complicated considerations on sequence redundancy and intensive computation will become necessary as the size of the database grows. Smaller non-redundant databases like RefSeq have been succesfully used in similar studies as a sufficient approximations of the human transcriptome [3, 7].

We chose the RefSeq database since it is the most highly curated, publicly available and moderately sized genome-wide definition of transcript sequence information. Only 25-mer perfect matches of probes to the RefSeq sequences were considered in this initial study and searched for by following the guidelines given in [3]. It should be noted that remarkable cross-hybridization to the probes by mRNAs with closely similar hut not perfectly matching sequences have been reported (see e.g. [9]).

Alternative mapping defines new probe sets that are named according to the mRNA target of the probes in the set. The total number of measured RefSeq sequences is smaller in our alternative mapping than the number of probe sets in the original array interpretation (16564 vs. 22283 on chip A; 7061 vs. 22645 on chip B). Some RefSeq sequences are not measured by any probe on either chip A or chip B. However, 96% and 76% of the measured RefSeq sequences have eight or more probes on chip A and chip B, respectively, making estimates of mRNA expression statistically more robust than with a lower number of probes.

A large number of *mistargeted* probes with no genomic targets and also a number of *non-specific* probes with two or more targets were identified on both chips A and B. 26% and 73% of the probes of chip A and chip B do not match any sequence in the RefSeq, respectively. A reason for the remarkably low number of matches in RefSeq with probes on chip B is probably that this chip is aimed at probing EST targets. Poorly annotated ESTs are absent from the RefSeq database that mainly contains well-annotated curated sequence data. Surprisingly high number

[1] ftp://ftp.ncbi.nih.gov/refseq/release/vertebrate_mammalian/

of mistargeted probes on chip A is probably due to incompleteness of the RefSeq data, too, and this conclusion is also supported by our preliminary studies on another genomic sequence data set. 17% and 4% of the probes were found to be non-specific on chips A and B, respectively. The mRNA targets of a given non-specific probe often seem to be related and even residing at the same genomic location. Alternative transcripts of the same gene can be responsible for such phenomenon, for example, and sometimes it may be misleading to consider such probes as non-specific. Despite the evident incompleteness, the RefSeq seems a valid choice for initial studies as most probes of chip A are *specific*, having exactly one target in the RefSeq and numerous probes of chip B could also be uniquely matched to RefSeq targets.

A straightforward comparison between the results from the original and alternative mapping are not possible due to the drastically different interpretation of the arrays in the two cases. For valid comparisons, we use original probe sets only, revisited with our RefSeq matching results. Expression values were computed in R by using the MAS5.0 algorithm with default parameters in the 'affy' package [2] of the Bioconductor project.

3.2 Number of mRNA targets affects the expression of a probe set

If all probes within a given probe set match identical RefSeq target(s) or have no detected targets in the RefSeq, the probe set is considered as *verified* and accepted for further study. Such probe sets are either mistargeted, specific, or non-specific, according to the number of RefSeq targets of the probes within the probe set. 87% of the original probe sets on chip A were accepted after removing control sets and sets with ambiguous targets.

Probe specificity seems to affect the expression value of a probe set (see Fig. 1). The mistargeted and the non-specific probe sets tend to have lower and slightly higher expressions than the specific probe sets, respectively. This shows that the matchings are sensible although they may not be complete. Expressions vary over a wide range of values in all specificity levels and it is not possible to draw conclusions of expression levels for individual probe sets, however.

3.3 Justified comparisons between different arrays by RefSeq matching

Comparison between different chip types is often based on probe set pairing by annotations provided by array manufacturer. Commonly used options include LocusID, UniGene and RefSeq identifier annotations. We used three different options for pairing probe sets between chips A and B, the (i) LocusID and (ii) RefSeq annotations provided by Affymetrix and (iii) confirmed RefSeq sequences from our own experiments.

To avoid overrepresentation of sets corresponding to any one identifier, just one probe set pair between chips A and B was randomly chosen for each unique identifier. Identical probe sets and control sets were excluded from pairing to avoid optimistic bias to the correlation results. Correlation of the paired expression values from the two chips was computed for all 21 samples available (see Fig. 2). Our RefSeq confirmed probe set pairs show highest correlation but the number of pairs is lower than for annotation-based probe set pairs. Although our objective here was not to make the chips A and B comparable, this result shows the potentiality of updated probe matching approach for justified inter-chip comparisons.

