# Supporting Text

## Data Processing

### Example I (Diabetes)

This data set was studied in Mootha *et al.* (1). The expression data from HG-U133A arrays for 43 patients (17 normal glucose tolerance, 18 Type 2 diabetes mellitus) are available at `http://www.broad.mit.edu/mpg/oxphos/`, as well as the lists of probe sets belonging to the 149 pathways. For preprocessing, we eliminated those genes whose values were $< 100$ in all samples.

### Example II (Inflammatory myopathies)

This data set contains 49 HG-U133A arrays collected from multiple studies, both published and unpublished. Dermatomyositis data were described in Greenberg *et al.* (2) and are available from Gene Expression Omnibus (`http://www.ncbi.nih.gov/geo`) with accession no. GSE1551; inclusion body myositis data were repeat experiments of the same samples initially done on HG-U95A arrays, described in Greenberg *et al.* (3), with GEO accession no. GDS198. A larger study containing all 49 samples used in this paper and more will be published elsewhere. For convenience, we have collected the data into a single table showing expression values processed through Microarray Analysis Suite 5.0 (MAS5), which is available from the authors upon request. For analysis, global normalization was applied using trimmed mean (2.5% from each end of the distribution), and those probe sets whose expression values were below the mean in every sample were removed, resulting in 10,526 probe sets.

### Example III (Alzheimers's Disease)

This data set was studied in Blalock *et al.* (4). Both CEL files and processed expression values are available from GEO with accession no. GDS810. HG-U133A arrays were used, and MAS5 was used to process the data. There are 22 postmortem subjects with Alzheimer's disease at various stages and 9 controls. The same global normalization and filtering used in Example II was applied, which resulted in 11137 probe sets from the original 22,286. These genes resulted in 939 gene sets whose size is between 20 and 500 (the number of gene sets are slightly different here from Example II because the probe sets are different after filtering).

## Choice of Optimal Weights for Testing $Q_2$.

To test whether a gene set contains any genes whose expression levels are associated with the phenotype of interest ($Q_2$), we proposed a procedure based on the statistic

$$E_k = \frac{1}{m_k} \sum_{i=1}^{B} G_{ki} t_i.$$

The power of this test against a certain alternative could be improved by using more general linear combination of association measure,

$$\frac{1}{m_k}\sum_{i=1}^{B} G_{ki} w_{ki} t_i,$$

where $w_{ki}$ are appropriate weights used to combine $m_k$ test statistics. The optimal choice of $w_{ki}$ depends on the alternative we want to detect.

Suppose our goal is to detect coordinated moderate associations for genes in the $k$-th gene set. It is reasonable to assume that under such alternative, $t_{k_i} \sim N(\Delta, 1)$. With $\{k_1, \cdots, k_{m_k}\}$ denoting the index set of $m_k$ genes in the $k$-th gene set, the objective is to find the optimal weight $\mathbf{w}_k = (w_{kk_1}, \cdots, w_{kk_{m_k}})'$ for linearly combining $m_k$ test statistics to detect the small shift $\Delta$, which is too small to be detected individually, while accounting for the correlations in $t_{k_i}$. If we let $\mathbf{\Sigma}_k$ be the covariance matrix of the random vector $\mathbf{t}_k = (t_{k_1}, \cdots, t_{k_{m_k}})'$ and $\mathbf{1}$ be the $m_k$-vector with all the components being 1, then $E_k(\mathbf{w}_k) \sim N(\Delta \mathbf{w}_k'\mathbf{1}, \mathbf{w}_k'\mathbf{\Sigma}_k\mathbf{w}_k)$ under the alternative, and the most powerful test in this class is the one with $\mathbf{w}_k = \mathbf{\Sigma}_k^{-1}\mathbf{1}$ (5).

