mvGST: Multivariate and Directional Gene Set Testing

John R. Stevens[1], Dennis S. Mecham[1], and Garrett Saunders[1,2]

October 17, 2016

1. Dept. of Mathematics and Statistics, Utah State University (`http://www.stat.usu.edu/jrstevens`)
2. Dept. of Mathematics, Brigham Young University – Idaho

Abstract

In a gene expression experiment (using oligo array, RNA-Seq, or other platform), researchers typically seek to characterize differentially expressed genes based on common gene function or pathway involvement. The field of gene set testing provides numerous characterization methods, some of which have proven to be more valid and powerful than others. Previous gene set testing methods have focused on experimental designs where there is a single null hypothesis (usually involving association with a continuous or categorical phenotype) for each gene. However, increasingly common experimental designs lead to multiple null hypotheses for each gene, and the characterization of these multivariately differentially expressed genes is of great interest.

The *mvGST* package provides tools to identify GO terms (gene sets) that are differentially active (up or down) in multiple comparisons (contrasts) of interest. These tools are platform-independent, so results from Affymetrix, next-gen sequencing, or subsequent gene expression technology can be handled. Given a matrix of p-values (rows for genes, columns for contrasts), the *mvGST* package uses statistical methods from the field of meta-analysis to combine p-values for all genes annotated to each gene set, and then classify each gene set (or biological process) as being significantly more active (1), less active (-1), or not significantly differentially active (0) in each contrast of interest. Where multiple contrasts are of interest, each gene set is assigned to a profile (across contrasts) of differential activity. Tools are also provided for visualizing (in a GO graph) the gene sets classified to a given profile.

Contents

1 Sample Data

1.1 Obatoclax (Affymetrix)

These data, publicly available as GSE36149 from the Gene Expression Omnibus website (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36149`), were reported by Urtishak et al. (2013). Briefly, tissue samples were taken from two human leukemia cell lines (Line; R = RS4:11, S = SEM-K2), originally cultured from infant leukemia blood. Three treatments were compared (Trt; C = control, L = low-dose obatoclax, H = high-dose obatoclax). Gene expression was measured in each replicate using the Affymetrix Human Genome U133 Plus 2.0 Array.

For the purposes of *mvGST* package demonstration, suppose that a research objective is to identify biological processes differentially active in one or more of the following four comparisons: low dose vs. control in RS4:11 cell line, high dose vs. control in RS4:11 cell line, low dose vs. control in SEM-K2 cell line, or high dose vs. control in SEM-K2 cell line. Each gene individually can be tested for differential expression in these four comparisons by constructing four contrasts within the framework of the following model:

$$E\left[Y_{ijk}\right] \;=\; \mu + Trt_i + Line_j + TrtLine_{ij}$$

Here, Y_{ijk} is the log-scale expression of the gene in replicate k of Trt i in Line j. The sample code in Appendix A shows how the tools of the *limma* package can be used to obtain p-values for each of these contrasts, for each gene individually. The `obatoclax.pvals` object provided with the *mvGST* package contains these results:

```
> library(mvGST)
> data(mvGSTsamples)
> head(obatoclax.pvals)
```

```
           Low.RS4  High.RS4 Low.SEMK2 High.SEMK2
1007_s_at 0.5044534 0.2717903 0.5389786 0.68724607
1053_at   0.4665344 0.3121148 0.1162036 0.53931978
117_at    0.8135495 0.7929617 0.9584846 0.65778015
121_at    0.3793150 0.1299377 0.5644299 0.38994582
1255_g_at 0.1318970 0.2182358 0.5302079 0.30793280
1294_at   0.1416885 0.5124254 0.4539475 0.09247627
```

Note that (as a result of their construction in Appendix A) these are "one-sided" or "one-tailed" p-values, as expected by the *mvGST* package. Using the first comparison as an example, the null hypothesis is "expression in low dose and control are the same in the RS4 cell line" while the alternative hypothesis is "expression in low dose exceeds that in control, in the RS4 cell line." As a result, very small p-values in the first column of `obatoclax.pvals` are evidence supporting "expression in low dose is greater than expression in control, in the RS4 cell line" while very large p-values are evidence supporting "expression in low dose is less than expression in control, in the RS4 cell line."

Also, note that the row names correspond to genes and the column names correspond to contrasts of interest. The '.' in the contrast names are important; the 'RS4' and 'SEMK2' that follow the '.' are considered by the *mvGST* package to be strata in the comparisons of interest.

