Increased Platform Concordance by Analyzing Gene Sets

Deeper interpretations of data produced using Illumina Expression BeadChips

Introduction

Alternative approaches can enhance and empower analysis of gene expression data derived from Illumina BeadArray™ technology and can go beyond the typical 'hit-list' of over- and under-expressed mRNA transcripts between two cell types. This Technical Note provides the motivation for, and descriptions of, two powerful techniques for analyzing pre-defined sets of genes for concordant differential expression between cellular conditions. These techniques facilitate the biological interpretation of data generated using Illumina BeadArray technology and ease the difficulty of cross-study microarray evaluations.

When the 'Hit-List' is Not Enough

Results from microarray platforms that examine differences between two cell types are typically reported as two hit-lists: one containing genes relatively over-expressed in one cell type and the other listing genes over-expressed in the contrasting cell type. These lists are informative for gene regulation cataloging but at least two major drawbacks exist when reporting array-based results in solely this format. First, hit-lists may contain thousands of genes, many of which are not involved in the core biological processes being affected. This complicates interpretation, which is necessarily subjective at best. Second, it has been suggested that hit-lists generated in different laboratories using the same or different array technologies may exhibit low overlap. This latter point makes the peer review of microarray-based experiments difficult or impossible.

These drawbacks are based on the assumption that a hit-list was generated in the first place. Due to the well-documented difficulty of multiple-hypothesis testing, stringent p-value thresholds are often required of each individual gene assay within a microarray experiment. This is done to control the number of false positives reported but may reduce hit-lists to just a handful of genes, if any. Furthermore, the differential phenotypes in the experimental setup may be caused by a collection of slight, concordant changes in gene expression from the genes of a particular metabolic pathway. Individually, these small changes would require a large sample size to reveal the significant genes in the traditional hit list.

A More Inclusive Solution

Gene set analysis addresses these concerns. In gene set analysis, predefined collections of genes are tested for significant concurrent aberration in lieu of directly testing individual genes. These sets may correspond to genes within the same cytological band, genes that belong to a particular signaling cascade, or even genes that were identified in a previous microarray study (possibly even using a different gene expression platform). Two important implementations demonstrating the utility of this approach are Gene Set Enrichment Analysis (GSEA)[1] and Parametric Analysis of Gene Set Enrichment (PAGE)[2].

GSEA begins by importing a list of test statistics, such as Student's t, RMA, and SAM q-value, for each gene between two cellular conditions. This gene list is sorted by the magnitude of the genes' test statistics. Then, all the genes for a pre-defined gene set are identified within the ranked list. Next, a running sum is computed from the top of the list to the bottom. This is computed by increasing the sum by an amount determined by nearness to the top of the list when a gene within the gene set is encountered, and decrementing the sum by a small amount otherwise. The gene set's enrichment score is computed as the maximum value of the running sum thusly calculated. The significance of the enrichment score is estimated by permutation testing, which simulates the null hypothesis of an even distribution of genes within the set throughout the ranked list.

PAGE is a much simpler procedure, yielding similar results as GSEA. As in GSEA, the first step is to import a list of genes along with some test statistic that discriminates between two biological classes. Next, the genes of a predefined gene set are isolated from this list. The mean of these genes' test statistics is computed and its significance assessed utilizing the calculus central limit theorem and the standard normal distribution. PAGE has the advantage of reduced computing time relative to GSEA since it does not require costly permutation testing to gauge significance.

Example Application

To illustrate the utility of GSEA and PAGE in a cross-platform comparison, two datasets were downloaded from the recent set of Microarray Quality Control Consortium (MAQC) experiments[3]. One of these datasets was generated using Illumina BeadArray technology and the other was obtained with the Affymetrix GeneChip technology. Both datasets examined the differences between human brain RNA and a control human universal RNA sample.

Each of 12,091 genes individually were tested for significant overexpression in the brain RNA samples on both platforms. We found generally good overlap between the two platforms: 3,523 genes were identified by the Illumina Expression BeadChips, 3,742 were identified by Affymetrix, and 3,067 were identified by both platforms. Given the number of genes found by each platform, one may have expected an overlap of 1,090 genes by chance. So, by scoring genes individually we achieve a three-fold boost over random chance.

