DATA QUALITY ISSUES IN A LARGE DATASET OF PUBLISHED MICROARRAYS (META-ANALYSIS)

Martin Dugas, Hans-Ulrich Klein, Joachim Gerß and Sylvia Merk
Department of Medical Informatics and Biomathematics, University of Münster, Germany

Abstract
A dataset with 5896 samples consisting of published HG-U133A microarrays was analysed. Data quality was assessed with respect to phenotype data, microarray quality control parameters, hierarchical clustering and Y-chromosome analysis. Phenotype information has substantial deficiencies regarding completeness and data coding. There is evidence for a relevant laboratory effect. Even though a highly standardized industrial chip platform was used, metaanalysis of this microarray dataset is problematic.

1. INTRODUCTION

The amount of published microarrray data is increasing continuously, which poses new opportunities and challenges for data analysis. CAMDA 2007 provided a dataset of almost 6000 arrays from one chip platform (Affymetrix GeneChip Human Genome HG-U133A). It consists of diseased and normal human samples and cell lines collected from ArrayExpress and GEO.
In this paper we address various aspects of data quality of this dataset - both regarding phenotype information and the microarray data itself.

2. METHODS

Data was processed using R [1], Bioconductor [2] and SAS [3]. We used a debian linux system with 64 GB Ram and 4 dual-core processors.
Phenotype information was analysed with frequency tables and manual curation.
Bioconductor routines were applied to calculate Affymetrix quality parameters, in particular percentage of present calls and 3' to 5' ratio (AFFX-HSAC07/X00351). A low present call rate may indicate limited chip quality, but also depends on tissue properties. High 3' to 5' ratios refer to low RNA quality (commonly used threshold value: 3).
Hierarchical clustering was performed with R-routines using average linkage and euclidean distance.
To analyse gene expression on the Y-chromosome, we selected 33 probesets (x_at and s_at-probesets were omitted). We calculated t-statistics male versus female for these probesets and selected the probeset with highest t-value.
Differential genes were determined with multtest package from Bioconductor.
We further analyzed the identified differential genes by investigating a possible laboratory effect of the documented gene expression. Because explicit laboratory information was not available, we used experiment information instead. Analysis of variance models were set up, including the laboratory originating a certain sample as a fixed effect. We calculated adjusted R-square values of the fitted models in order to quantify the percentage of variation of gene expression that is explained by the laboratory. Analyses were performed separately for each gene and for the subgroup of normal and tumor tissue samples, respectively.

3. RESULTS

Analysis of phenotype

Overall, the dataset comprises 5896 samples from 252 experiments. The number of samples per experiment has a skewed distribution with an average of 23.4 samples and a median of 8. There are 1142 cell line samples and 4754 organism part samples. Cell type is missing for 4419 samples and has 121 different categories; distribution of cell type is skewed with an average of 12.2 samples per cell type and a median of 3.

Disease state is missing for 1868 samples and has 193 textual categories with a skewed distribution (average 20.9 samples per category, median 8). In addition, many of these categories are (partially) synonymous. For instance, there are 8 different texts for "breast cancer" like "breast tumor" or "breast carcinoma". By manual curation, consolidated categories for normal, tumor, leukemia, breast cancer and colon cancer were established. Disease stage is missing for 4932 samples and has 17 categories with a skewed distribution (mean 56.7, median 16 samples per category).

Developmental stage is missing for 5095 samples. BioSource type is missing in 2809 samples and has 12, partially synonymous categories. There are 97 different cell lines with a skewed distribution of categories (mean 11.8, median 5 samples per cell line).

Gender is missing for 4221 samples and has 7 (!) categories. Organism part is missing in 3056 samples and has 190 categories (most frequent: bone marrow with 607 samples).

Microarray quality control parameters

Figure 1 presents percentage of present calls and 3' to 5' ratio for each microarray by size of experiment. It becomes evident that there is a substantial variability of these quality parameters within the dataset. In addition, there appears to be less variability and better overall values of quality parameters (= high percentage of present calls and low 3' to 5' ratios) in larger experiments.