The *mvGST* package can use these gene-level results to identify biological processes differentially active (up or down) in one or more of the comparisons of interest. This is demonstrated in Section 3.1.

1.2 Parathyroid (RNA-Seq)

These data, publicly available in the *parathyroidSE* package, were reported by Haglund et al. (2012). Briefly, cell cultures of parathyroid tumors were taken from four patients (Patient; 1, 2, 3, 4), and exposed to one of three treatments (Treatment; DPN = diarylpropionitrite, OHT = 4-hydroxytamoxifen, Control). Samples were taken from each treated cell culture, and gene expression was measured using RNA-Seq, with ENSEMBL gene names used.

For the purposes of *mvGST* package demonstration, suppose that a research objective is to identify biological processes differentially active in one or more of the following three pairwise treatment comparisons: OHT vs. DPN, OHT vs. Control, and DPN vs. Control. Each gene individually can be tested for differential expression in these three comparisons by constructing three contrasts within the framework of the following model:

$$\log\left(E\left[Y_{ijk}\right]\right) \;=\; \mu + Patient_i + Treatment_j$$

Here, Y_{ijk} is the mapped sequence count of the gene in replicate k of Treatment j in Patient i. The sample code in Appendix B shows how the tools of the *DESeq2* package can be used to obtain p-values for each of these contrasts, for each gene individually. The `parathyroid.pvals` object provided with the *mvGST* package contains these results:

```
> head(parathyroid.pvals)
```

```
                   OHT_DPN OHT_Control DPN_Control
ENSG00000000003 0.77615783  0.77615783  0.77615783
ENSG00000000005 0.40872077  0.40872077  0.40872077
ENSG00000000419 0.08509812  0.08509812  0.08509812
ENSG00000000457 0.25179363  0.25179363  0.25179363
ENSG00000000460 0.36515566  0.36515566  0.36515566
ENSG00000000938 0.73911620  0.73911620  0.73911620
```

Again, note that (as a result of their construction in Appendix B) these are "one-sided" or "one-tailed" p-values, as expected by the *mvGST* package. Using the first comparison as an example, the null hypothesis is "expression in OHT and DPN are the same" while the alternative hypothesis is "expression in OHT is greater than expression in DPN." As a result, very small p-values in the first column of `parathyroid.pvals` are evidence supporting "expression in OHT is greater than expression in DPN" while very large p-values are evidence supporting "expression in OHT is less than expression in DPN."

Also, note that the row names correspond to genes and the column names correspond to contrasts of interest. The lack of '.' in the contrast names is important, as this tells the *mvGST* package that there are no strata in the comparisons of interest.

The *mvGST* package can use these gene-level results to identify biological processes differentially active (up or down) in one or more of the comparisons of interest. This is demonstrated in Section 3.2.

2 Multivariate and Directional Gene Set Testing

The statistical methods employed for gene set testing by the *mvGST* package are discussed in Stevens and Isom (2012) and Mecham (2014), and the key points are summarized in the bullet points below. Here, italics are used to indicate text cited from Stevens and Isom (2012), and the obatoclax example from Section 1.1 is used along with the biological process ontology as an example.

- "Multivariate":
  - Expression data of genes annotated to a particular GO term are used as proxy for the activity level of the corresponding biological process in a given treatment condition.
  - Multiple comparisons can be of simultaneous interest, as in seeking to identify biological processes that are more active in high dose than control in the RS4:11 cell line, but not differentially active between low dose and control in the RS4:11 cell line.
- "Directional":
  - *A gene is annotated to a biological process only when the gene's product "contributes to" the biological process (Hill et al. 2008). (Consequently, there is no annotation if a gene's product impedes or inhibits the biological process.) Then for a biological process to proceed, it is not necessarily sufficient for "at least one" of the contributing genes to be active. In fact, lower activity by any of the genes annotated to a biological process will "disturb" the biological process (Hill et al. 2008). Thus a more meaningful alternative in gene set testing would be that there is a consensus of activity among gene set members – for example, that there is "collective support" (Rice 1990) that the genes annotated to the biological process are more active in* high dose than control in the RS4:11 cell line.
  - Using one-sided p-values (i.e., from a one-sided test) allows statements of directional activity differences, such as that a biological process is more active in high dose than control in the RS4:11 cell line.
- For a given set of genes for each comparison of interest, the p-values for the genes can be meaningfully combined using Stouffer's method (Stouffer et al. 1949) from the field of meta-analysis, to arrive at a single p-value for the corresponding biological process. *While Fisher's p-value combination method was found previously to be most*

powerful (Fridley et al. 2010), it seems that it may be most powerful for a less meaningful alternative hypothesis. In cases were directionality is meaningful, consensus is the desired alternative, and there Stouffer's method has been shown superior to competing methods (Whitlock 2005).

