

Two-Group Comparisons with Permutation Analysis for Differential Expression (PADE)

This Technical Note provides an overview and detailed description of the standard output that Expression Analysis provides for Group Comparisons. It employs a new method that supersedes and improves on the method of using p-values to determine statistical significance of differential expression.

Expression Analysis provides a two-group comparison analysis to meet the needs of those clients wishing to detect and estimate changes in expression between two experimental groups that are each represented by multiple (≥ 3) specimens or arrays per group, as in a replicated experiment.

The two-group comparison is best for experimental designs that involve a baseline group and an experimental group that may represent control vs. treated specimens, or specimens that are negative vs. positive for an outcome of interest. Our revised two-group comparison now incorporates a permutation analysis for differential expression (PADE). This analysis greatly helps to mitigate false positives (falsely detecting differential expression), a very important consideration when analyzing whole genome chips where potentially tens of thousands of statistical tests are examined in parallel (i.e., one test for each transcript). This permutation analysis allows us to estimate the False Discovery Rate (FDR) for a set of potential differentially expressed transcripts, which are often the most useful summary for comparisons of this type.

This comparison does not account for other experimental design features that may exist such as relevant cofactors or a time course. It can still be used in these situations by comparing a series of appropriate subsets of specimens. However, it is not likely to be the most powerful test available in such cases. More customized or exploratory analyses can be conducted individually through our inhouse statistical consulting services.

Description and Content

The comparison results are provided in an Excel spreadsheet with one row per probeset (transcript). An example excerpted from the top portion of a spreadsheet is shown on the next page (in this Technical Note, data lines are wrapped to allow for easier viewing--in the actual report file, all information pertaining to a transcript is contained in a single row).

In brief, the spreadsheet provides:

- The Affymetrix probe set identifier, which can be linked to other databases containing information on each gene, as well as other annotation information in additional columns;
- An overall estimate of transcript abundance for the specimens in each group (control and experiment);
- An estimate of the \log_2 ratio and, equivalently, the fold change based on the control group relative to the experiment group, with a negative \log_2 ratio/fold change indicating a reduction in abundance from control to experiment group. [Note: in many contexts, providing \log_2 ratio values are preferable to fold change since a log ratio is both monotonically related to fold change and the log ratio is continuous, while fold change is "undefined" on $(-1, 1)$ and a fold change of -1 and 1 are equivalent (no change). Using base 2 (\log_2) is preferable to other bases since a \log_2 ratio = 1 indicates a change in expression to 2-fold or twice the level of expression as the control group;]
- PADE (permutation analysis for differential expression) output:
 - Delta - the observed absolute difference between the observed and expected test statistic from permutation analysis
 - A series of transcript sets of increasing size corresponding to an estimated False Discovery Rate group
 - An estimate of the False Discovery Rate (FDR) for each set
 - An estimate of the marginal FDR for the transcripts unique to set n versus set $n-1$.
- A p-value using a two-sample t-test for those transcripts where individual tests are needed. a



Spreadsheet Layout

| Probe Set ID | Control Group Signal (MAS5) | Exp. Group Signal (MAS5) | Est. Log Ratio | Raw Est. Fold Change | Affy PA Calls Cntrl -Exp | Delta (PADE) | Set (PADE) | FDR for Accumulated Sets (PADE) | FDR for Specific Set (PADE) | Individual Transcript p-value |
|-------------------|-----------------------------|--------------------------|----------------|----------------------|--------------------------|--------------|------------|---------------------------------|-----------------------------|-------------------------------|
| Group size | 3 | 3 | | | | | | | | |
| 101918_at | 137.6 | 1128.0 | 3.04 | 8.20 | PPP-PPP | 2.1 | 1 | .012 | .012 | 0.00342 |
| 94057_g_at | 24.2 | 517.6 | 4.42 | 21.39 | AAA-PPP | 2.1 | 1 | .012 | .012 | 0.00364 |
| 101578_f_at | 97.7 | 772.2 | 2.98 | 7.90 | PMP-PPP | 2.1 | 1 | .012 | .012 | 0.00250 |
| 100300_at | 208.9 | 1473.5 | 2.82 | 7.05 | PPP-PPP | 2.1 | 1 | .012 | .012 | 0.00205 |
| 100444_at | 454.4 | 65.2 | -2.80 | -6.97 | PPP-PPP | 2.0 | 2 | .035 | .065 | 0.00632 |
| 94296_s_at | 48.8 | 366.3 | 2.91 | 7.51 | PPA-PPP | 2.0 | 2 | .035 | .065 | 0.00920 |
| 99459_f_at | 19.0 | 263.9 | 3.80 | 13.89 | AAA-PPP | 2.0 | 2 | .035 | .065 | 0.00406 |
| 101871_f_at | 24.6 | 191.6 | 2.96 | 7.79 | AAP-PPP | 1.9 | 3 | .042 | .091 | 0.00050 |