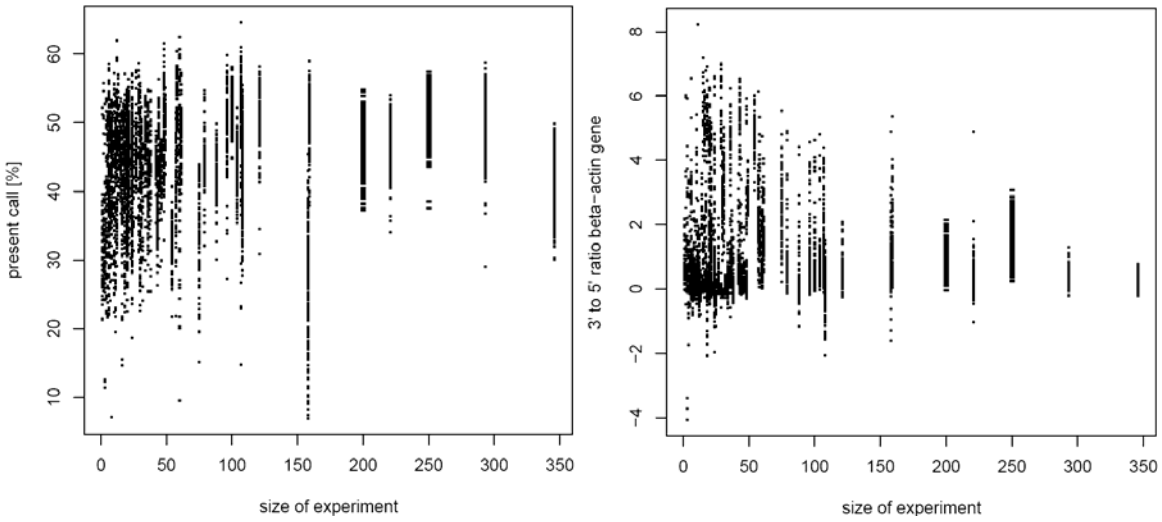

Figure 1: percentage present call and 3' to 5' ratio by size of experiment

Hierarchical clustering

Hierarchical clustering with color-coding of samples belonging to the same experiment revealed that many samples are clustered by experiment. Inspection of the distance matrix revealed that there are duplicate samples in the dataset (distance 0). We identified 498 duplicated samples. Interestingly, within these duplicates we found cel-files with identical content but different file name.

Y-chromosome analysis

From available probesets located on the Y-chromosome we selected 205000_at, because it was most differentially expressed between males and females according to a t-statistic criterion. Figure 2 shows frequency distribution for males (n=907) and females (n=424) for this probeset. There are 32 of 424 females with expression level above 4, even though they are supposed to have no Y-chromosome.

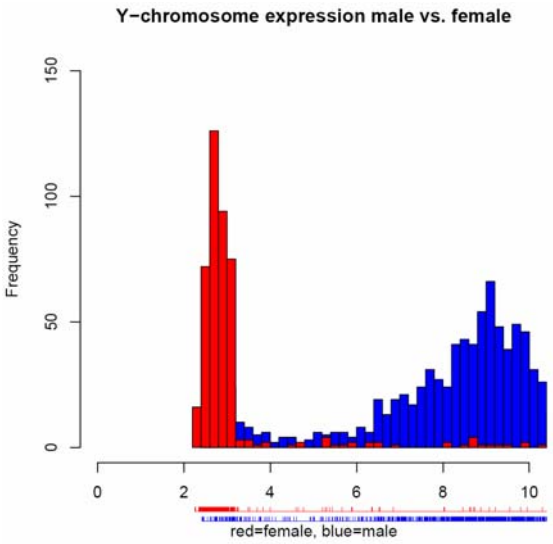

Figure 2: Histogram of Y-chromosome expression (205000_at)

Tumor signature

To estimate the order of magnitude of a laboratory effect, we performed the following analysis: We removed duplicate samples from the dataset and defined by manual curation a "tumor" and a "normal" group of samples, because most diseases in the dataset are various types of tumors. 2335 samples could be assigned to tumor. Only 83 non-duplicated samples were annotated as "normal" or "healthy", therefore samples without disease annotation were considered normal. We excluded all cell lines from "normal", because many cell lines are tumor cell lines. By this approach 1160 samples were assigned normal.

Using Bioconductors multtest we determined differential genes tumor versus normal.