3 *mvGST* Package Demonstration

3.1 Obatoclax Demonstration

3.1.1 Obatoclax: `profileTable`

The following code chunk uses the `obatoclax.pvals` object introduced in Section 1.1 to classify biological processes into multivariate profiles across the four comparisons of interest, while restricting attention to only biological processes with between 10 and 200 genes annotated thereto. Because the gene names (row names in `obatoclax.pvals`) are Affymetrix probe set identifiers from the hgu133plus2 array version (corresponding to the human genome), the `gene.ID`, `affy.chip`, and `organism` arguments are as specified in the call to function `profileTable`.

```
> library(hgu133plus2.db)
> test1 <- profileTable(obatoclax.pvals, gene.ID='affy',
   affy.chip='hgu133plus2', organism='hsapiens',
   minsize=10, maxsize=200)
> test1
```

```
 Low  High  RS4    SEMK2
  0    0     6916   6557
  0    1     12     305
  0   -1     34     150
  1    0     61     36
  1    1     26     51
 -1   -1     50     14
 -1    0     18     4
```

Recall the brief discussion of the contrast names in Section 1.1: the 'RS4' and 'SEMK2' that follow the '.' are considered by the *mvGST* package to be strata in the comparisons of interest. This can be seen in the above output, where the profiles were stratified by cell line (RS4 or SEMK2). Within each cell line, there were two contrasts of interest – Low−Control and High−Control. Within each comparison, each biological process is classified as a −1 if the tested contrast is significantly negative, as a 1 if the tested contrast is significantly positive, and as a 0 otherwise. Based on the preceding output, there are 4

biological processes in the SEMK2 cell line that are classified as -1 in the Low vs. Control comparison (meaning they are significantly less active in the low dosage group than in the control group) and as 0 in the High vs. Control comparison (meaning they have no significant activity difference between the high dosage and control groups). This can be thought of as the (-1, 0) multivariate profile.

3.1.2 Obatoclax: `pickOut`

Note that these 4 biological processes correspond to row 7 and stratum 2 of the `test1` table output above. The `pickOut` function can be used to see which biological processes these are, by picking them out of the table. The object returned by `pickOut` is a data frame, with the first two columns being the GO identifier and description, followed by columns of p-values for each of the comparisons of interest. In the following code chunk, the "head" of this object is trimmed to ensure it will fit on the vignette page:

```
> res <- pickOut(test1, row=7, col=2)
> as.data.frame(apply(head(res),2,strtrim,width=60))
```

```
       GO.ID                    GO.Description Low.RS4 High.RS4
1 GO:0021545         cranial nerve development     0.5      0.5
2 GO:0021602       cranial nerve morphogenesis     0.5      0.5
3 GO:0048745 smooth muscle tissue development     0.5      0.5
4 GO:2001222   regulation of neuron migration     0.5      0.5
          Low.SEMK2        High.SEMK2
1 0.994022294914266 0.897311878049151
2 0.986394924986406 0.830062802207708
3 0.989563785474559               0.5
4 0.991639371222025 0.680537480066285
```

These GO-level p-values are the result of Stouffer's combination of the p-values of all genes in the gene set, and are returned as "one-sided" or "one-tailed" p-values. For example, very small p-values in the `Low.SEMK2` column of the `pickOut` output are evidence supporting "activity in low dose is greater than activity in control, in the SEMK2 cell line" while very large p-values (as for the biological processes in the preceding output) are evidence supporting "activity in low dose is less than activity in control, in the SEMK2 cell line."

3.1.3 Obatoclax: `go2Profile`

If there are certain GO terms of interest, the `go2Profile` function can be used to identify their profile classification. Note that a profile of NA values (and a warning message) will be returned if a GO term of supposed interest (such as "GO:dummmy" in the following code chunk) is not among the gene sets that were actually tested:

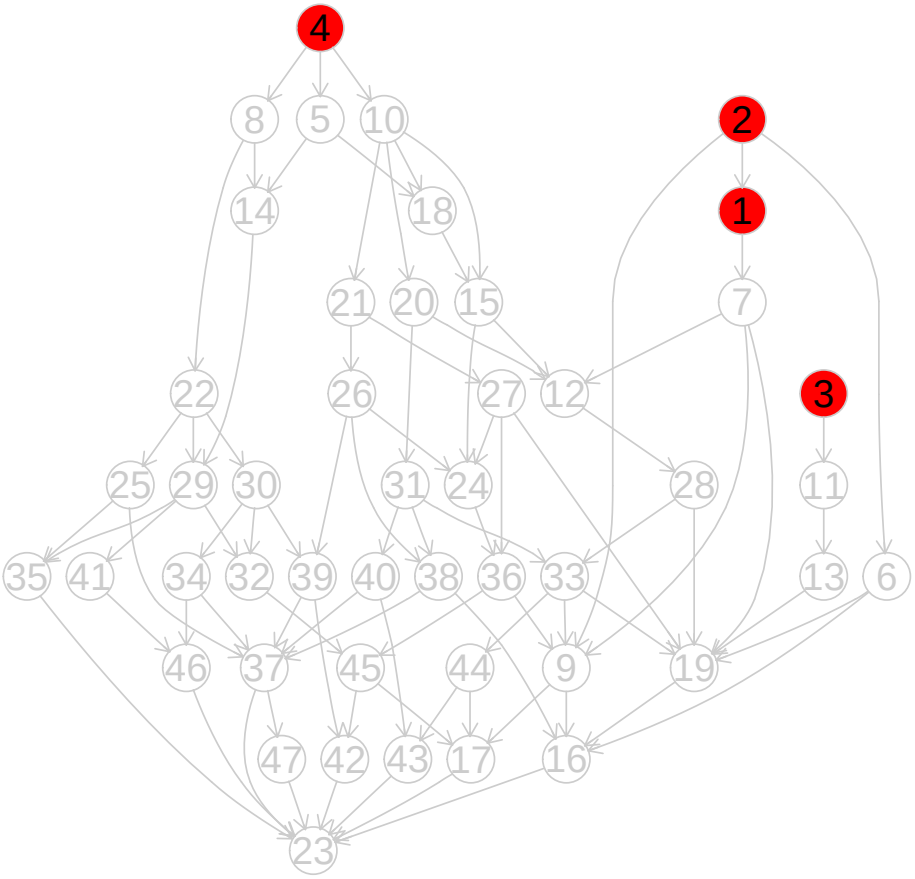
```
> temp <- go2Profile(c("GO:0002274", "GO:0002544", "GO:dummy"), test1)
> temp

$`GO:0002274`
     Low High RS4 SEMK2
[1,]  NA   NA   1     1

$`GO:0002544`
 Low  High  RS4    SEMK2
  0    0     1       1

$`GO:dummy`
     Low High RS4 SEMK2
[1,]  NA   NA   1     1
```

This output shows, for each of the requested GO terms, the (Low vs. Control, High vs. Control) multivariate profile to which they were classified, for each strata. For example, the row of the `$`GO:0002544`` output where `RS4` is 1 indicates the profile to which that GO term was classified in the RS4 cell line; it is the (0, 0) profile.

3.1.4 Obatoclax: `graphCell`

The `graphCell` function can be used to visualize (in a GO graph) the GO terms that were classified to a particular profile. The function name is derived from the fact that it graphs GO terms from a specified cell in the `profileTable` output. For example, the following code chunk visualizes (as red nodes) the 4 biological processes classified to the (-1, 0) multivariate profile (row 7 of `test1` table output) in the SEMK2 cell line (stratum 2).

```
> graphCell(test1, row=7, col=2, print.legend=FALSE, interact=FALSE)
```



In this preceding code chunk, the `bg.col` argument is used to force the GO graph portions of lesser interest (i.e., GO terms other than the 4 classified to the (-1, 0) multivariate profile) into the "background" by using a lighter color. The `print.legend` and `interact` arguments are set to `FALSE` just for convenience in creating this vignette. If set to `TRUE`, they allow interactivity with the graph (click on or near a node to see its description, ESC to end interactivity) and a printed summary of the graph (IDs and descriptions for all nodes).

