1 METHODS AND DATASETS

Sequence-based mask

All probes were mapped to human and chimp genomes (human genome hg18, chimp genome version 2.1.45 and cDNA (from NCBI36.45 build) with the PatMan (Prüfer et al., 2008) algorithm, allowing up to one mismatch. In order to have exact error estimates for our mask, we conservatively defined probes as having a sequence difference only when a probe had exactly one perfect match to the genome in one species and in another species lacks a perfect match, but had exactly one match with a single basepair difference. We defined probes without any sequence difference as probes with exactly one perfect match in both species. If there was no match to the genome sequence, then any number of matches in cDNA were also treated as a match.

Other masks

All expression based masks and the following computations were implemented in R, with help of Bioconductor “affy” package Gentleman et al. (2004). To construct a mask with the (Greenhall et al., 2007) method we used their webpage interface. In mask construction and in downstream analysis we use only probesets expressed in both groups, where a probeset is defined as expressed if there is a max5 “P” call in at least 90% of samples in each group.

Errors in detecting BAD probes

We calculated the ratio of not detected BAD probes to all BAD probes (type 1 error) and the fraction of probes not differing in predicted binding affinity, detected as such (type 2 error). For the human-chimp dataset we used sequence-based mask as binding affinity prediction, for simulated datasets we used the information of which probes were flipped.

Probes with high-low GC content

Probes (all, or probes with desired middle nucleotide) were ordered according to their GC content and first/last 20% of them were used to simulate BAD probes.

Analysis of gene expression differences

Gene expression was estimated with standard RMA method and quantile normalization, with affy Bioconductor package (Gautier et al., 2004). Differentially expressed probesets were found by a t-test, at a significance level 0.05 (lower cutoffs, and using instead a Wilcoxon rank test gave very similar results) We define as type 2 errors the fraction of probesets differing in expression between groups in the original data, which were not detected as differentially expressed after simulating BAD probes and masking. We define as type 1 errors the fraction of probesets not differing in expression in the original dataset that are detected as differentially expressed after simulation and masking.

Other methods tested

In constructing the method described in the paper, we tested several other methods for their ability to detect sequence differences. The methods tested were the following:

1. ANCOVA. The basic question we used the ANCOVA for is whether it is significantly better to explain the underlying data with two groups, two regression lines - or with just one group - one regression model. The model is given by

\[ y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon \]  

where \( x \) and \( y \) are the probes, \( z \) is a group label. The ANCOVA compares the variances of the complete model and the model where \( \beta_2 = \beta_3 = 0 \).

2. Methods fitting a single model without groups We used different methods using only one regression model. The regression fit is done based on the whole data, not taking into account the groups (e.g. species assignment). One method is to compare the distances from the datapoints to the regression line taking into account whether a point is above or below the line by giving it a negative or positive distance. The distances are compared by a t-test between the two groups. A second method is to look at the amount of points above and below the line for each group. Assuming a binomial distribution behavior one can test the observed ratio of points above the line against the null hypothesis that the expected frequency is 50%.

3. Other methods We also tested several other ad hoc methods: Comparing the slopes of the linear fit between the two groups. A method based on the Principle Component Analysis (PCA). Comparing the distance of the two linear fits for each group to the fit for the whole data.

Furthermore, we also tested if background correcting and/or normalizing the raw data is beneficial, and found that the raw fluorescence levels gave us the best detection rates as compared to sequence data.

Datasets

All datasets used are publicly available. Khaitovich et al. (2004) human-chimpzee dataset: Array Express E-ABM4-2.

Single-tissue dataset

(Ryan et al., 2006) GEO GSE5392 healthy individuals from bipolar disorder study postmortem brain tissue (orbitofrontal cortex)

| GSM123212 | GSM123213 | GSM123214 | GSM123215 |
| GSM123216 | GSM123217 | GSM123218 | GSM123219 |
| GSM123220 | GSM123221 | GSM123222 | GSM123223 |
| GSM123224 | GSM123225 | GSM123226 | GSM123227 |
| GSM123228 | GSM123229 | GSM123230 | GSM123231 |
| GSM123232 | GSM123233 | GSM123234 | GSM123235 |
| GSM123236 | GSM123237 | GSM123238 | GSM123239 |
| GSM123240 | GSM123241 | GSM123242 |

Two-tissue dataset

(Hodges et al., 2006) GEO Series GSE3790.