Since $\mathbf{\Sigma}_k$ is unknown and its estimator $\hat{\mathbf{\Sigma}}_k$ often is singular, we propose to use

$$E_k(\lambda_k) = \mathbf{1}'(\hat{\mathbf{\Sigma}}_k + \lambda_k \mathbf{I}_{m_k})^{-1}\mathbf{t}_k,$$

where $\mathbf{I}_{m_k}$ is the $m_k$-identity matrix and $\lambda_k > 0$ is an appropriate shrinkage parameter. The choice of $\lambda_k$ is important for the performance of the resultant test procedure. When $\lambda_k \to \infty$, $E_k(\lambda_k)$ degenerates to the simple average $E_k$. Since it is desirable to have all the weights be positive in terms of interpretability and robust performance against other alternatives, we set $\lambda_k$ as the smallest positive constant, such that all the components of vector $(\hat{\mathbf{\Sigma}}_k + \lambda_k \mathbf{I}_{m_k})^{-1}\mathbf{1}$ be nonnegative.

Once we decide $\lambda_k$ for $k = 1, \cdots, K$, we can compute $(E_1(\lambda_1), \cdots, E_K(\lambda_K))$ and approximate its null distribution by the empirical distribution of $(E_1^*(\lambda_1), \cdots, E_K^*(\lambda_K))$, where $E_k^*(\lambda_k)$ is the test statistic for the $k$-th gene set based on permuted phenotype measurements $(z_1^*, \cdots, z_n^*)$. To examine the effect of proposed weighting in $E_k(\lambda_k)$ for testing $Q_2$, we calculated $E_k(\lambda_k)$ when the analysis based on $E_k$ failed to discover sufficient number of significant gene sets. For example, we have run the parallel analysis based on $E_k(\lambda_k)$ in the diabetes example. There is no significant gene set at the $q$ value level of 0.05 based on the simple $NE_k$; on the other hand, $NE_k(\lambda_k)$ based tests identify two significant gene sets. It seems that although $NE_k(\lambda_k)$ produces similar rankings for gene sets, the introduction of weights can improve the sensitivity and power of the testing procedure.

## Generation of gene sets

We have collected the gene sets from five publicly available sources:

- Gene Ontology (`www.geneontology.org`);
- Biocarta (`www.biocarta.com`);
- KEGG (`www.genome.jp/kegg`);
- BioCyc (`biocyc.org`);
- custom arrays (`www.superarray.com`).

To map the probe sets on each array type to the gene sets, we converted each possible gene set from the databases (GO category, pathway, etc.) into a set of corresponding LocusLink IDs (now superceded by Entrez Gene). We then mapped every probe to its LocusLink IDs and then matched them against the gene sets identified in the first step. The conversion from probe ID to LocusLink ID was performed using the annotations packages available through Bioconductor

(`www.bioconductor.org`), using the statistical language *R*. Our gene sets are based on the annotations last compiled in September, 2004. Due to updates in the databases, small discrepancies should be expected at later dates.

Because the data sets used in the manuscript were Affymetrix human data, we have curated the gene sets for Affymetrix HG-U95A and HG-U133A arrays. These are available from the authors upon request as an object that can imported to the R software package. `load(''GenesetsU133a.Robject'')` will create a object `G`; `G[[i]]` can be used to view the *i*th gene set. Each gene set contains a complete annotation, including its source, corresponding LocusLink IDs, biological description, and probe identifiers.

# References


1. Mootha, V. K., Lindgren, C. M., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* (2003) *Nature Genetics* **34**, 267–273.

2. Greenberg, S. A., Pinkus, J. L., Pinkus, G. S., Burleson, T., Sanoudou, D., Tawil, R., Barohn, R. J., Saperstein, D. S., Briemberg, H. R., Ericsson, M., Park, P. J., & Amato, A. A. (2005) *Annals of Neurology* **57**, 664–78.

3. Greenberg, S. A., Sanoudou, D., Haslett, J. N., Kohane, I. S., Kunkel, L. M., Beggs, A. H., & Amato, A. A. (2002) *Neurology* **59**, 1170–82.

4. Blalock, E. M., Geddes, J. W., Chen, K. C., Porter, N. M., Markesbery, W. R., & Landfield, P. W. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 2173–8.

5. Wei, L. J. & Johnson, W. (1985) *Biometrika* **72**, 359–364.