An Analysis to Identify Differentially Expressed and Alternately Spliced Genes

affymetrix

Abstract

In order to identify genes with differential gene expression or alternative splicing between the groups Nctx, Hipp, Striat, Thal, and Cbllm we study 95 hybridizations on the HumanExon10ST array using mixed model analysis of variance. 7911 genes with significant gene expression differences between the groups and 5818 genes with significant exon-group interaction (a symptom of alternative splicing) were found, including 4291 genes with both gene and possible splicing differences ($p<0.01$).

Contingency table analysis of the set of studied genes and a dataset of known pathways and gene classifications revealed that the set of alternatively spliced and expressed genes were found to be significantly over-represented in groups of the GOMolFn, GOProcess, GOCellLoc, and Pathway classes ($p<0.01$).

Project Information and Data Set

The following analysis was run on March 1, 2008 by affymetrix with XRAY (version 2.63) software, the Excel add-in from Biotique Systems Inc. (Burke, 2007). This document was auto-generated by XRAY. Project files are located in the D:\MattJ\ 5reg_263 directory.

The 95 input CEL files were analyzed with the Affymetrix HumanExon10ST array to identify genes that were significantly differentially expressed or displayed (Gardina, et al 2006; Huang, et al 2007; Clark et al 2007) significant differential alternative splicing between the groups of interest. The input files were assigned to 5 groups:

Group	CEL File
Nctx	DLPFC_18_Left.CEL
Nctx	DLPFC_18_Right.CEL
Nctx	DLPFC_19_Left.CEL
Nctx	DLPFC_19_Right.CEL
Nctx	DLPFC_21_Left.CEL
Nctx	DLPFC_21_Right.CEL
Nctx	DLPFC_23_Left.CEL
Nctx	DLPFC_23_Right.CEL
Nctx	MedPFC_18_Left.CEL
Nctx	MedPFC_18_Right.CEL
Nctx	MedPFC_19_Left.CEL
Nctx	MedPFC_19_Right.CEL
Nctx	MedPFC_21_Left.CEL
Nctx	MedPFC_21_Right.CEL
Nctx	MedPFC_23_Left.CEL

Nctx	MedPFC_23_Right.CEL
Nctx	MotorSens_18_Left.CEL
Nctx	MotorSens_18_Right.CEL
Nctx	MotorSens_19_Left.CEL
Nctx	MotorSens_19_Right.CEL
Nctx	MotorSens_21_Left.CEL
Nctx	MotorSens_21_Right.CEL
Nctx	MotorSens_23_Left.CEL
Nctx	MotorSens_23_Right.CEL
Nctx	Occipital_18_Left.CEL
Nctx	Occipital_19_Left.CEL
Nctx	Occipital_19_Right.CEL
Nctx	Occipital_21_Left.CEL
Nctx	Occipital_21_Right.CEL
Nctx	Occipital_23_Left.CEL
Nctx	Occipital_23_Right.CEL
Nctx	Occiptial_18_Right.CEL
Nctx	OrbFC_19_Left.CEL
Nctx	OrbFC_19_Right.CEL
Nctx	Parietal_18_Left.CEL
Nctx	Parietal_18_Right.CEL
Nctx	Parietal_19_Left.CEL
Nctx	Parietal_19_Right.CEL
Nctx	Parietal_21_Left.CEL
Nctx	Parietal_21_Right.CEL
Nctx	Parietal_23_Left.CEL
Nctx	Parietal_23_Right.CEL
Nctx	TempAssoc_18_Left.CEL
Nctx	TempAssoc_18_Right.CEL
Nctx	TempAssoc_19_Left.CEL
Nctx	TempAssoc_19_Right.CEL
Nctx	TempAssoc_21_Left.CEL
Nctx	TempAssoc_21_Right.CEL
Nctx	TempAssoc_23_Left.CEL
Nctx	TempAssoc_23_Right.CEL
Nctx	TempAud_18_Left.CEL
Nctx	TempAud_18_Right.CEL
Nctx	TempAud_19_Left.CEL
Nctx	TempAud_19_Right.CEL
Nctx	TempAud_21_Left.CEL
Nctx	TempAud_21_Right.CEL
Nctx	TempAud_23_Left.CEL
Nctx	TempAud_23_Right.CEL
Nctx	VLPFC_18_Left.CEL
Nctx	VLPFC_18_Right.CEL
Nctx	VLPFC_19_Left.CEL
Nctx	VLPFC_19_Right.CEL
Nctx	VLPFC_21_Left.CEL
Nctx	VLPFC_21_Right.CEL
Nctx	VLPFC_23_Left.CEL
Nctx	VLPFC_23_Right.CEL
Hipp	Hippocamp_18_Left.CEL
Hipp	Hippocamp_18_Right.CEL
Hipp	Hippocamp_19_Left.CEL
Hipp	Hippocamp_19_Right.CEL
Hipp	Hippocamp_21_Left.CEL

Hipp	Hippocamp_21_Right.CEL
Hipp	Hippocamp_23_Left.CEL
Hipp	Hippocamp_23_Right.CEL
Striat	Striatum_18_Left.CEL
Striat	Striatum_18_Right.CEL
Striat	Striatum_19_Left.CEL
Striat	Striatum_19_Right.CEL
Striat	Striatum_21_Left.CEL
Striat	Striatum_21_Right.CEL
Striat	Striatum_23_Left.CEL
Striat	Striatum_23_Right.CEL
Thal	Thalamus_18_Left.CEL
Thal	Thalamus_18_Right.CEL
Thal	Thalamus_19_Left.CEL
Thal	Thalamus_19_Right.CEL
Thal	Thalamus_21_Left.CEL
Thal	Thalamus_21_Right.CEL
Thal	Thalamus_23_Left.CEL
Thal	Thalamus_23_Right.CEL
Cbllm	Cbllm_18.CEL
Cbllm	Cbllm_19_Left.CEL
Cbllm	Cbllm_19_Right.CEL
Cbllm	Cbllm_21.CEL
Cbllm	Cbllm_23.CEL

and analyzed for differential gene expression and alternative splicing as detailed below.

Methods

Array Normalization

The input files were normalized with full quantile normalization (Irizarry et al 2003). For each input array, for each probe expression value, the array ith percentile probe value was replaced with the average of all array ith percentile points.

Low Level Data Handling

Next, the 6,553,590 probes were manipulated into the analysis values as follows. Probes with GC count less than 6 and greater than 17 were excluded from the analysis. Probe scores were then transformed by taking the Natural Logarithm of 0.1 plus the probe score.

Background Correction

Exon arrays do not use individual mis-match probes. Background is established from a pool of probes designed for that purpose. Background probes are stratified by CG content and are defined in the HumanExon10ST_antigenomic.bgp file. BGP files can also be downloaded from www.affymetrix.com. Each probe score was corrected for background by subtracting the median expression score of background probes with similar GC content.

Probe-set Expression Scores

The HumanExon10ST array contains 1,404,693 probe-sets (typically, but not always, groups of four probes).

Probe-set Expression Scores and Annotation Filtering

The expression score for a probe-set was defined to be the median of its probe expression scores and probe-sets with fewer than 3 probes (that pass all of the tests defined above) are excluded from further analysis. Exon Array Probes are designed off of genomic sequence and hence the reliability of probes and probe-sets correspond to the quality of their parent genomic annotations. Probe-set reliability is ranked from more to less reliable as Core, Extended, or Full. For example 'Core' probe-sets include probe-sets that correspond to high quality genomic features like RefSeq (www.ncbi.nlm.nih.gov) or Ensembl (www.ensembl.org) transcripts while 'full' and 'extended' probe-sets match less reliable annotations like EST hits and gene prediction algorithms. For this analysis, only 'Core' probe-sets were analyzed.