3.2 Parathyroid Demonstration

3.2.1 Parathyroid: `profileTable`

The following code chunk uses the `parathyroid.pvals` object introduced in Section 1.2 to classify biological processes into multivariate profiles across the three comparisons of interest. Because the gene names (row names in `parathyroid.pvals`) are ENSEMBL identifiers and these are human genes, the `gene.ID` and `organism` arguments are as specified in the

call to function `profileTable`.

```
> test2 <- profileTable(parathyroid.pvals, gene.ID='ensembl',
    organism='hsapiens')
> test2
```

```
 OHT_DPN  OHT_Control  DPN_Control  BP
  0        0            0           13228
  1        1            1           1497
 -1       -1           -1           102
```

Because there was no '.' in the contrast names (see Section 1.2), there are no strata here – so `profileTable` adds a column `BP` for a single "pseudo-stratum" (biological processes). Based on the preceding output, there are 102 biological processes classified to the (-1, -1, -1) multivariate profile, with comparisons in order (OHT−DPN, OHT−Control, DPN−Control). In other words, there are 102 biological processes significantly less active in OHT than in DPN, less active in OHT than in Control, and less active in DPN than in Control. Put another way, for these 102 biological processes, activity is less in OHT than in DPN, and in DPN than in Control.

3.2.2 Parathyroid: `pickOut`

The `pickOut` function can be used to identify these 102 biological processes. In the following code chunk, the "head" of the resulting object is trimmed to ensure it will fit on the vignette page:

```
> res <- pickOut(test2, row=3, col=1)
> as.data.frame(apply(head(res),2,strtrim,width=60))
```

```
       GO.ID                                              GO.Description
1 GO:0000184 nuclear-transcribed mRNA catabolic process, nonsense-mediate
2 GO:0006986                                  response to unfolded protein
3 GO:0006000                                    fructose metabolic process
4 GO:0006082                                organic acid metabolic process
5 GO:0006083                                     acetate metabolic process
6 GO:0055114                                  oxidation-reduction process
         OHT_DPN.BP    OHT_Control.BP    DPN_Control.BP
1  0.99999568862456  0.99999568862456  0.99999568862456
2 0.998869676931867 0.998869676931867 0.998869676931867
3  0.99898942218904  0.99898942218904  0.99898942218904
4 0.991512440747123 0.991512440747123 0.991512440747123
```

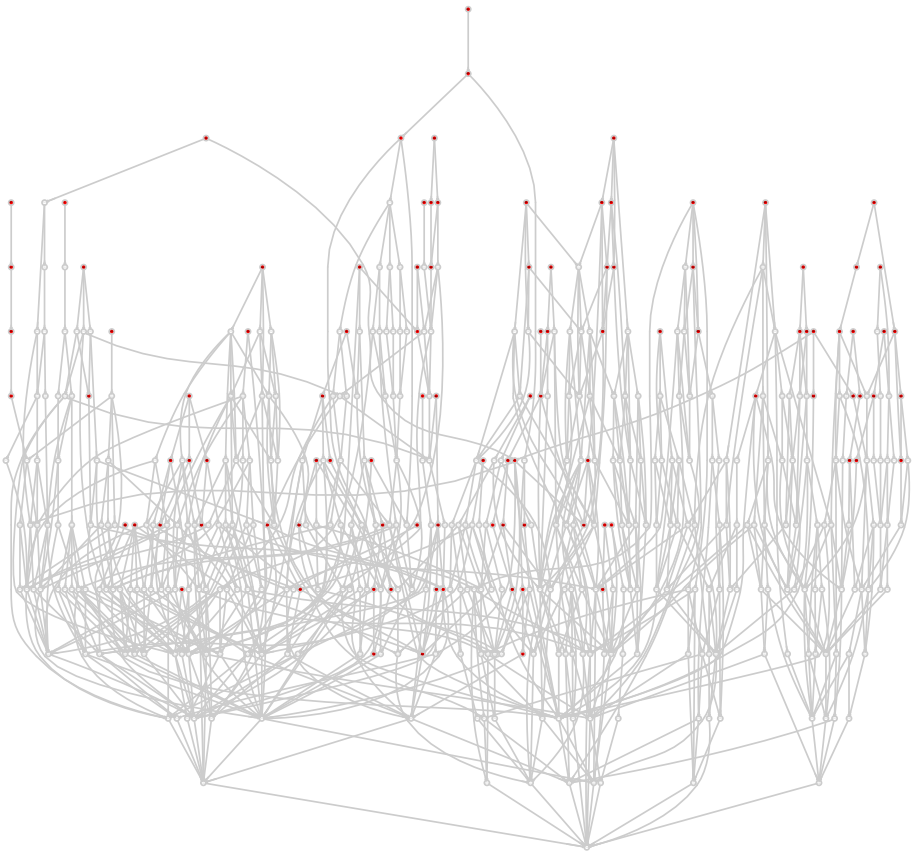
```
5 0.999143226012738 0.999143226012738 0.999143226012738
6 0.999709192115971 0.999709192115971 0.999709192115971
```

The three p-value columns in the preceding output represent "one-sided" or "one-tailed" p-values from the Stouffer's combination for each GO term in the comparison named. For example, very small p-values in the `DPN_Control.BP` column of the `pickOut` output (as for the biological processes in the preceding output) are evidence supporting "activity in DPN is greater than activity in Control" while very large p-values would be evidence supporting "activity in DPN is less than activity in Control."