| Gene Descriptor | Unigene | Entrez Gene | Gene Symbol | Seq Derived From | GO Bio Process |
|------------------------------------|-----------|-----------------------|-------------|------------------|-------------------------------------------------|
| transforming growth factor, beta 1 | Mm.9154 | 21803 | Tgfb1 | AJ009862 | 42306 // regulation of protein-nucleus ... |
| stearoyl-Coenzyme A desaturase 1 | Mm.267377 | 20249 | Scd1 | M21285 | 6633 // fatty acid biosynthesis // inferred ... |
| actin, beta, cytoplasmic | Mm.297 | 11461 | Actb | M12481 | 7010 // cytoskeleton organization and ... |
| cytochrome b-245, beta polypeptide | Mm.200362 | 13058 | Cybb | U43384 | 6118 // electron transport // inferred from ... |
| cyclin-dependent kinase 5 | Mm.4818 | 12568 | Cdk5 | D29678 | 7160 // cell-matrix adhesion // inferred ... |
| general transcription factor II I | Mm.22593 | 14886 | Gtf2i | AF043220 | 6355 // regulation of transcription, DNA-... |
| ELKL motif kinase | Mm.258986 | 13728 | Mark2 | X70764 | 6468 // protein amino acid phosphorylation . |
| immunoglobulin heavy chain 4 ... | Mm.259062 | 16017 | Igh-4 | X88902 | 6959 // humoral immune response // ... |

| GO Cell Component | GO Molecular Function |
|-----------------------------------------------|-------------------------------------------------------|
| 5615 // extracellular space // traceable ... | 8083 // growth factor activity // inferred from ... |
| 16021 // integral to membrane // traceable | 5506 // iron ion binding // inferred from ... |
| 5856 // cytoskeleton // inferred from ... | 5198 // structural molecule activity // inferred ... |
| 16021 // integral to membrane // ... | 5216 // ion channel activity // inferred from ... |
| 5829 // cytosol // inferred from direct ... | 5515 // protein binding // inferred from ... |
| 5667 // transcription factor complex // ... | 3700 // transcription factor activity // inferred ... |
| inferred from electronic annotation | 16740 // transferase activity // inferred from ... |
| 16021 // integral to membrane // inferred ... | 3823 // antigen binding // inferred from ... |



More Detailed Descriptions of Columns

- The control and experimental group signal values are computed using log signal averages of the specimens from each group which are then converted back to signal units (i.e., they are geometric averages);
- PADE
 - o Delta - The derivation of Delta (Δ) is described in more detail in the Methods section. It is a threshold for determining differential expression and is also used to associate the PADE spreadsheet data with the differential expression (DE) and FDR graphs.
 - o Set – In the PADE analysis, Affymetrix probes sets are analyzed in sets of transcripts, rather than individually. The PADE transcript sets are incrementally labeled 1, 2, 3 ... These sets are distinct but can be accumulated consecutively for some calculations.
 - o False Discovery Rate (FDR) – The ratio of the typical number (over all permutations) of transcripts “discovered” at random (false discoveries) versus the number detected in the experiment. Lower values indicate more evidence for differential expression.
 - o FDR for Accumulated Sets – This value is overall transcript FDR for a group of consecutive transcript sets starting with set 1. For example, the Accumulated FDR related to set 3 is the overall FDR for transcripts in sets 3, 2, and 1 combined.
 - o FDR for Specific Set – This value is the FDR for those transcripts in the specific set given the previous “discoveries” of the differential transcripts in the lower numbered sets. It is by definition always greater than or equal to the FDR for Accumulated Sets. Note that for Set 1, the FDR for Specific Set is equal to the FDR for Accumulated Sets.
- p-values - These values indicate the level of statistical significance when individually testing the equality of the mean expression of the control and experimental populations (equal variances are assumed). Lower values indicate more evidence for differences in the means of the two populations. The p-value indicates the probability that the observed differences between the groups has occurred purely by chance alone, assuming no real difference between the groups. Often a p-value of 0.05 or less is taken to represent statistical significance; however, note that by this rule one would expect 5% of the transcripts with no real change in expression to be incorrectly classified as differentially expressed. When testing 40,000+ transcripts simultaneously, as is often the case with whole genome arrays, this process could lead to gene lists having a size of 2000 or more transcripts equivalent in quality to a list of transcripts chosen at random. This difficult aspect of traditional statistical testing is the primary reason for providing PADE: to create a list of transcripts where one has an estimate of the quality of the list of transcripts. One cannot “combine” p-values to perform the same function.
- Annotation of Probe Sets:
 - o UniGene – The identifier number associated with each probe set.
 - o Entrez Gene – The identifier number associated with each probe set. This entry is hyperlinked to the NCBI website for that particular probe set.
 - o Online Inheritance in Man (OMIM) – This entry is the Identifier number in the OMIM database. It is included only for human probe sets
 - o Sequence Derived From – The accession number of the single sequence or representative sequence used in probe selection. This identifier is typically a GenBank accession number or it may be from a species-specific databank.
 - o Gene Ontology (GO) – Three GO categories are provided: biological process, cellular component and molecular function. Each GO annotation consists of three parts –“accession number // description // evidence”. Multiple annotations for the same probe set are separated by ‘//’