Probe-set Presence/Absence and the Removal of Non-expressed Probe-sets

Non-expressed probes can cause tests for alternative splicing to find false positives (because they cause 'non-parallel' expression patterns across the gene). A probe-set is judged to be expressed above background if for any group:

> Integral from T0 to Infinity of the standard normal distribution < Significance (0.001)
>
> Where:
>
> T0 = Sqr(GroupSize) (T - P) / Sqr(Pvar),
> GroupSize = Number of CEL files in the group,

T = Average of probe scores in probe-set,
P = Average of Background probes averages of GC content, and
Pvar = Sum of Background probe variances / (Number of probes in probe-set)^2,

Hence we test that the average of probe-sets in a group is greater than the average expression of background probes of similar gc content as the probe-set probes as the center of background for the probe-set and derive its dispersion from the background probe-set variance.

Filtering Invariant Probe-sets

Low-variance probe-sets are excluded from the analysis via a Chi-Square test. A probe-set is considered to be low-variance if its transformed variance is to the left of the 90 percent confidence interval of the Chi-Squared distribution with (N-1) degrees of freedom.

(N-1) * Probe-set Variance / (Gene Probe-set Variance) ~ Chi-Sq(N-1)

where N is the number of input CEL files, (N-1) is the degrees of freedom for the Chi-Squared distribution, and the 'probe-set variance for the gene' is the average of probe-set variances across the gene. Although, in practice, this method works well, it should be noted that the Chi-Square test of variance is usually applied to test a variance against a constant value and we are using it to test probe-set variance against a random variable (probe-set variance across gene) ; furthermore, the probe-set and probe-set across gene are not independent.

The following table summarizes the results of filtering.

Filtering Step	**Filter**	**Probes**	**Probe-Sets**	**Transcript Clusters**
0	Total on Chip	6,553,590	1,404,693	312,368
1	Core Probe-Sets	1,070,573	281,191	17,421(*)
2	Pass Filter 1 and Probes with GC Count between 6 and 17	979,086	246,666(**)	
3	Pass Filters 1, 2, and Probe-Sets Expressed Above Background	727,982	183,323(**)	
4	No Absolute Score Filter Used	727,982	183,323(**)	
5	Pass Filters 1, 2, 3, and 4, and Pass the Invariant Probe Filter	676,053	170,259(**)	13,223(*)

(*) Transcript clusters with between 4 and 200 passing probe-sets.
(**) Probe-sets contain at least 3 passing probes.

The 13,223 genes are passed on to the analysis to detect gene expression and alternative splicing differences between the groups. The number of genes tested may be much less than the number of transcript clusters on the chip because probe-set annotation level filtering (Core) and the removal of probe-sets not expressed in the groups Nctx, Hipp, Striat, Thal, or Cbllm can leave fewer transcript-clusters with more than 4 probe-sets.

Identification of Group Specific Gene Expression and Alternative Splicing

Mixed Model, Nested Analysis of Variance (Montgomery, 2006) was used to identify genes with group specific gene expression or alternative splicing. The nested model is appropriate because data is not sampled in a truly randomized fashion because expression points are harvested in batches defined by hybridizations (or individual CEL files). The mixed model is used since CEL files are random factors (i.e. we are not interested in the effect of individual Cel files since we are sampling from the many arrays that have been manufactured). Exons and groups are fixed effects. To justify the designation of states are fixed or random consider that if we were to redo the experiment we would use the same groups and exons but we would use different CEL files).

The data generated above are analyzed with Analysis of Variance (ANOVA) according to the linear model

$$Y[ijk] = M + d[i] + e[j] + c[k(i)] + ed[ij] + ec[jk(i)] + err[jk(i)]$$

where M is a global mean, d(i) is the effect attributable to group i, e(j) is the effect of exon j, and ec and ed are interaction effects. c, which is the hybridization (or kth chip) effect, is a random factor and all other factors are fixed. Note that the CEL file effect, c, is nested inside tissue state (d). Genes with significant D (tissue or group) effect are said to show significant differential gene expression difference between the study groups. Genes with significant Exon-Tissue interaction (ED effect) are said to show signs of tissue specific alternative splicing (p-value < 0.01). Interactions appear as "non-parallel" lines when average group expression is plotted over the exons.

Multiple Tests Correction

Under proper randomization conditions, for each gene we test, the probability of a 'false-positive' (or 'Type I Error') is 0.01. Because we are testing a large number of "independent" genes, this significance value is misleading since the probability of finding a false-positive will grow as we test more genes (Glantz, 1996). To correct for this we use the Benjamini and Hochberg False Discovery Rate (FDR) method outlined by Benjamini and Hochberg (1995) and originally proposed by Simes (1986) that controls the family wide error rate in a weak sense (The "False Discovery Rate" of expected proportion of false positives is controlled - by contrast methods like the Bonferroni correction control false discoveries in the "strong sense" and bound the probability of occurrence of ANY false positives these methods tend to be too conservative and have low power for these types of studies). Benjamini and Hochberg outline the Simes procedure as follows. The gene-level p-values are sorted in ascending order and then corrected:

$$p\_corrected = p\_value * 1$$

for the largest p-value

$$p\_corrected = p\_value * (N/N-1)$$

for the second largest p-value, and

$$p\_corrected = p\_value * (N/N-2)$$

for the third largest p-value, etc... Where N is the total number of genes tested. The False Discovery Rate, say R, for the project can then be established by removing all genes where p_corrected > R. Alternatively, all the pre-correction significant results can be retained and the FDR can be set as the maximum p_corrected. The individual tests are assumed to be

independent and, while this method is regarded as a standard for expression analysis, it does not account for correlations between genes.

Determining Tissue Presence/Absence Using Group Expression Above Background

To establish the presence or absence of expression for a particular gene in a group (or tissue) we derive a p-value to test the null hypothesis that the average of CEL files belonging to the group is not above background. In more precise terms the p-value is the likelihood of observing the gene-wide expression level under the null hypothesis that the tissue is not expressed above background. Rejection of the null hypothesis occurs when the p-value is less than the significance level 0.01 in which case we infer that the gene is most likely expressed in the given tissue. Given the probes in a particular gene this significance is assessed by evaluating [1 - CumulativeStandardNormal(N0)] at N0 = Sum(probe - GCBackground) / Var where the sum is over all probe scores belonging to CEL files in the group, GCBackground is the median score at a particular probe GC count, and Var is the average background variance.

Group Expression Level Filters to Reduce Alternative Splicing False Positives

Large differences between tissue expression levels in a gene can cause false positives for alternative splicing by introducing non-linear behavior that departs from the above model of expression. Specifically, the situation when exon expression of one group approaches background (or saturation) while other groups remain in the dynamic range causes "non-parallel" expression between the groups because the expression values in the dynamic range are free to vary while samples near saturation or background are "dampened". Such non-parallel behavior may register (falsely) as group specific alternative exon usage. We attempt to filter out these cases by using p-values for group expression in the gene.

