Normalization Protocol for Morozova *et al.*

James C Costello[1,2], Mehmet M Dalkilic[2,3], Justen R Andrews[1,2,3]

[1]Dept. of Biology, [2]School of Informatics, [3]Center for Genomics and Bioinformatics
Indiana University, Bloomington, IN

Data Acquisition

Data were downloaded from Gene Expression Omnibus (GEO) as raw Affymetrix .CEL files under the accession id, GSE5382:
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5382

All data are referenced in:

> Morozova TV, Anholt RR, Mackay TF. *Transcriptional response to alcohol exposure in Drosophila melanogaster.* Genome Biol. 2006. 7(10):R95.

Affymetrix Normalization Protocol

This experiment used the Affymetrix Drosophila version 2 platform. The "fullmodel" **GCRMA** [1] normalization was run using the **Affy** and **GCRMA** packages in the **R** statistical programming environment[1], as follows:

1. Using the **readAffy()** function in the **Affy** package, read in all the .CEL files.
2. The **GCRMA** normalization method requires that an affinity score be calculated for each probe on the array. The affinity scores are calculated with respect to the melting temperature of the oligonucleotide sequence. This is done with the **compute.affinities( )** function.
3. Perform the **GCRMA** normalization with type = "fullmodel", optical correction is turned on, and the previously calculated probe affinities are included.
4. The final output of **GCRMA** is the log-transformed intensity values for the background corrected probe sets.

Affymetrix Presence/Absence calls

The presence/absence calls provide a statistical measure of the presence of a transcript within the tested biological sample. Details of the calculations can be found in the Affymetrix Microarray Suite User Guide [2]. Briefly, *p*-values for each probe set with respect to both the perfect match and mismatch probes are calculated using the Wilcoxon signed rank test. If the *p*-value is less than 0.04 the probe is considered to be present (P). If the p-value is in the interval [0.06, 0.04], then the probe is considered marginal (M), and if the *p*-value is greater than 0.06, the probe is called absent (A). The Affymetrix Expression Console software was used to make presence/absence calls. The raw data are read in this software and the **MAS5.0** normalization method was run. Of the resulting

[1] http://www.r-project.org/

output only the presense/absence calls were extracted. These calls were then mapped to the expression values output from the **GCRMA**.

Mapping Affymetrix Probe Sets to FlyBase Annotated Genes

Drosophila melanogaster sequences were downloaded from FlyBase[2]. All sequences associated with the printed probes were searched against the version 5 genome assembly of the *D. melanogaster* genome using BLASTn (E-value < 10^{-3}). BLAST results were processed with custom Perl scripts. The physical coordinates of transcripts associated with a FlyBase annotated gene were from version 5.10 of the *D. melanogaster* genome annotation. An Affymetrix probe set was mapped to a FlyBase gene ID if the BLAST results mapped to transcripts from a unique gene and only one gene. Strandedness was also considered.

Merging Probe Sets Across Replicate Arrays

There are multiple hybridizations done for the same condition and the intensity values are merged by simply taking the average for a particular probe set across the arrays under the same condition.

The *p*-values associated with presence/absence calls are treated a bit differently. They are combined using Fisher's combined probability test. The *p*-values for a probe set are combined using the following equation:

$$X^2_{2k} = -2\sum_{i=1}^{k} \ln(p_i)$$

The resulting test statistics is then interpreted using the chi-square distribution to arrive at a new *p*-value with 2*k* degrees of freedom. For probe sets on individual arrays, if the *p*-value is less than 0.04 the probe is considered to be present (P). If the *p*-value is in the interval [0.06, 0.04], then the probe is considered marginal (M), and if the *p*-value is greater than 0.06, the probe is called absent (A). The presence/absence calls are made on the newly calculated *p*-values by recalculating the thresholds according to the number of *k* arrays under one condition. For example, if $k = 4$ arrays hybridized under one condition, then the new interval becomes [0.001158, 0.004058], where $p_i = 0.04$ and $p_i = 0.06$, respectively.

References

1. Wu, Z. and R.A. Irizarry, *Preprocessing of oligonucleotide array data.* Nature Biotechnol. 2004. **22**(6):656-658.
2. Affymetrix, *Affymetrix Microarray Suite User Guide*. 2001, Affymetrix: Santa Clara, CA.

[2] ftp://www.flybase.net/genomes/Drosophila_melanogaster/