By ANOVA analysis we estimated the explained variance by the laboratory for differential genes. Table 1 presents adjusted R-square values of one-factorial ANOVA models investigating the laboratory effect for top 10 differential genes. The factor laboratory explains at least 60% of variance in normal samples. The corresponding values of tumor tissue were lower by about 10 percentage points. Presumably this is due to a higher amout of variation among tumor tissue samples, that results from different tumor entities included. Normal tissue samples appear to be more homogeneous.

Adjusted R^2	Tumor	Normal
201417_at	0.5681464	0.6234135
213668_s_at	0.5480438	0.6587130
210719_s_at	0.5719505	0.6030735
213491_x_at	0.4805231	0.6227067
212115_at	0.6684609	0.6602535
203462_x_at	0.4515008	0.6023647
213399_x_at	0.5003343	0.6225354
201416_at	0.6475034	0.7435285
208688_x_at	0.4218399	0.6366382
208650_s_at	0.5651868	0.6053727

Table 1: Adjusted R-square values of one-factorial ANOVA models including the laboratory effect for top 10 differential genes

4. DISCUSSION

The available phenotype information for the META-analysis dataset has major deficiencies, in particular many missing values and non-coded information. Missing gender information in 71.5% of cases and 7 categories for gender highlights the need for data monitoring to improve data quality.

Free text description of diseases apparently is not suited for large-scale datasets: There are 193 textual categories for diseases with many synonyms, for instance 8 different texts for breast cancer, and many ambiguous terms (e.g. colorectal tumor: benign or cancer?).

From an analysis perspective, precise disease classification is needed. International coding schemes like international classification of diseases (ICD [4]) should be applied. In addition, disease and stage of each disease should be clearly separated. For instance, lung adenocarcinoma is listed as 4 different disease entities, corresponding to stage I to IV. Frequencies of several item categories are highly skewed; this restricts data analysis. Again, coded data values based on established classifications instead of free text would be helpful to enable merging of similar categories or exclusion of samples.

All experiments used the same chip type and presumably followed manufacturer's instructions. However, analysis of quality control parameters like percentage present calls and 3' to 5' ratio revealed substantial variability. This can be attributed to the diversity of biological samples as well as differences between laboratories. There seems to be a trend that experiments with larger sample sizes have better QC parameters.

Y-chromosome analysis also sheds some light on data quality issues. Even for such clear cut questions like "patient has Y-chromosome - yes or no?" there is no easy answer. There are more than 30 probesets for Y-chromsomal genes on the HG-U133A chip. Some of them are almost not differentially expressed between males and females (data not shown), but even when selecting the most discriminative probeset regarding gender, there is a relevant discordance with gender information. This might be a chip problem or a phenotyping problem and should be subject to additional validation.

From a medical viewpoint, our tumor signature is very probably invalid, because different tumor entities are not represented proportionally within the dataset. Controlled studies are needed to establish such tumor signatures ([5],[6]). A sufficient number of valid control samples is important for data analysis. Duplication and

multiple usage of the same cel-file is probably not a valid approach - at least it must be transparent to the user.
The aim of our heuristic approach was to estimate the order of magnitude of the laboratory effect. Consistent with hierarchical cluster analysis, the laboratory effect seems to be relevant even for a highly standardized industrial chip platform and therefore should not be ignored. Tissue type may be a confounder of laboratory effect and experiment effect, but unfortunately could not be entered into the model due to insufficient phenotype coding.
Overall, data quality both with respect to phenotype information and microarray data remains an important issue, in particular when we aim to combine results from different microarray experiments.

5. REFERENCES

(1) R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0]
(2) Bioconductor. http://www.bioconductor.org (accessed 18 October 2007)
(3) SAS. http://www.sas.com (accessed 18 October 2007)
(4) International classification of diseases. http://www.who.int/classifications/icd/en/ (accessed 19 October 2007)
(5) Dugas M, Weninger F, Merk S, Kohlmann A, Haferlach TA. Generic Concept for Large-scale Microarray Analysis Dedicated to Medical Diagnostics. Methods Inf Med. 2006;45(2): 146-152.
(6) Haferlach T, Kohlmann A, Schnittger S, Dugas M, Hiddemann W, Kern W, Schoch C. Global approach to the diagnosis of leukemia using gene expression profiling. Blood. 2005;106(4):1189-98.