3.2.3 Parathyroid: `graphCell`

The following code chunk visualizes (as red nodes) the 102 biological processes classified to the (-1, -1, -1) multivariate profile (row 3 of `test2` table output). Because there are no strata here, the column number is 1 (for the single "pseudo-stratum" `BP` in the `profileTable` output).

```
> graphCell(test2, row=3, col=1, print.legend=FALSE, interact=FALSE)
```



Such a large GO graph is probably most useful when the number of GO terms of interest is more manageable (i.e., smaller).

4 Multiple Comparison Adjustments

4.1 Default: False Discovery Rate

With thousands of GO terms tested in possibly multiple comparisons of interest, attention must be paid to multiple comparisons adjustments and thresholds of significance. The default in the *mvGST* package (and in the `profileTable` function in particular) is to control the FDR at 0.05 using the Benjamini-Yekutieli adjustment (Benjamini & Yekutieli 2001) within each comparison (contrast) of interest. This threshold can be modified using the `sig.level` argument of the `profileTable` function. This Benjamini-Yekutieli adjustment allows for dependence among p-values, which is certainly the case with nested GO terms.

4.2 Short Focus Level Demonstration: Parathyroid Data

For those interested in controlling the family-wise error rate at a specified level within the structure of the GO graph, the *mvGST* package includes an interface to the Short Focus Level method (Saunders 2014; Saunders, Stevens, and Isom 2014). The following code chunk is not run here to conserve computation time in the creation of this vignette document:

```
> test3 <- profileTable(parathyroid.pvals, gene.ID='ensembl',
   organism='hsapiens', mult.adj='SFL')
```

It takes about 4 hours on a desktop PC to run this full example.

Note that the Short Focus Level adjustment requires all ancestor and offspring nodes of the GO terms of interest to be included in the set of tested GO terms (Saunders 2014; Saunders, Stevens, and Isom 2014), so the `minsize` and `maxsize` arguments are not used.

For demonstration purposes of the `p.adjust.SFL` in this vignette, suppose we were only interested in the OHT vs. DPN comparison, and in the following set of GO terms that are ancestors of GO:0001775 and GO:0007275:

```
> library(GO.db)
> xx <- as.list(GOBPANCESTOR)
> ancs <- sort( union( xx$`GO:0001775`, xx$`GO:0007275` ) )[-1]
> GOids <- c('GO:0001775','GO:0007275', ancs)
> GOids

 [1] "GO:0001775" "GO:0007275" "GO:0009987" "GO:0032501" "GO:0032502"
 [6] "GO:0044699" "GO:0044707" "GO:0044763" "GO:0044767" "GO:0048856"
[11] "all"
```

We can get the p-values for the OHT vs. DPN comparison for each of these GO terms from the `test2` object created in Section 3.2.1:

```
> t <- is.element(test2$group.names, GOids)
> frame <- as.data.frame(test2$grouped.raw[t,])
> pvals <- frame$OHT_DPN.BP
> names(pvals) <- test2$group.names[t]
```

Note that the names of the p-values vector is the GO term IDs. Then the `p.adjust.SFL` function can be called:

```
> SFL.pvals <- p.adjust.SFL(pvals, ontology='BP', sig.level=.10)
> cbind(pvals, SFL.pvals)
```

```
                  pvals    SFL.pvals
GO:0001775 7.343604e-20 2.203081e-19
GO:0007275 2.839373e-58 8.277425e-44
GO:0009987 1.476203e-49 1.000000e+00
GO:0032501 9.522402e-55 1.000000e+00
GO:0032502 1.451138e-63 1.000000e+00
GO:0044699 8.277425e-44 8.277425e-44
GO:0044707 1.058243e-59 8.277425e-44
GO:0044763 3.720789e-42 1.116237e-41
GO:0044767 1.331981e-62 8.277425e-44
GO:0048856 2.953661e-65 1.000000e+00
```

Calling GO terms significant when `SFL.pvals` is less than 0.10 controls the family-wise error rate at 0.10, within the context of the GO graph.

References

[1] Benjamini Y. and Yekutieli D. (2001) "The control of the false discovery rate in multiple testing under dependencey," *Annals of Statistics* 29:1165-1188.

[2] Fisher R.A. (1932) *Statistical Methods for Research Workers*, 4th ed. Oliver and Boyd, Edinburgh.

[3] Fridley B.L., Jenkins G.D., and Biernacka J.M. (2010) "Self-contained gene set analysis of expression data: An evaluation of existing and novel methods," *PLoS ONE* 5(9):e12693.

[4] Haglund F., Ma R., Huss M., Sulaiman L., Lu M., Nilsson I.L., Hoog A., Juhlin C.C., Hartman J., and Larsson C. (2012) "Evidence of a Functional Estrogen Receptor in Parathyroid Adenomas," *The Journal of Clinical Endocrinology & Metabolism* 97(12):4631-9. PMID: 23024189

[5] Hill D.P., Smith B., McAndrews-Hill M.S., and Blake J.A. (2008) "Gene ontology annotations: what they mean and where they come from," *BMC Bioinformatics* 9(Suppl 5):S2.

[6] Mecham D. S. (2014) "mvGST: Multivariate and Directional Gene Set Testing," M.S. Project, Utah State University, Department of Mathematics and Statistics. `http://digitalcommons.usu.edu/gradreports/382/`

[7] Rice W. R. (1990) "A consensus combined p-value test and the family-wide significance of component tests," *Biometrics* 46(2):303-308.

[8] Saunders G. (2014) "Family-wise error rate control in QTL mapping and gene ontology graphs with remarks on family selection," Ph.D. thesis, Utah State University, Department of Mathematics and Statistics. `http://digitalcommons.usu.edu/etd/2164/`

[9] Saunders G., Stevens J.R., and Isom S.C. (2014) "A shortcut for multiple testing on the directed acyclic graph of Gene Ontology," *BMC Bioinformatics* (under review).

[10] Stevens J.R. and Isom S.C. (2012) "Gene Set Testing to Characterize Multivariately Differentially Expressed Genes," Proceedings of Conference on Applied Statistics in Agriculture, pp. 125-137.

[11] Stouffer S. A., Suchman E.A., DeVinney L.C., Star S. A., and Williams R.M.J. (1949) *The American Soldier, Vol. 1: Adjustment during Army Life*. Princeton University Press, Princeton.

[12] Urtishak K.A., Edwards A.Y., Wang L.S., Hudome A., et al. (2013) "Potent obatoclax cytotoxicity and activation of triple death mode killing across infant acute lymphoblastic leukemia," *Blood* 121(14):2689-703. PMID: 23393050

[13] Whitlock M.C. (2005) "Combining probability from independent tests: the weighted z-method is superior to Fisher's approach," *Journal of Evolutionary Biology* 18:1368-1373.

Appendix A: Constructing `obatoclax.pvals` object

The `obatoclax.pvals` object was introduced in Section 1.1.

After downloading the .CEL files from `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36149` and saving them to a directory (say "C:\folder\data"), the .CEL files were renamed as follows to facilitate interpretation of the constructed contrasts:

Sample	Trt	Line	Rep	CELfile
GSM881823	C	R	1	CR1.CEL
GSM881824	L	R	1	LR1.CEL
GSM881825	H	R	1	HR1.CEL
GSM881826	C	S	1	CS1.CEL
GSM881827	L	S	1	LS1.CEL
GSM881828	H	S	1	HS1.CEL
GSM881829	C	R	2	CR2.CEL
GSM881830	L	R	2	LR2.CEL
GSM881831	H	R	2	HR2.CEL
GSM881832	C	S	2	CS2.CEL
GSM881833	L	S	2	LS2.CEL
GSM881834	H	S	2	HS2.CEL