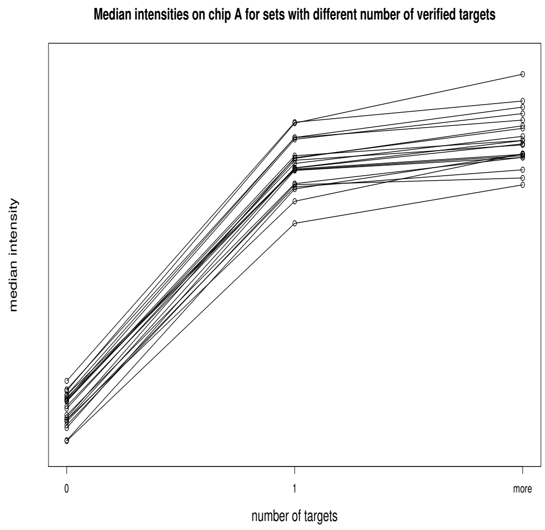

Figure 1: Median probe set intensities computed for chip chip A sets with 0, 1 or more hits to RefSeq sequences. Data for all 21 samples are shown. Probe sets with no matches on RefSeq sequences (5612) tend to have a lower intensity than other sets ($p < 10^{-16}$, one-sided Mann-Whitney test). The difference is not so clear ($p < 0.05$ for 13/21 samples) between sets with one and multiple matches (11005 and 2741 sets, respectively).

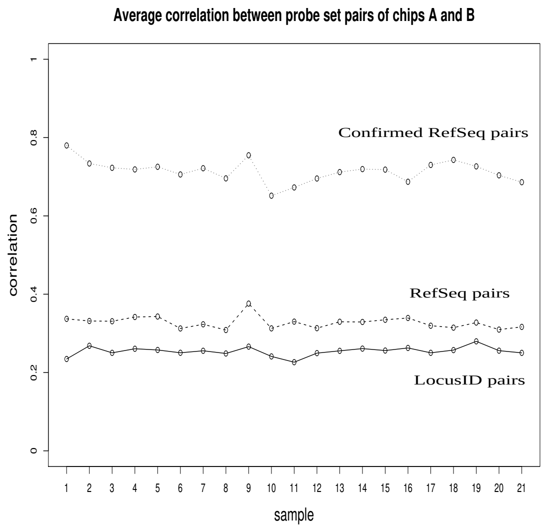

Figure 2: Correlation of probe set pairs between chips A and B across the 21 samples. Solid line connects the correlations between LocusID paired probe sets (5042 pairs). Dashed line corresponds to the pairs formed by RefSeq annotations provided by Affymetrix (4799 pairs). Correlations connected with the dotted line are clearly higher and belong to our RefSeq confirmed probe set pairs. However, number of pairs is lower here (949 pairs).

4 Conclusion

Our initial results confirm that mistargeted and non-specific probes may significantly affect the results of a microarray study and that the RefSeq-based matching of probes to mRNA targets makes sense. It has been suggested in earlier studies that probe verification can be used either to remove bad probes or probe sets from subsequent analyses or to build a totally new mapping of probes to mRNA targets, with the objective of getting more accurate results out of a microarray analysis [3, 5, 8, 9].

In addition to presenting probe matching statistics against the RefSeq database, we have shown that the number of confirmed mRNA targets generally affects the expression level of a probe set. Moreover, increased correlation of RefSeq verified probe set pairs between chips A and B shows that the updated matching approach may lead to more justified comparisons between different array types than ordinary comparisons. However, the RefSeq database seems insufficient for studies on the EST-focused chip B and the use of a more comprehensive mRNA database could lead to more accurate results also for chip A.

In future studies we will use the alternative mapping approach combined with other methods for analyzing and comparing microarray data, with the main objective of obtaining more accurate results from individual expression analyses and enhanced comparability between the data from different array types. A more comprehensive genomic database will be used, possibly combined with EST sequence data and information on genomic locations of the targets.

Acknowledgments

I would like to thank Tero Aittokallio of the idea for this study and Laura Elo for useful discussion and literature tips. I am also thankful for Heli Skottman for getting the stem cell line data for this study, and Riitta Lahesmaa for the facilities for this research work.

References

[1] Affymetrix. Array design for the genechip human genome u133 set. Technical report, Affymetrix, 2003.

[2] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[3] L. Gautier, M. Moller, L. Friis-Hansen, and S. Knudsen. Alternative mapping of probes to genes for affymetrix chips. *BMC Bioinformatics*, 5(1):111, 2004.

[4] W. Huber and R. Gentleman. matchprobes: a Bioconductor package for the sequence-matching of microarray probe elements. *Bioinformatics*, 20(10):1651–1652, 2004.

[5] K.-B. Hwang, S. W. Kong, S. A. Greenberg, and P. J. Park. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, 5:159, 2004.

[6] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[7] B. Mecham, G. Klus, J. Strovel, M. Augustus, B. P. Byrne, D. and, D. Wetmore, T. Mariani, I. Kohane, and Z. Szallasi. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research*, 32:e74, 2004.

[8] B. H. Mecham, D. Z. Wetmore, Z. Szallasi, Y. Sadovsky, I. Kohane, and T. J. Mariani. Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics*, 18(3):308–315, 2004.

[9] J. Zhang, R. P. Finney, R. J. Clifford, L. K. Derr, and K. H. Buetow. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics*, 85:297–308, 2005.