Only One Group Expressed

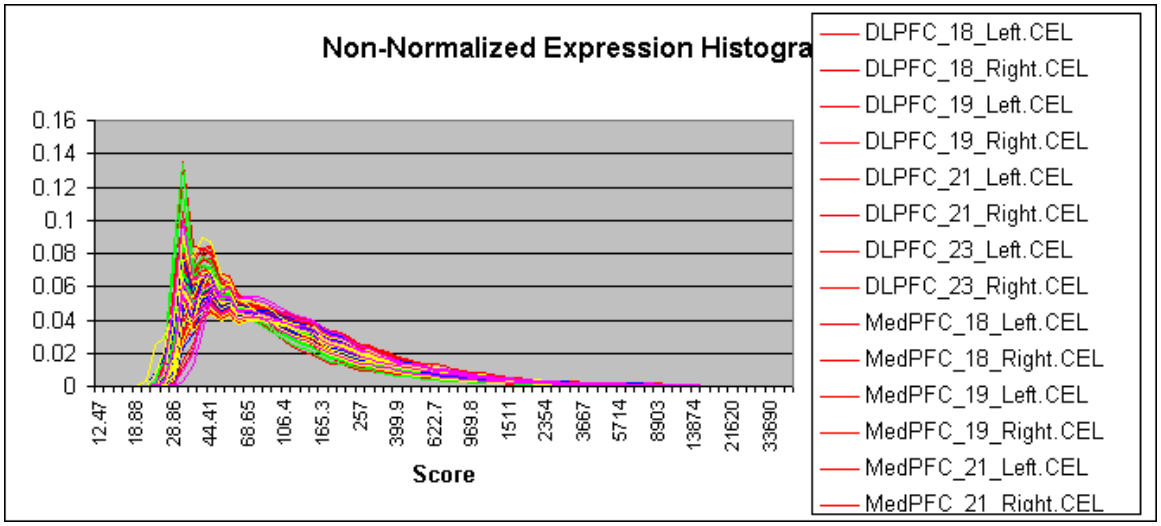
For each gene, generate p-values for the presence of 5 groups as desribed above. Remove CEL files belonging to groups not significantly expressed above background. Do not test genes for alternative splicing if only one group is expressed.

Input Data Quality Reports

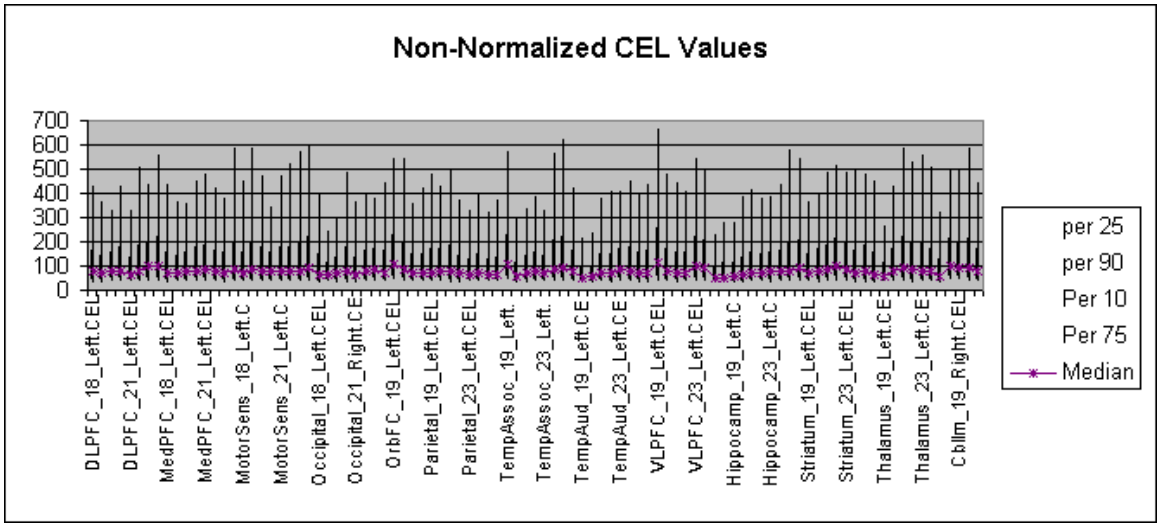
This section presents some quality plots for each array. These include methods run before and after normalization.

Before Normalization

These views are generated with non-normalized, non-background corrected, untransformed probe-level data.



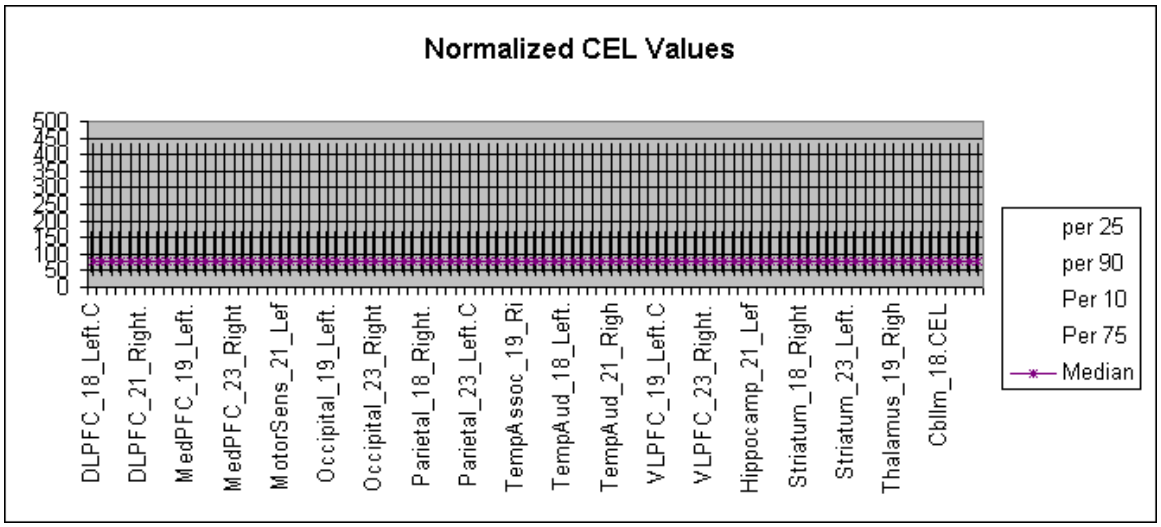
This graph shows the distribution of scores for each array. Each array has a line, the x axis represents score, and the y-axis represents the number of probes with score in the range divided by the total number of probes.



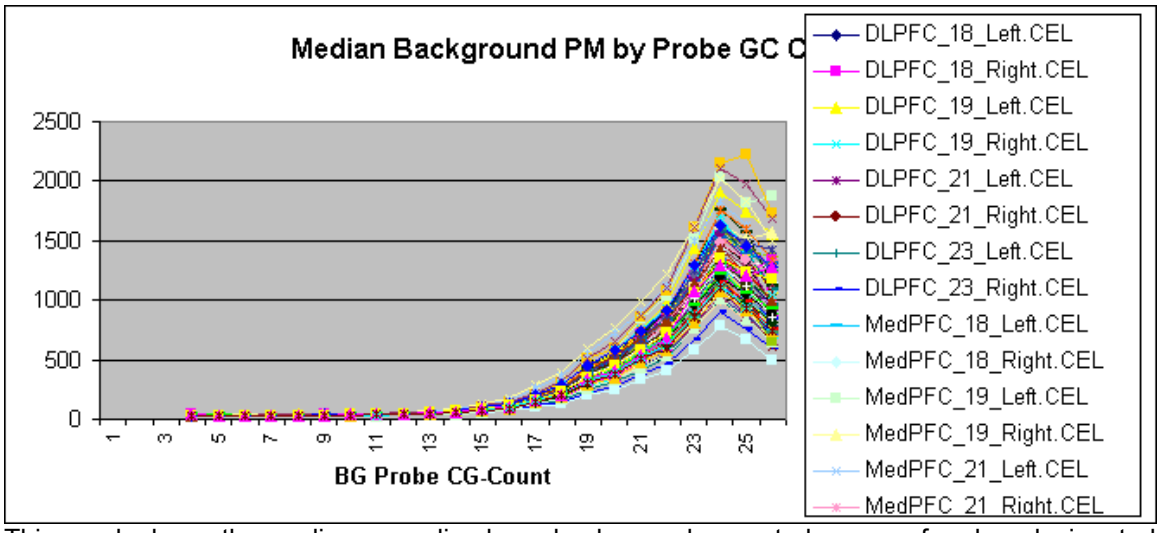
This graph shows the distribution of scores for each array. Each array is represented by a box plot where the middle (joined by lines to aid comparison) is the probe score median, the box top and bottom are the 25th and 75th percentiles of probe score, and the top and bottom lines are the 10th and 90th percentile of probe score.

After Normalization

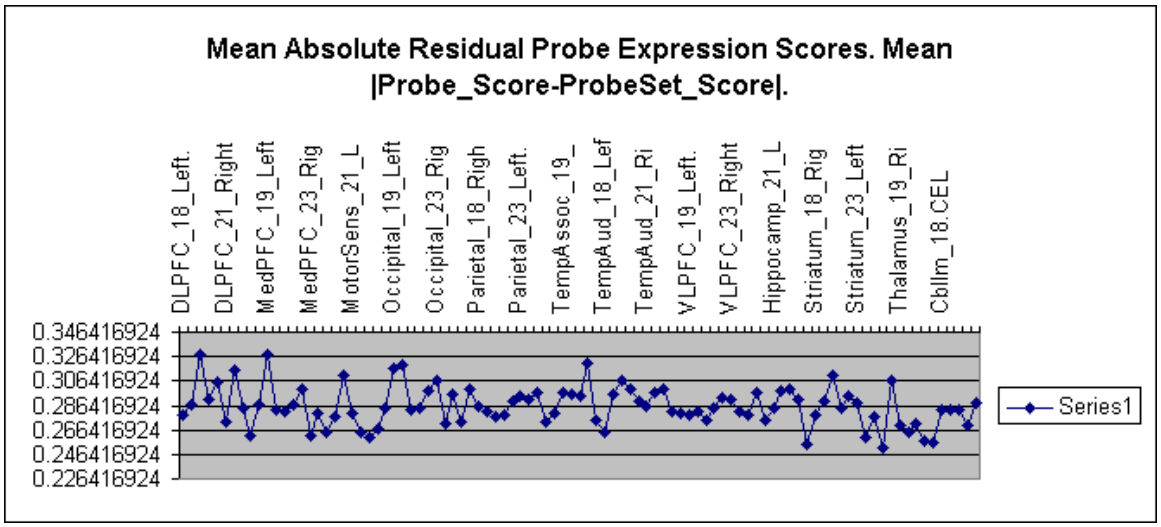
These views are generated with normalized, non-background corrected, untransformed probe-level data. For details of the normalizations see the 'Normalization' section in 'Methods'.



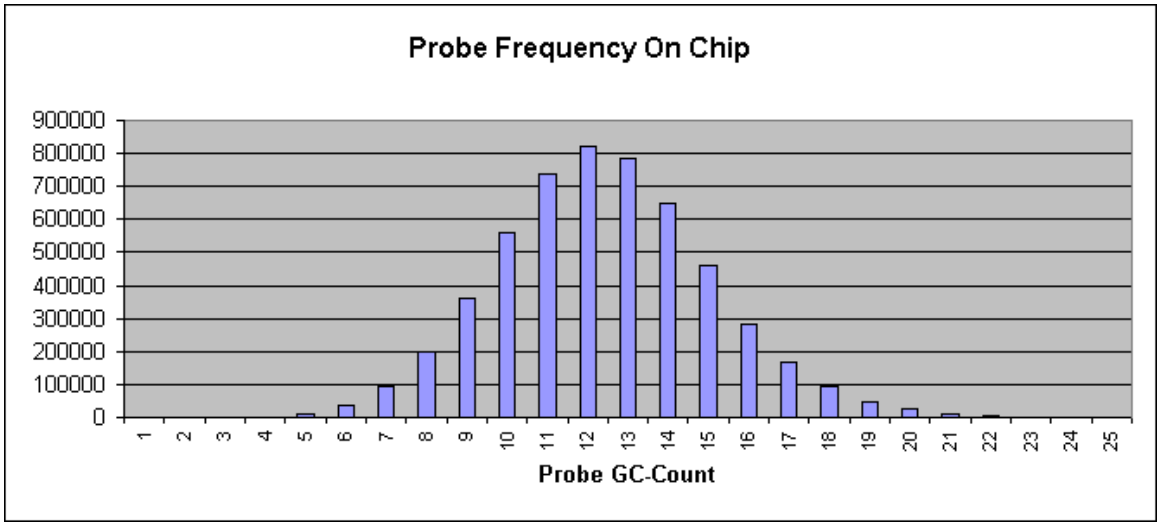
This graph shows the distribution of scores for each array. Each array is represented by a box plot where the middle (joined by lines to aid comparison) is the probe score median, the box top and bottom are the 25th and 75th percentiles of probe score, and the top and bottom lines are the 10th and 90th percentile of probe score. All box plots are identical because full quantile normalization forced all input arrays to have identical "shapes".



This graph shows the median normalized non-background corrected scores of probes designated as background. Each chip has a line, the x-axis represents GC count and the y axis is probe-score.



This graph shows the mean absolute difference between transformed, background-corrected (if specified), and normalized probe scores and summarized probe-set scores. The x-axis represents array and each array has a point on the graph. This is a quality score given by the ExACT package from Affymetrix.



This graph shows a histogram where the x-axis is probe GC count and the y-axis is the frequency on the input arrays. This graph is, of course, the same before and after normalization.

Results

Tissue Distribution of Expression

The HumanExon10ST array contains a total of 312,368 transcript clusters. After the filters described above were applied 13,223 contained between 4 and 200 probe-sets. These were tested for differential gene expression and alternative splicing using the statistical methods described above. For the tested transcript clusters, the following table contains a summary of the number of genes (transcript clusters) expressed in each group.

Group	**Number of transcript clusters with significant expression in group**
Nctx	8,876 67.1% of genes tested
Hipp	9,586 72.5% of genes tested
Striat	9,764 73.8% of genes tested
Thal	9,866 74.6% of genes tested
Cbllm	10,020 75.8% of genes tested

Using the same test, the following table summarizes frequencies of pair-wise co-expression between the study groups.

-	Nctx	Hipp	Striat	Thal	Cbllm
Nctx	8,876**(0,659)**	7,528(0,062)	7,628(0,075)	7,643(0,070)	7,639(0,188)
Hipp	-	9,586**(0,114)**	9,054(0,032)	9,058(0,038)	8,891(0,065)
Striat	-	-	9,764**(0,083)**	9,227(0,067)	9,036(0,081)
Thal	-	-	-	9,866**(0,135)**	9,092(0,109)
Cbllm	-	-	-	-	10,020**(0,316)**

If present, the numbers in parentheses are exclusive and the plain number is inclusive so, for example, there are 8,876 genes in which the Nctx group is expressed significantly above background while 0,659 genes has this group and no other groups expressed above background. All patterns of co-expression is summarized in the following table. Frequencies are exclusive so, for example there are 6924 genes where the group Nctx+Hipp+Striat+Thal+Cbllm and no other tissues are expressed.

Nctx+Hipp+Striat+Thal+Cbllm 6924
Nctx 659
Nctx+Striat+Thal 58
Striat+Thal 67
Hipp+Striat+Thal+Cbllm 1528
Hipp+Striat+Thal 136
Striat+Cbllm 81
Nctx+Hipp+Striat+Thal 234
Nctx+Hipp+Thal36
Nctx+Thal 70
Cbllm 316
Nctx+Hipp+Striat+Cbllm85
Nctx+Striat+Thal+Cbllm 143
Thal 135
Nctx+Cbllm 188
Striat+Thal+Cbllm 137

Thal+Cbllm	109
Nctx+Hipp+Cbllm	55
Nctx+Hipp+Thal+Cbllm	89
Nctx+Striat+Cbllm	66
Nctx+Hipp	62
Hipp	**114**
Hipp+Striat+Cbllm	72
Nctx+Hipp+Striat	43
Striat	**83**
Nctx+Thal+Cbllm	89
Hipp+Thal	38
Hipp+Thal+Cbllm	73
Hipp+Striat	32
Nctx+Striat	75
Hipp+Cbllm	65

Differential Gene Expression and Alternative Splicing

The statistical analysis explained in the 'Methods' section resulted in 7911 genes with significant gene expression differences between the groups and 5818 genes with significant exon-group interaction (a symptom of alternative splicing) including 4291 genes with both gene differences and interaction.

Gene Symbol	TCluster ID	Description	Fold Change	Differential Expression p-value
SLA	3154263	Src-like-adaptor	18.71	7.90E-70
ATBF1	3698256	AT-binding transcription factor 1	7.01	2.34E-58
MAB21L1	3509411	mab-21-like 1 (C. elegans)	23.91	1.94E-58
SATB2	2594089	SATB homeobox 2	17.11	7.41E-58
TBR1	2512752	T-box brain 1	22.71	5.94E-57
SLITRK6	3519840	SLIT and NTRK-like family member 6	7.11	5.21E-57
NEUROD6	3044518	neurogenic differentiation 6	47.01	9.96E-56
KIAA1772	3780981	KIAA1772	6.21	3.24E-55
TCF7L2	3264621	transcription factor 7-like 2 (T-cell sp	32.61	2.08E-54
FAT2	2882026	FAT tumor suppressor homolog 2 (Drosophi	3.81	6.57E-53

Table of top 10 fold changes in genes with significant differential gene expression. Fold change is in terms of the normalized untransformed data.

Gene Symbol	TCluster ID	Description	Exon-Tissue Interaction p-value
ATBF1	3698256	AT-binding transcription factor 1	0.00E+00
TBR1	2512752	T-box brain 1	0.00E+00
SLITRK6	3519840	SLIT and NTRK-like family member 6	0.00E+00
ZNF521	3801943	zinc finger protein 521	0.00E+00
EBF1	2883878	early B-cell factor 1	0.00E+00
EFCBP2	3671552	EF-hand calcium binding protein 2	0.00E+00
DACH1	3517251	dachshund homolog 1 (Drosophila)	0.00E+00
ZIC4	2699844	Zic family member 4	0.00E+00
ALK	2546409	anaplastic lymphoma kinase (Ki-1)	0.00E+00

KIAA1622	3549605	KIAA1622	0.00E+00

Table of top 10 with significant differential alternative splicing.

False Discovery Rate

The sequential step-down procedure described above was used to calculate that the false discovery rate for this project is less than 1.32E+04 for differential alternative splicing and gene expression tests.

Comparison of Differentially Expressed genes and Exons to Known Gene Classifications

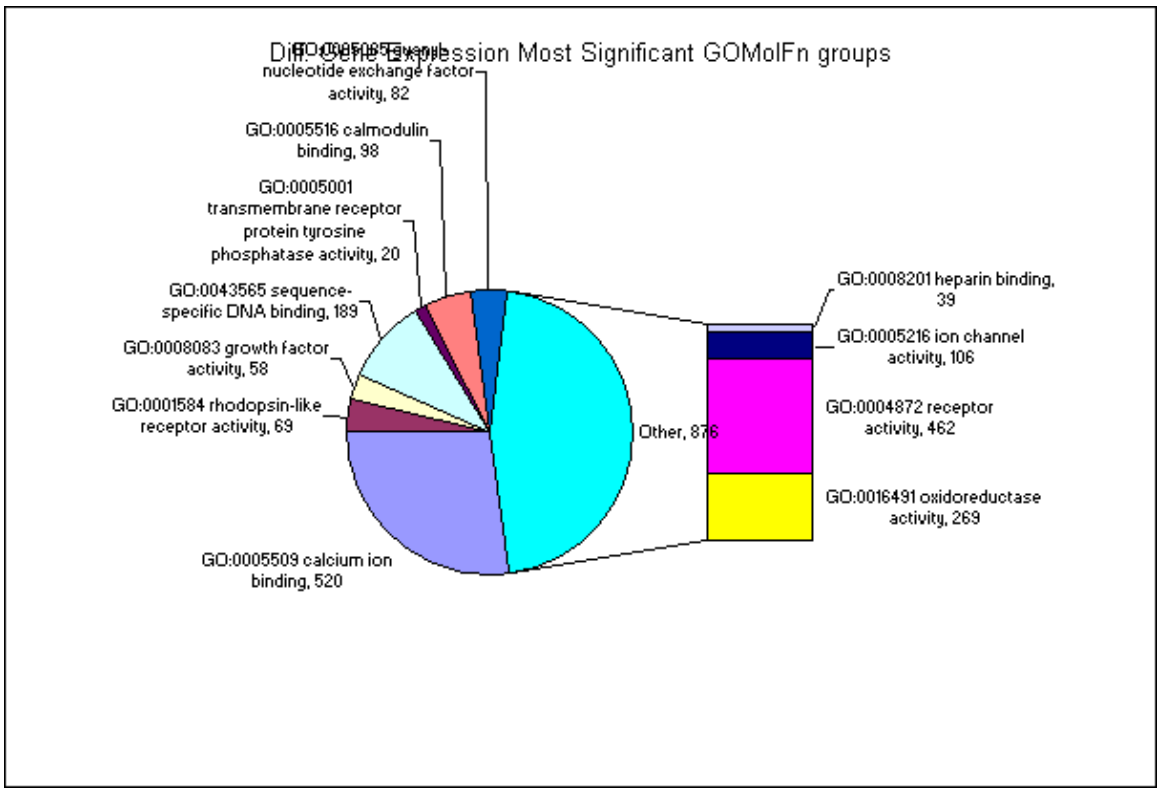
The 13,223 genes tested for differential alternative splicing and gene expression were compared to known gene classifications (contained in the file HumanExon10ST.info) to identify significant over-representation in groups of the GOMolFn, GOProcess, GOCellLoc, or Pathway classes .Contingency table analysis (Agresti, 1990) was used to identify groups in which the genes with significant splicing or expression differences were over-represented. The number of significant genes in a group follows a hyper-geometric distribution under random conditions (Ross, 1989) and the probability of seeing the given number or more significant genes in a group can be approximated as:

1 - Cumulative Normal(x,mean=0, variance=1)
where x = (a - (n * p)) / Sqr(n * p * (1-p)),
a = Number of Significant genes in group,
n = Number of genes in group, and
p = Total Number of Significant Genes / Total Number of Genes

This is similar to a one sided version of the Fisher Exact Test. The exact statistic could be calculated by summing over the discreet hypergeometric pdf values. The above calculation was performed for each group and the significant groups (where the above function returns less than 0.01 for gene expression or alternative splicing) are given for each annotation class below.

Significant Representation in group of the GOMolFn classification

The GOMolFn gene classification had 38 groups that were significantly over-represented in the set of differentially spliced or expressed genes (differential splicing and gene expression determined as described above). The top 30 groups are shown in the table below. There is one group per row and the three columns contain the number of tested genes found to have significant differential gene expression (with p-value of over-representation), the number of genes found to have significant differential splicing (with p-value of over-representation), and the group name respectively.