GSEA was also used to analyze 521 pathways for significant upregulation in each the two microarray datasets. When testing with the data generated using Illumina Expression BeadChips, we found fifteen significant pathways. When testing with the data produced using Affymetrix arrays, fourteen pathways were identified. Importantly, when the intersection of the two sets of identified pathways was applied, nine pathways remained significant. If a set of fifteen pathways were selected at random, and then another set of fourteen pathways were

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

randomly and independently selected, the expectation would be zero or one pathway in common. Therefore, GSEA gives us a platform overlap nine times better than random chance. This is significantly better than the three-fold enrichment previously found at the gene level.

PAGE analysis of the same 521 pathways was carried out. Fourteen pathways coordinately up-regulated in brain samples were found when using the Illumina platform and twenty significant pathways were identified when using the Affymetrix platform. After the intersection of the platforms' pathways were taken, we were left with thirteen significant pathways. We would have expected zero or maybe one pathway given random pathways—a thirteen-fold enrichment over happen chance. This represents a further improvement over the three-fold betterment achieved when analyzing genes individually.

How to Use The Gene Set Approach

Software implementing GSEA can be obtained from the Broad Institute at http://www.broad.mit.edu/gsea/. Broad distributes the program as a Java application that requires the Java Runtime Environment (JRE) be installed on your computer. If you don't already have this installed, a self-installer which includes JRE is available from the same site.

Launch the GSEA software from the GSEA desktop icon. Once GSEA is running, input your BeadArray-derived data. This is accomplished by first clicking on the 'Load Data' icon within the GSEA window and then browsing for the various files in your dataset. Your data must be formatted in a certain way; directions for this are available at the GSEA website. Once all your data are loaded into memory, click on the 'Run GSEA' icon. This opens a panel where you select which of your loaded datasets you want to analyze with GSEA, and adjust some of the algorithm's default parameters if desired. Finally, click 'Run' to apply GSEA to your BeadArray expression data. GSEA conveniently outputs the results in a set of HTML files that you can browse in a web browser.

To facilitate the use of PAGE, its authors provide a Python script for your convenience. This is available by contacting the original publication's first author. To use the program, you must first have Python installed. Python is available for all major operating systems and can be freely downloaded at http://www.python.org/. Once Python is installed on your computer and you've obtained the PAGE software, run it by typing 'python page.py <DATA> <GENE SETS> <OUTPUT>' where <DATA> is your scored BeadArray data, <GENE SETS> is a file defining gene sets to be analyzed, and <OUTPUT> is the file name for the results file. Your scored data consists of two columns: the first contains gene symbols and the second contains statistics indicative of differential expression. The gene set file is conveniently of the same format as GSEA's and can be found at Broad's GSEA website. The results file, which is easily imported into Microsoft Excel, will give you a score for each pathway tested with PAGE.

Summary

Using GSEA to analyze BeadArray-derived data complements the traditional hit-list reporting strategy. GSEA has the following advantages:

- **Testing gene sets automatically moves the analysis towards biological themes, thus accelerating the discovery process.**
- **The problem of inter-study reproducibility is more directly addressed.**
- **Multiple hypothesis testing is lessened due to testing a reduced number of gene sets to test rather than genes. If desired, the focused testing of genes within an identified gene set can provide the same resolution as the hit-list, but with the benefit of a diminished false positive component.**
- **Small, concordant changes in gene expression within a pathway can be identified, even if none of their effects are observable with the study's preset sample size.**

Software implementations of gene set analysis are freely available and ready for Illumina Expression BeadChip-based applications.

References

Implementations

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102: 15545-50. Epub 2005 Sep 30.
2. Kim S-Y, Volsky DJ (2005) PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics 6: 144.
3. MAQC Consortium (2006) The MicroArray Quality Control (MAQC) project shows interplatform reproducibility of gene expression measurements. Nature Biotechnol 24(9): 1151-1161.

Studies Using Gene Set Analysis

Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, et al. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer.
Cell 114:323-34.

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Shihad S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34:267-73.

Shepard JL, Amatruda JF, Stern HM, Subramanian A, Finkelstein D, et al. (2005) A zebrafish bmyb mutation causes genome instability and increased cancer susceptibility. Proc Natl Acad Sci U S A 102: 13194-9, Epub Sep 6, 2005.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 470-2007-001 Current as of 13 September 2007

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA