

Summary of numerical methods used on Affymetrix chips, public version.

rork@umich.edu

Our group has used a variety of Affymetrix ([Affy top page](#)) chips. This document describes our recent methods for estimating probe-set signal strength, for normalizing chips to each other, and for annotating probe-sets, and has links to all the software, including the code, free for almost any use by anyone. Some recent publications using these methods are listed at [PubList](#). Ongoing alterations in Affymetrix's own algorithms and annotation should also be considered by people interested in these methods -see their web pages. (We have made some comparisons of our methods with Affymetrix MAS 5.0 software, and 5.0 is much better than their previous version, but not so good that we want to switch to it.) Note that our normalization and annotating software is not specific to Affymetrix chips, and could be useful for other analyses.

HuFL (Hu6800, HuGeneFL) chips consist of 287296 24x24 um features, which are 25 base long single stranded DNA. After hybridization steps the chips are scanned at 3 um per pixel resolution. For each feature, our understanding is that the Affymetrix software ignores a 1 pixel border, and the 75-th percentile of the remaining pixels is stored in a file (.CEL files). There are typically 20 pairs of features (probe-pairs) on HuFL chip for each transcript (probe-set), 20 of which are designed to be complementary to a specific sequence (perfect match = PM features), and the other 20 being identical except that the central base has been altered (mismatch = MM features). On the more recent chips produced by Affymetrix (human U95 series, Mouse U74 series) there are more features, these occupying 20 um squares, and there are fewer probe-pairs per probe-set on average (16 typically). On the even more recent arrays (U133, MOE430, RAE230), the features are 18 um with 11 probe-pairs per probe-set being typical, and most recently there are designs with even smaller features. On all but the oldest chip designs the probe-pairs are scattered over the area of the chip, rather than right next to each other as they are in the HuFL, but on all chips we use the MM feature is directly below the PM feature on the chip.

We have developed software to read the .CEL files and their descriptions and perform some processing of the data (written in C). The code performs most of the calculations described below except for the quantile-normalization. Downloads and more documentation are at readaffy.html. The idea of a standard chip is used, which is usually selected to be a chip with high signals and low background. Probe-pairs for which either the PM or MM feature are saturated (pure white = large pixel values) in the image of the standard, or for

which $PM-MM < C$ on the standard are excluded from further consideration. These probe-pairs usually give highly negative PM-MM values on all chips. (C was usually chosen to be -1000 prior to a reduction of the voltage across the PMT of the scanner. For subsequent scans, -100 is a typical choice.). Saturated features (measures at least 98% of the maximum pixel value for the chip) on other chips are imputed separately for PM or MM values. For a saturated PM value the ratios of non-saturating PM values for the chip divided by the standard are averaged for a probe-set by taking the anti-logarithm of the mean of the log ratios. This factor is multiplied by the PM values of the standard to obtain imputed values for the chip under consideration. (The original values are replaced by the imputed values only if the imputed values are larger.) The MM values are imputed similarly. Doing something about saturated features was very important previously, but recently we have very few or no such difficulties.

A one-sided Wilcoxon signed-rank test is performed on the PM-MM differences to help judge if the transcript represented by the probe-set is being expressed. This is analogous to present/absent calls produced by the Affymetrix software, except that a P-value is produced rather than a call, and that the test used is completely described, simple, and conventional. (Affymetrix's new MAS 5.0 software now performs a signed-rank test as well, but is rather more complicated and should be studied carefully.)

The average intensity for each probe-set is computed as the mean of the PM-MM differences, after trimming away the 25% highest and lowest differences. This is sometimes referred to as the trimmed-mean to distinguish it from the analogous "average difference" computed by the Affymetrix software. (There is now an option to trim about 20% rather than 25%, and we have evidence that 20% is better for the most modern chip designs.)

Normalization is usually by one of two methods. For method 1, we select a set of reference probe-sets which are used to normalize each chip by adjusting by a single scale factor. For several human tumor data sets a group of 1300 reference probe-sets that were infrequently among the 5% of probe-sets with largest average intensity or 40% of probe-sets with smallest average intensity are used (less than two such occurrences in a set of 71 lung, colon, brain, and ovary tumor and normal samples). For other data sets a reference set is chosen by a similar method, or by asking that a certain minimum percentage of the chips give small P-values for signed-rank test for "present". However the set is chosen, a normalization factor is obtained using the reference probe-sets by computing the anti-logarithm of the mean log ratios of the trimmed means for the selected chip divided by the standard. In other studies reference probe-sets

can be selected by other means. We currently always use the quantile-normalization procedures described next however.

For method 2 (Kerby Shedden's quantile normalization) the distribution of trimmed-means is adjusted to more nearly match that of a standard chip by making 100 (or 20 or 50, etc) individual quantiles have the same values, using a piece-wise linear function. (The "standard chip" being normalized to can be a real chip's trimmed-means, or some artificially computed standard such as the median value for each quantile over a set of chips.) The first and last interval are normalized by fitting a regression line through these largest (or smallest) values for the two chips. This method is sometimes used after removing a set of 65 (varies with the chip) probe-sets that serve quality control purposes or give highly negative average intensity values. A more detailed description and discussion is available along with software downloads (kerby_norm.htm). Quantile normalization is performed by a separate C++ program, and thus can also be used for data obtained from other kinds of assays. On March 11 2001, Kerby put out another version of this software. "The primary extension is that you can specify a bound such that genes whose rank range exceeds the bound will not be used to calculate the quantile adjustment." We almost always have used the original Shedden algorithm. It is simple and good. More recently, other groups have adopted some similar methods, or performed quantile-normalization on the probe intensities rather than the derived probe-set intensities.

For comparisons of two chips, the software can perform two-sided Wilcoxon signed-rank tests on the differences of the PM-MM values for each probe-set, after normalization with method 1. Probe-pairs for which either chip has saturated values for PM or MM are excluded, excepting cases where a PM value is saturated on the chip that gave a larger PM-MM value (even before imputing it), in which case the imputed PM-MM value is used. These tests are somewhat analogous to certain calls obtainable from the Affymetrix software. For larger experiments that include some replication of the samples in the design we usually fit models to the quantile-adjusted (and log-transformed) trimmed-mean data rather than considering the probe-pair level data, since signed-rank tests at the probe-pair level treat the PM-MM values for probe-pairs on the same chip as independent assays, which is not true. We suggest using these signed-rank tests only on preliminary experiments to help decide if it is worth running additional samples.

For most chips there is high-quality annotation of the probe-sets made by the software of Jean-Marie Rouillard, that is usually attached to all spreadsheets. For documentation and downloads see [JMR Affy annotation](#)

Recently Affymetrix has web sites (netaffx) with good annotation and powerful query interfaces, however as of Feb 2002, these have some problems in my view, having to do with what gene is being represented by a probe-set. Our methods above are also not perfect though. See [Affy affx annot](#) .

On June 23 2004, they have finally done something better about computing the annotation. "Genomic alignments of consensus sequences were used as the primary means to assign genes to probe sets (instead of an accession-based method) for the human, mouse, and rat arrays" is all the announcement on their web said. I find nothing further on the subject.

See [SIF index.html](#) for probe-set sequences given to us by Affymetrix. This link also has files that hold the probe sequences (25 bases at a time). This link is not as critical now, since these sequences can be very conveniently obtained directly from Affymetrix web sites.

If you find any of this software useful, please give us some credit, and consider having the modified code be public as we have done.