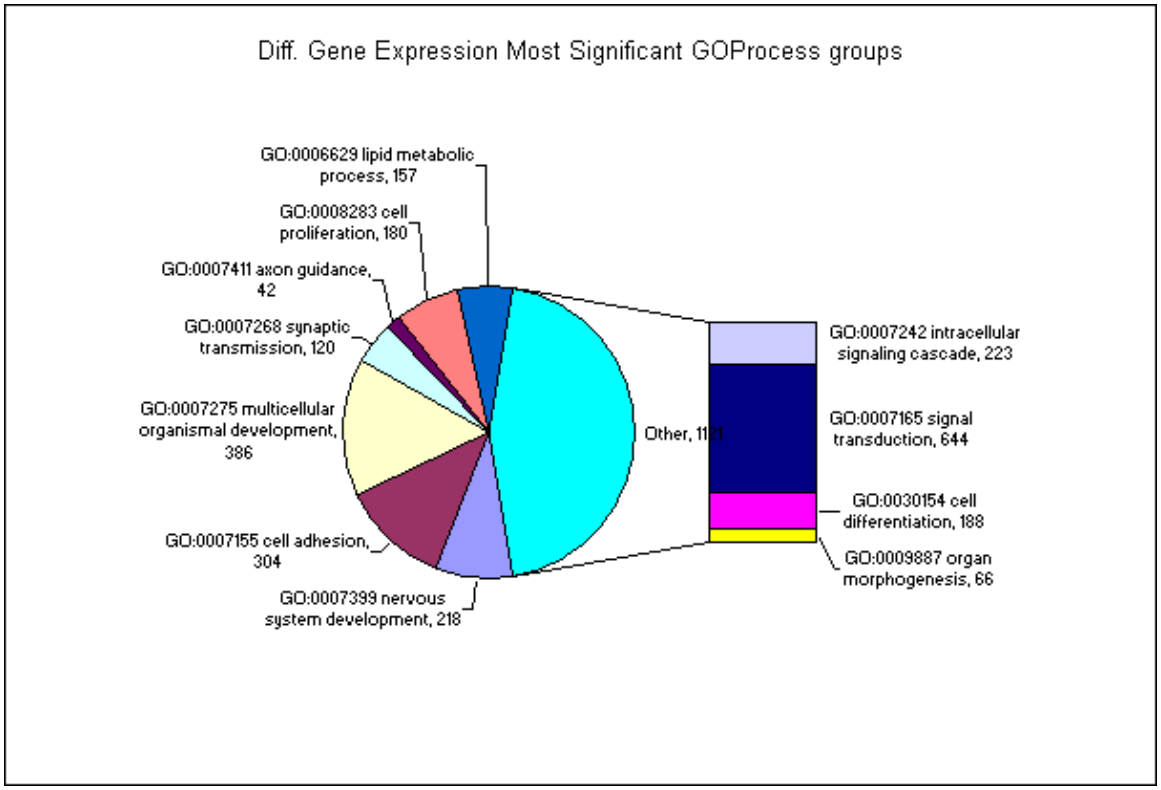


Number GE	Number AS	Group Name
520(2.29E-11)	376(8.50E-06)	GO:0005509 calcium ion binding
2456(1.94E-01)	1925(5.85E-06)	GO:0005515 protein binding
46(5.61E-01)	53(1.02E-05)	GO:0003682 chromatin binding
69(2.97E-05)	40(2.85E-01)	GO:0001584 rhodopsin-like receptor activ
13(2.22E-01)	17(3.26E-05)	GO:0018024 histone-lysine N-methyltransf
58(8.58E-05)	39(3.18E-02)	GO:0008083 growth factor activity
189(1.78E-04)	138(6.73E-03)	GO:0043565 sequence-specific DNA binding
19(8.41E-02)	20(3.60E-04)	GO:0008092 cytoskeletal protein binding
94(1.10E-01)	84(3.63E-04)	GO:0005096 GTPase activator activity
25(3.10E-02)	24(4.43E-04)	GO:0004714 transmembrane receptor protei
20(4.66E-04)	15(5.67E-03)	GO:0005001 transmembrane receptor protei
80(1.66E-02)	67(1.03E-03)	GO:0042803 protein homodimerization acti
98(1.24E-03)	70(3.30E-02)	GO:0005516 calmodulin binding
82(1.27E-03)	62(5.93E-03)	GO:0005085 guanyl-nucleotide exchange fa
39(2.39E-03)	28(3.19E-02)	GO:0008201 heparin binding
106(2.42E-03)	67(4.06E-01)	GO:0005216 ion channel activity
462(2.54E-03)	334(5.49E-02)	GO:0004872 receptor activity
714(2.06E-02)	547(2.62E-03)	GO:0016740 transferase activity
269(2.80E-03)	193(6.33E-02)	GO:0016491 oxidoreductase activity
0(1.00E+00)	6(2.86E-03)	GO:0004556 alpha-amylase activity
45(2.90E-03)	32(4.32E-02)	GO:0005201 extracellular matrix structur
13(5.85E-03)	11(4.58E-03)	GO:0004114 3'-5'-cyclic-nucleotide phosp
159(5.64E-03)	120(1.24E-02)	GO:0003779 actin binding
5(3.35E-02)	5(5.82E-03)	GO:0005242 inward rectifier potassium ch
5(3.35E-02)	5(5.82E-03)	GO:0005007 fibroblast growth factor

		rece
13(5.85E-03)	8(1.61E-01)	GO:0042826 histone deacetylase binding
16(5.95E-03)	13(7.93E-03)	GO:0004435 phosphoinositide phospholipas
16(5.95E-03)	8(4.85E-01)	GO:0015171 amino acid transporter activi
6(9.66E-01)	12(6.24E-03)	GO:0000049 tRNA binding
6(1.91E-01)	7(6.59E-03)	GO:0043531 ADP binding

Significant Representation in group of the GOProcess classification

The GOProcess gene classification had 55 groups that were significantly over-represented in the set of differentially spliced or expressed genes (differential splicing and gene expression determined as described above). The top 30 groups are shown in the table below. There is one group per row and the three columns contain the number of tested genes found to have significant differential gene expression (with p-value of over-representation), the number of genes found to have significant differential splicing (with p-value of over-representation), and the group name respectively.