Caudate nucleus samples

| GSM86816 | GSM86817 | GSM86818 | GSM86819 |
| GSM86820 | GSM86821 | GSM86822 | GSM86824 |
| GSM86825 | GSM86826 | GSM86827 | GSM86828 |
| GSM86829 | GSM86831 |
Mouse dataset
(Hovatta et al., 2005). GEO Series GSE3327.
A/J:
GSM74915 GSM74916 GSM74929 GSM74930
GSM74943 GSM74944 GSM74945 GSM74957
GSM74958 GSM74972 GSM74973 GSM74987
GSM74988 GSM74999 GSM75000
C57BL/6J:
GSM74917 GSM74918 GSM74931 GSM74932
GSM74946 GSM74947 GSM74959 GSM74960
GSM74974 GSM74975 GSM74989 GSM74990
GSM75001 GSM75002

Supplementary Figures

Fig. 1: An example of a BAD probe in the mouse dataset. Each point is a sample, green for A/J, blue for C57BL/6J. On the left, fluorescence level for probe 2 vs. probe 7 in probeset 94486_at. No apparent affinity difference can be seen between blue and green dots for these probes. On the right, probe 6 vs. probe 2 for the same probeset. Even though none of the probes contains a known SNP between the strains, it is obvious that there is a strong binding affinity difference for probe 6 between A/J and C57BL/6J.

Fig. 2: Detection of sequence differences and their location within a probe. The fraction of sequence differences that are detected as BAD probes is plotted against the position of the sequence difference along the probes. The graph shows the result for masking 30% of the probes from the human-chimpanzee dataset; 95% confidence intervals are marked with dashed lines.

Fig. 3: Detection of BAD probes and their nucleotide constitution. Type 1 error is the fraction of probes without sequence differences, but detected as BAD. Type 2 error is the fraction of probes containing a sequence difference, but not detected as BAD. The probes from the two-tissue dataset were divided: (a) in quartiles, by the GC content of a probe, and (b) according to the middle nucleotide of a probe. The 95% confidence intervals are marked with dashed lines; the human-chimpanzee dataset is shown in black.
Fig. 4: Effect of the number of individuals used to build a mask on its accuracy. Type 1 vs. type 2 error for masks constructed with a different number of individuals per group for the human-chimpanzee dataset.

Fig. 5: Effect of the number of individuals used to build a mask based on differential expression detection. Errors in detecting differential expression are marked against mask stringency. Black dot: perfect mask (information about all probes simulated as BAD) (a) Single-tissue dataset, fraction of the genes that were originally classified as non-differentially expressed that are classified as differentially expressed after masking. (b) Two-tissue dataset, fraction of the genes originally differentially expressed that are classified as non-differentially expressed after masking. (c) Two-tissues dataset, fraction of the genes that were originally non-differentially expressed that are classified as differentially expressed after masking.
Fig. 6: Effect of the number of probes simulated as BAD on their detection. (a) Single-tissue dataset (b) Two-tissues dataset.

Fig. 7: Performance of our method compared to the previously published Greenhall et al. method. Type 1 (fraction of probes without sequence differences, but detected as BAD) and type 2 (fraction of probes containing a sequence difference, but not detected as BAD) errors for the human-chimpanzee dataset, when masked using each method.

Fig. 8: Comparison of several masking methods evaluated for this paper. Type 1 (fraction of probes without sequence differences, but detected as BAD) and type 2 (fraction of probes containing a sequence difference, but not detected as BAD) errors for the human-chimpanzee dataset, when masked using each method.

Fig. 9: Effect of number of probes per probeset on the error rates of detecting bad probes. We artificially created probesets with fewer probes in the human chimpanzee dataset, and measured the error rates vs. sequence data.
Fig. 10: Frequency distribution of the number of probes left after masking in three datasets. (a) Human-chimpanzee brain dataset. This set uses the U95A array, so all probesets start with 16 probes. (b) Single-tissue dataset. (c) Two-tissue dataset. In both datasets most probesets have 11 probes, and only a few have 16.
REFERENCES