All gene annotation information is provided by Affymetrix and updated quarterly (http://www.affymetrix.com/support/technical/manual/taf_manual.affx). Note that the first data row of the spreadsheet indicates the number of arrays or replicates included in each group. The results are automatically sorted by the transcripts sets created from the permutation analysis. Within the permutation analysis they are sorted by absolute log ratios.

Methods

The following information is provided for the benefit of those clients who require a more detailed accounting of the precise methodology used in this analysis.

The two-group comparison analysis is conducted on normalized expression values that are individually transformed using the base 2 logarithm of the relevant expression index (e.g., MAS5). A floor of 1.0 is used to avoid negative transformed values $\{\log_2(\max(1.0, \text{Signal}))\}$.



After computing the mean for each group using the log transformed signal values, the mean is then converted back to the original-probeset signal units. Averaging on the logarithm scale and then inverse transforming (also known as the geometric averaging) provides a more robust estimate of overall expression that is less impacted by outliers or skewed expression levels relative to a simple arithmetic average of the raw expression values from each array.

Raw fold change is calculated as the simple ratio of overall expression values from the two groups. The higher overall expression is divided by the lower overall expression. If the baseline group expression is higher, then fold change is designated to have a negative value. [Note: for some expression indices such as MAS5, very high fold change in absolute value when the average of both groups is small (< ~100 on arrays with a mean signal of 500) tend to be very unreliable unless you have a large sample size (e.g., > 10) in each group]

PADE is a specific implementation of permutation testing dealing with the very large dimensional aspect (transcripts) to each individual sample (array). In general, a statistical reference distribution is constructed of the transcripts using the following test statistic:

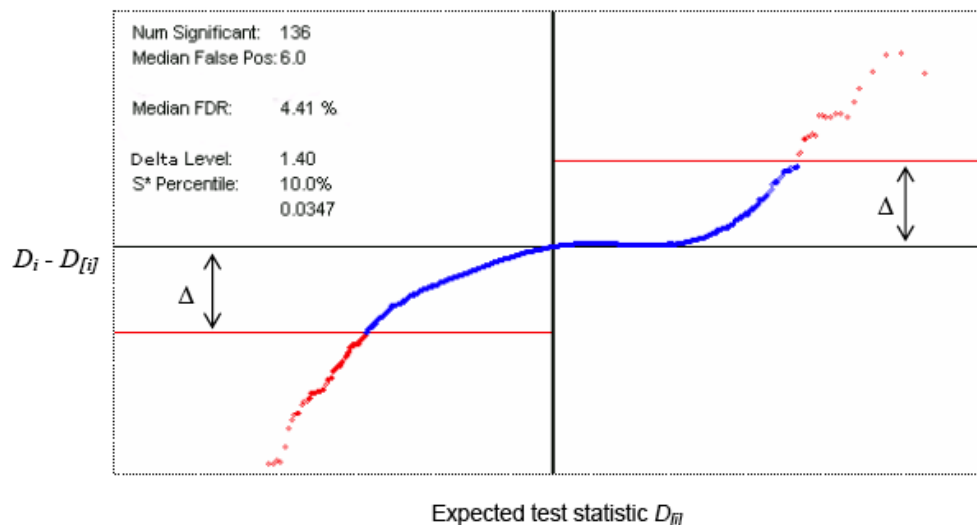
$$D_i = \frac{\bar{X}_i^{exp} - \bar{X}_i^{control}}{\hat{\sigma}_i^{pooled} \left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) + s^*_{10\%}}$$

where $\hat{\sigma}_i^{pooled}$ is the pooled estimated standard deviation and $s^*_{10\%}$ is the 10th %ile of values from $\hat{\sigma}_i^{pooled} \left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$

for transcripts $i=1, 2, \dots, n$ from the chip using your samples. The $s^*_{10\%}$ is a shrinkage factor that guards against random and undesirably large test statistics, primarily attributable to very small values being occasionally present in both the numerator and especially the denominator of the traditional t statistic. At first examination, the test statistic is nearly identical to the comparable t statistic. However, with permutation analysis, we repeatedly recompute this statistic by temporarily ignoring the actual groups each sample is associated with and assign each sample randomly to one of after all assignments are completed, will have the same number of samples as its original assignment. We continue with this rerandomization until we enumerate all potential reassignments or until the number of reassignments is sufficiently large (e.g., > 150 or more). After all permutations are completed, we build a reference distribution $D_{[i]}$ by finding the i th-ordered D_j values, $j=1, 2, \dots, n$, from each permutation instance then averaging each i th-ordered value across all permutations. The reference distribution indicates the expected value that we would get by chance when calculating the i th-ordered value from the true assignment of samples to groups. We then compare the observed with the expected ($D_i - D_{[i]} = \Delta_i$ (Delta) for each transcript i to determine if we detect unusual or nonrandom differences between groups.

For example, consider the following plot that we term the differential expression (DE) graph provided with the PADE output:

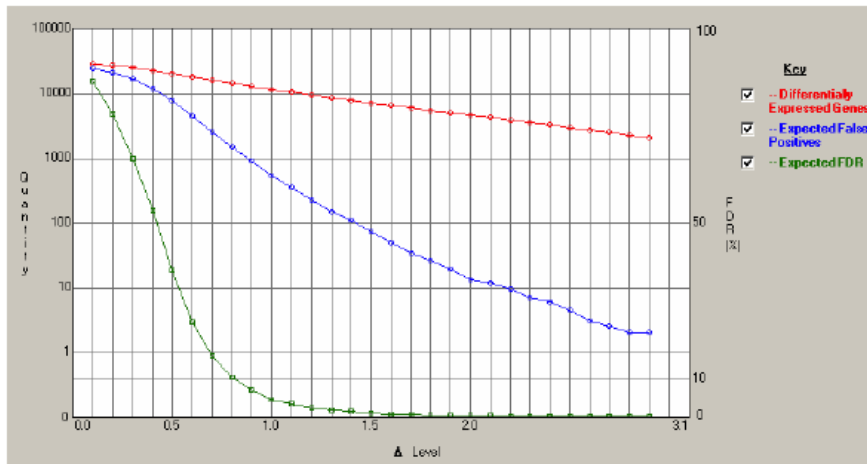
Differential Expression (DE) Graph



Each transcript is one point on the graph. With this simple graph we see evidence of specific increased (on the right) and decreased (on the left) differential expression. Transcripts that are undifferentiated between groups will be blue points between the Δ thresholds and the X-axis. The DE graph provides a summary of the number of transcripts called "significant" (136), the median-based estimated number of false positives from this set, also known as false discoveries (6 out of 136), the median-estimated FDR (4.41% = 6/136), and the shrinkage factor ($s^*_{10\%} = 0.0347$). The Delta (Δ) level indicates a threshold (1.4 in this experiment) for the difference between the observed test statistic and the expected test statistic (in absolute value) where we will call transcripts exceeding this threshold as differentially expressed, with an associated FDR. The appropriate Δ level for your experiment will depend on several factors, including the associated FDR. Typical values chosen for Δ are between 0.5 and 2 but values higher than this are sometimes selected to greatly reduce the FDR. To help determine the appropriate Δ level for your experiment, we provide another diagnostic called the FDR graph which is explained in the next paragraph.

An additional graph, termed the FDR graph and displayed below, is also provided with the PADE output and shows the effect of varying the Δ level relative to increasing or decreasing the differential transcript set size (shown in Red) and the resultant estimated FDR. [Note: The set sizes are on a log10 scale on the left while the estimated FDR uses a linear scale on the right.] For example, in the FDR graph below, which is from a different experiment than the DE graph shown earlier, the graph indicates that using a Δ level of 1.2 implies a total (differentiated transcript) set size of about 10,000 transcripts of which < 1000 are estimated to be false positives, yielding an FDR less than 10%. However, a Δ level of 3 implies a total (differentiated transcript) set size of > 2000 transcripts of which less than 10 are estimated to be false positives, yielding an FDR < 0.5%.

False Discovery Rate (FDR) Graph



Summary

This two-group comparison using permutation analysis with multiple samples provides very useful information related to the detection and estimation of differential expression between two groups, each group consisting of multiple arrays. The analysis provides statistically valid summary measures including false discovery rates along with transcript sets that are typically much more useful and indicative of differential expression between your groups than using techniques such as p-values alone, corrected p-values, or p-values with fold change estimates.

Comparisons can be requested at any time by communicating with an Expression Analysis representative. Expression Analysis also offers custom statistical analysis and consulting services to support your more exploratory or individualized analysis needs.