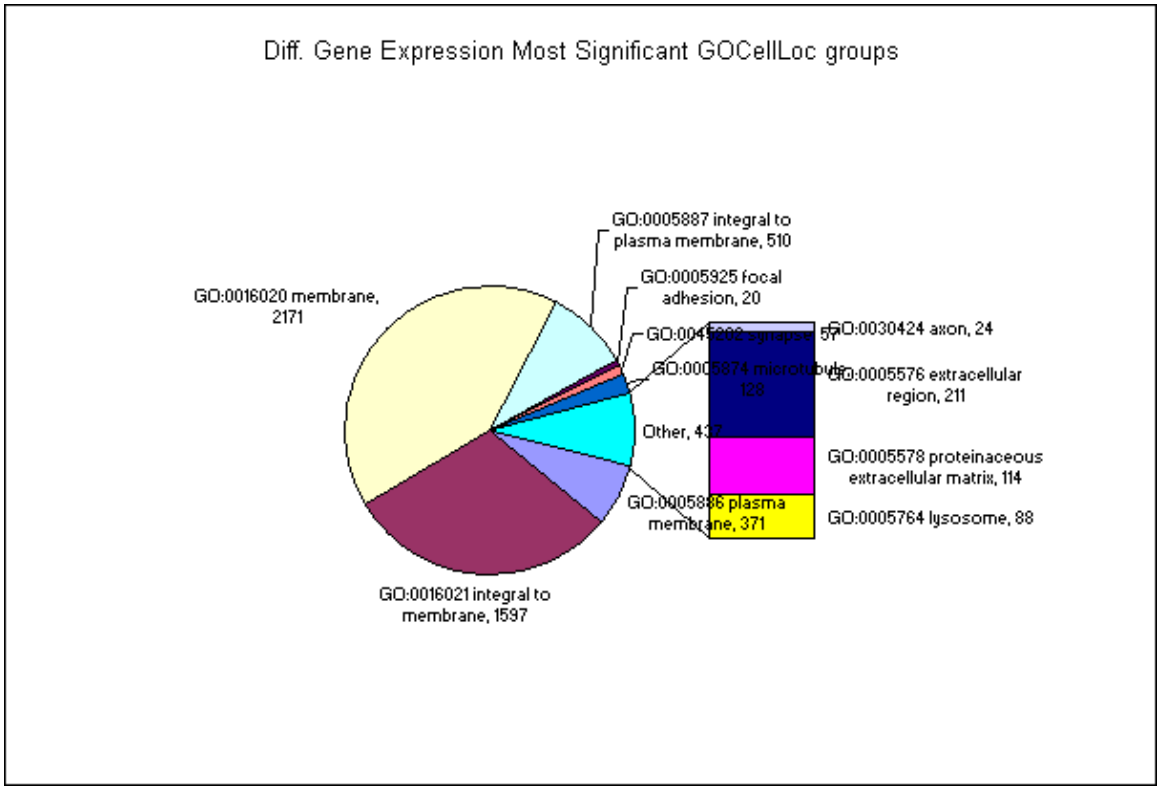


Number GE	Number AS	Group Name
218(1.29E-10)	168(1.70E-08)	GO:0007399 nervous system development
304(2.58E-08)	218(3.30E-04)	GO:0007155 cell adhesion
386(2.72E-07)	280(4.74E-04)	GO:0007275 multicellular organismal deve
120(4.25E-07)	85(1.17E-03)	GO:0007268 synaptic transmission
42(6.66E-06)	34(2.18E-05)	GO:0007411 axon guidance
180(3.02E-05)	124(3.26E-02)	GO:0008283 cell proliferation
157(8.40E-05)	106(7.51E-02)	GO:0006629 lipid metabolic process
41(1.30E-02)	38(8.88E-05)	GO:0007420 brain development
223(1.20E-04)	154(6.19E-02)	GO:0007242 intracellular signaling casca
644(1.62E-04)	478(1.91E-03)	GO:0007165 signal transduction
188(2.09E-04)	143(8.26E-04)	GO:0030154 cell differentiation
68(8.71E-01)	74(2.18E-04)	GO:0016568 chromatin modification
13(4.02E-02)	14(2.28E-04)	GO:0001764 neuron migration
66(2.29E-04)	49(4.06E-03)	GO:0009887 organ morphogenesis
25(2.58E-04)	14(2.06E-01)	GO:0006006 glucose metabolic process
26(5.26E-04)	13(4.64E-01)	GO:0009968 negative regulation of signal

26(5.26E-04)	20(3.38E-03)	GO:0016049 cell growth
255(6.23E-04)	175(1.49E-01)	GO:0006811 ion transport
15(8.44E-03)	14(7.22E-04)	GO:0051258 protein polymerization
354(7.92E-04)	266(2.65E-03)	GO:0006468 protein amino acid phosphoryl
12(2.41E-02)	12(8.32E-04)	GO:0009953 dorsal/ventral pattern format
59(5.54E-03)	49(9.41E-04)	GO:0007169 transmembrane receptor protei
40(1.64E-02)	35(1.03E-03)	GO:0051056 regulation of small GTPase me
62(1.08E-03)	40(1.65E-01)	GO:0007018 microtubule-based movement
70(4.45E-03)	57(1.20E-03)	GO:0007156 homophilic cell adhesion
20(1.47E-03)	13(7.69E-02)	GO:0007626 locomotory behavior
5(7.37E-01)	9(1.69E-03)	GO:0006378 mRNA polyadenylation
84(1.92E-03)	57(1.15E-01)	GO:0006813 potassium ion transport
59(8.36E-02)	52(2.17E-03)	GO:0030036 actin cytoskeleton organizati
20(8.62E-01)	26(2.17E-03)	GO:0043087 regulation of GTPase activity

Significant Representation in group of the GOCellLoc classification

The GOCellLoc gene classification had 25 groups that were significantly over-represented in the set of differentially spliced or expressed genes (differential splicing and gene expression determined as described above). The top 25 groups are shown in the table below. There is one group per row and the three columns contain the number of tested genes found to have significant differential gene expression (with p-value of over-representation), the number of genes found to have significant differential splicing (with p-value of over-representation), and the group name respectively.