Then the following R code (which is not run in this vignette, simply to avoid needing the .CEL files with this *mvGST* package) was used on July 10, 2014 to construct the `obatoclax.pvals` object for the *mvGST* package:

```
> ### Objective is to identify gene sets differentially active
> ### in one or more of the following comparisons:
> ## G1 = RS4:11 cell line at low dose   (vs. control)
> ## G2 = RS4:11 cell line at high dose  (vs. control)
> ## G3 = SEM-K2 cell line at low dose   (vs. control)
> ## G4 = SEM-K2 cell line at high dose  (vs. control)
> #
> ## Read in data
> library(affy)
> data <- ReadAffy(celfile.path="C:\\folder\\data")
> eset <- exprs(rma(data))
> colnames(eset)
> # [1] "CR1.CEL" "CR2.CEL" "CS1.CEL" "CS2.CEL" "HR1.CEL" "HR2.CEL" "HS1.CEL"
> # [8] "HS2.CEL" "LR1.CEL" "LR2.CEL" "LS1.CEL" "LS2.CEL"
> #
> # Define simple function to convert two-tailed p-values to one-tailed,
> #  based on means of comparison groups
> # - this assumes null: Mean2=Mean1 and alt: Mean2>Mean1, and
```

```
> #   diff = Mean2-Mean1
> p2.p1 <- function(p,diff)
 {
   p1 <- rep(NA,length(p))
   t <- diff >=0
   p1[t] <- p[t]/2
   p1[!t] <- 1-p[!t]/2
   return(p1)
 }
> #
> # Define function to return one-tailed p-values for a specific contrast,
> # sorted in order of geneNames
> p1.ctrst <- function(ctr)
 {
   ctr <<- ctr
   ctrst <- makeContrasts(ctr, levels=design)
   fit.ctrst <- contrasts.fit(fit, ctrst)
   final.fit.ctrst <- eBayes(fit.ctrst)
   top.ctrst <- topTableF(final.fit.ctrst, n=nrow(eset))
   p1 <- p2.p1(top.ctrst$P.Value, top.ctrst[,1])
   gn <- rownames(top.ctrst)
   names(p1) <- gn
   t <- order(gn)
   return(p1[t])
 }
> #
> ## Fit model
> library(limma)
> trt <- rep(c('C','H','L'),each=4)
> line <- rep(rep(c('R','S'),each=2),3)
> design <- model.matrix(~0+trt:line)
> head(design)
> colnames(design) <- c('CR','HR','LR','CS','HS','LS')
> fit <- lmFit(eset, design)
> #
> ## Create contrasts
> # R: L vs. C   (G1)
> Low.RS4 <- p1.ctrst(ctr="LR-CR")
> # R: H vs. C   (G2)
> High.RS4 <- p1.ctrst("HR-CR")
> # S: L vs. C   (G3)
> Low.SEMK2 <- p1.ctrst("LS-CS")
> # S: H vs. C   (G4)
```

```
> High.SEMK2 <- p1.ctrst("HS-CS")
> #
> ## Assemble object for mvGST
> GN <- names(Low.RS4)
> o.pvals <- cbind(Low.RS4, High.RS4, Low.SEMK2, High.SEMK2)
> rownames(o.pvals) <- GN
> obatoclax.pvals <- o.pvals
```

Appendix B: Constructing `parathyroid.pvals` object

The `parathyroid.pvals` object was introduced in Section 1.2. The following R code (which is not run in this vignette, simply to avoid needing the *parathyroidSE* and *DESeq2* packages with this *mvGST* package) was used on July 10, 2014 to construct the `parathyroid.pvals` object for the *mvGST* package:

```
> # Load data
> library("parathyroidSE")
> data("parathyroidGenesSE")
> se <- parathyroidGenesSE
> colnames(se) <- colData(se)$run
> #
> # Fit model
> library("DESeq2")
> dds <- DESeqDataSet(se = se, design = ~ patient + treatment)
> design(dds) <- ~ patient + treatment
> ddsCtrst1 <- DESeq(dds)
> resultsNames(ddsCtrst1)
> #
> # Create contrasts
> res1 <- results(ddsCtrst1, contrast=c("treatment", "OHT", "DPN"))
> res2 <- results(ddsCtrst1, contrast=c("treatment", "OHT", "Control"))
> res3 <- results(ddsCtrst1, contrast=c("treatment", "DPN", "Control"))
> #
> # Assemble object for mvGST
> r1 <- res1[!is.na(res1$pvalue),]
> r2 <- res1[!is.na(res2$pvalue),]
> r3 <- res1[!is.na(res3$pvalue),]
> OHT_DPN <- p2.p1(r1$pvalue,r1$log2FoldChange)
> OHT_Control <- p2.p1(r2$pvalue,r2$log2FoldChange)
> DPN_Control <- p2.p1(r3$pvalue,r3$log2FoldChange)
> p.pvals <- cbind(OHT_DPN,OHT_Control,DPN_Control)
> GN <- rownames(r1)
> rownames(p.pvals) <- GN
> parathyroid.pvals <- p.pvals
```

This code is based on code found in the *DESeq2* package vignette. Note that the `p2.p1` function was defined in Appendix A .