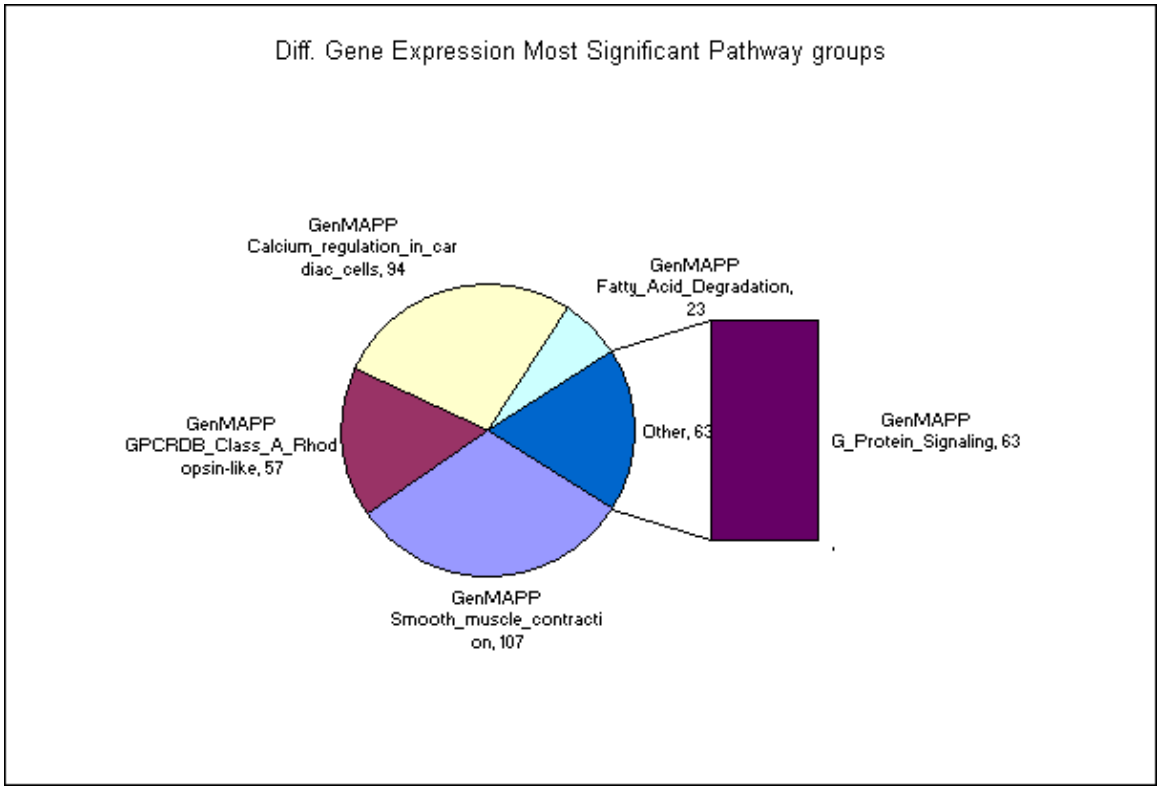


Number GE	Number AS	Group Name
371(4.63E-06)	275(3.27E-04)	GO:0005886 plasma membrane
1597(9.15E-06)	1093(5.69E-01)	GO:0016021 integral to membrane
2171(1.18E-05)	1518(3.58E-01)	GO:0016020 membrane
510(2.13E-05)	353(8.70E-02)	GO:0005887 integral to plasma membrane
848(8.97E-01)	714(5.92E-05)	GO:0005737 cytoplasm
20(1.24E-04)	11(1.61E-01)	GO:0005925 focal adhesion
57(2.19E-04)	38(5.30E-02)	GO:0045202 synapse
128(2.41E-04)	97(1.49E-03)	GO:0005874 microtubule
24(3.65E-04)	16(3.58E-02)	GO:0030424 axon
211(4.31E-04)	124(8.80E-01)	GO:0005576 extracellular region
114(5.43E-04)	79(5.53E-02)	GO:0005578 proteinaceous extracellular m
11(7.63E-02)	12(8.32E-04)	GO:0005884 actin filament
13(1.42E-01)	14(1.94E-03)	GO:0000228 nuclear chromosome
74(1.92E-01)	66(2.57E-03)	GO:0015629 actin cytoskeleton
88(2.79E-03)	58(2.15E-01)	GO:0005764 lysosome
18(2.96E-03)	14(9.58E-03)	GO:0005891 voltage-gated calcium channel
49(3.16E-03)	33(1.11E-01)	GO:0008076 voltage-gated potassium chann

17(4.20E-03)	12(4.62E-02)	GO:0043025 cell soma
10(4.78E-03)	7(4.88E-02)	GO:0042734 presynaptic membrane
36(5.38E-03)	27(2.23E-02)	GO:0005923 tight junction
46(4.51E-01)	44(7.33E-03)	GO:0000785 chromatin
226(7.57E-03)	165(5.13E-02)	GO:0005856 cytoskeleton
53(8.52E-03)	38(6.67E-02)	GO:0045211 postsynaptic membrane
1929(1.00E+00)	1631(8.66E-03)	GO:0005634 nucleus
195(7.27E-01)	169(8.71E-03)	GO:0005829 cytosol

Significant Representation in group of the Pathway classification

The Pathway gene classification had 5 groups that were significantly over-represented in the set of differentially spliced or expressed genes (differential splicing and gene expression determined as described above). The top 5 groups are shown in the table below. There is one group per row and the three columns contain the number of tested genes found to have significant differential gene expression (with p-value of over-representation), the number of genes found to have significant differential splicing (with p-value of over-representation), and the group name respectively.



Number GE	Number AS	Group Name
107(5.02E-05)	73(3.15E-02)	GenMAPP Smooth_muscle_contraction
57(1.14E-04)	31(4.81E-01)	GenMAPP GPCRDB_Class_A_Rhodopsin-like
94(1.85E-03)	67(4.17E-02)	GenMAPP Calcium_regulation_in_cardiac_ce
23(8.01E-03)	13(3.98E-01)	GenMAPP Fatty_Acid_Degradation
63(8.32E-03)	44(1.08E-01)	GenMAPP G_Protein_Signaling

References

Agresti, A., (1990). 'Categorical Data Analysis' Wiley Inter-Science.

Bolstad, B. M., Irizarry R. A., Astrand, M, and Speed, T. P. (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19,2,pp 185-193

Burke, J. Biotique Systems (2006,2007). XRAY Software from Biotique Systems. www.biotiquesystems.com, www.orderxray.com.

Clark, T.A., Schweitzer, A.C., Chen, T.X., Staples, M.K., Lu, G., Wang, H., Williams, A., Blume, J.E. (2007). 'Discovery of Tissue-specific Exons Using Comprehensive Human Exon Microarrays.' Genome Biology, 2007. 8:R64.

Benjamini, Y. and Hochberg, Y. (1995). 'Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.' Journal of the Royal Statistical Society B, Vol 57, Num 1, 289-300.

Crawley, M.J (2003). 'Statistical Computing. An Introduction to Data Analysis Using S-Plus', Wiley.

Durbin, R., Eddy, S., Krough, A., Mitchison, G. (1998). 'Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids', Cambridge University Press.

Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, W., Davies, C., Williams, A., Turpaz, Y. (2006). Alternative Splicing and Differential Gene Expression in Colon Cancer Detected by a Whole Genome Exon Array. BMC Genomics. 7:325.

Glantz, S. (1996). 'Primer of Bio-Statistics.' 4th Edition., McGraw Hill.

Holm, S. (1979). 'A Simple Sequentially Rejective Bonferroni Test Procedure.' Scandinavian Journal of Statistics, 6, 65-70.

Huang, S.R., Duan, S., Bleibel, W.K., Kistner, E.O., Zhang, W., Clark, T.A., Chen, T.X., Schweitzer, A.C., Blume, J.E., Cox, N.J., Dolan, E.M. (2007). 'A Genome-wide Approach to Identify Genetic Variants that Contribute to Etoposide-induced Cytotoxicity.' PNAS Vol. 104, No. 23.

Irizarry, R.A., Hobbs B., Collin F., Beazer-Barclay, Y.D., Antonellis, K.UJ., Scherf, U., and Speed T.P. (2003). 'Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.' Biostatistics, 4, 249-264.

Irizarry, R.A., Bolstad B.M., Collin, F., Cope, L.M., Hobbs B. and Speed, T.P. (2003). 'Summaries of Affymetrix GeneChip Probe Level Data.' Nucleic Acids Research 31(4)

Montgomery D.C., (2006). 'Design and Analysis of Experiments.' 5th Edition. Wiley and Sons.

Ross S.M., (1989). 'Introduction to Probability Models.' 4th Edition. Academic Press.

Simes, R.J., (1986). 'An improved Bonferroni Procedure for Multiple Tests of Significance.' Biometrika, 77, 663-665

Auxiliary Files

The following files can be found in the directory D:\MattJ\5reg_263.

5reg_263.xls

This is an excel sheet in which the XRAY project was run. the full list of results can be found here.

5reg_263_methods.doc

The title of this document. Describes methods used and results found by XRAY analysis.

5reg_263_gene.txt

For each gene (transcript cluster) that had the minimum number of probe-sets at the proper annotation level (core, extended, or full), lists the transcript cluster ID, gene symbol (symbol derived from NetAffx .csv annotation files), differential gene expression between groups pvalue, differential alternative splicing between groups pvalue, and for each group, the group median probe expression and group presence pvalue. All scores are normalized and background corrected.

5reg_263_probe.txt

For each probe belonging to probe-sets at the proper annotation level (core, extended, or full), lists the transcript cluster ID, probe-set ID, probe-set level, GC count, and for each input file the probe score. All scores are normalized and background corrected.

5reg_263_results.ipa

This is a file suitable for input into the Ingenuity (C) Pathway Analysis tools (TM). For each gene (transcript cluster) that had the minimum number of probe-sets at the proper annotation level (core, extended, or full), lists the transcript cluster ID, GenBank idenfifier, gene symbol (symbol derived from NetAffx .csv annotation files), differential gene expression between groups pvalue, differential alternative splicing between groups pvalue, and for each group, the group median probe expression and group presence pvalue. All scores are normalized and background corrected.