# R/GSEPD Tutorial
# Gene Set Enrichment and Projection Displays

Karl Stamm – karl.stamm@gmail.com –

May 15, 2016

## 1 Introduction

`GSEPD` is a package for streamlining RNA-Seq data analysis, targeting complex samples with low replicate count such as human tissues, where all factors (metabolic, genetic, etc) cannot be controlled statistically. As a prerequisite, you need only your multiple samples' count data as a matrix whose columns are samples and rows are RefSeq NM and NR transcript identifiers. A second matrix associates sample identifiers with treatment/condition. Given both datasets, GSEPD will automate differential expression via DESeq2 [1], functional analysis via GOSeq [3], generate heatmaps of gene expression for significantly differentially expressed genes, and subsets of genes defined by the significantly enriched GO Terms.

After gene sets are detected from a differential expression analysis, the results are merged into a novel 'projection display' wherein each sample is scored according to each condition's multidimensional average expression. When the treatment samples are found to have a perturbed expression profile for a particular GO Term (geneset), all samples are scored on an axis ranging from control to treatment condition, and outliers or anomalous samples are readily apparent. Clustering quality of samples in a given geneset-space is quantified by the cluster's "Validity score" [2] and an empirical permutation derived p-value. GO Terms with more genes than samples in your comparison will randomly appear enriched, so the Segregation P-value is used to determine if a GO Term is significantly segregating your samples.

## 2 Usage

You'll need a prepared matrix of read-counts per transcript (Table 1.) You can use HTSeq or RSEM or coverageBed, or any other generation method, so long as it ends with a table of counts by transcript ID. This software comes pre-packaged with a dataset based on the IlluminaBodyMap project, counted with coverageBed. The second prerequisite is a metadata table associating sample identifiers with their test-condition and a nickname to annotate figures with (see Table 2 for the included sample). Alternatively, the manual/Vignette for DESeq2 describes how to generate a dataset object from HTSeq read counts, you can also initialize GSEPD with the DESeqDataSet object instead of the counts matrix.

### 2.1 Naming Conventions

In `R`, column names are not allowed to have spaces or certain special characters like + or . As your sample names are the columns of the count table, this implies your sample names may not have special characters or spaces. When you load a table with invalid column headers, `R` may silently convert invalid characters into periods, thus "Sample 1" becomes "Sample.1". If your metadata table (where samples must be annotated) matches the original count table, it won't match this converted name, and you'll either get an error about samples not being found, or worse, an apparently successful run with invalid data. In later stages of the `GSEPD` process your test conditions also become column headers, and spaces will cause an error before completion. It's better to compare groups `` `test' `` vs `` `control' `` than `` `lung tissue' `` vs `` `healthy(-ish) person' ``.

```
library(rgsepd)
```

```
## Loading required package:  DESeq2
## Loading required package:  S4Vectors
## Loading required package:  stats4
## Loading required package:  BiocGenerics
## Loading required package:  parallel
##
## Attaching package:  'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##    clusterExport, clusterMap, parApply, parCapply, parLapply,
##    parLapplyLB, parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##    IQR, mad, xtabs
## The following objects are masked from 'package:base':
##
##    Filter, Find, Map, Position, Reduce, anyDuplicated, append,
##    as.data.frame, cbind, colnames, do.call, duplicated, eval,
##    evalq, get, grep, grepl, intersect, is.unsorted, lapply,
##    lengths, mapply, match, mget, order, paste, pmax, pmax.int,
##    pmin, pmin.int, rank, rbind, rownames, sapply, setdiff, sort,
##    table, tapply, union, unique, unsplit
##
## Attaching package:  'S4Vectors'
## The following objects are masked from 'package:base':
##
##    colMeans, colSums, expand.grid, rowMeans, rowSums
## Loading required package:  IRanges
## Loading required package:  GenomicRanges
## Loading required package:  GenomeInfoDb
## Loading required package:  SummarizedExperiment
## Loading required package:  Biobase
## Welcome to Bioconductor
##
##    Vignettes contain introductory material; view with
##    'browseVignettes()'.  To cite Bioconductor, see
##    'citation("Biobase")', and for packages 'citation("pkgname")'.
## Loading required package:  goseq
## Loading required package:  BiasedUrn
## Loading required package:  geneLenDataBase
##
##
## Loading R/GSEPD 1.4.2
## Building human gene name caches
## 'select()' returned 1:1 mapping between keys and columns
## 'select()' returned 1:1 mapping between keys and columns
```

```
data("IlluminaBodymap" , package="rgsepd")
data("IlluminaBodymapMeta" , package="rgsepd")
```

Next we'll sub-select $1,000$ genes from this set for speed. Initialize the GSEPD object with the `GSEPD_INIT`

| | adipose.1 | adipose.2 | adipose.3 | adipose.4 | blood.1 | heart.1 | skeletal_muscle.1 |
|---|---|---|---|---|---|---|---|
| NM_000014 | 91717 | 91406 | 91667 | 91788 | 290 | 67285 | 15609 |
| NM_000015 | 7 | 7 | 7 | 0 | 0 | 3 | 0 |
| NM_000016 | 1395 | 1410 | 1505 | 1368 | 2592 | 15851 | 2255 |
| NM_000017 | 1400 | 1439 | 1393 | 1334 | 460 | 1992 | 2759 |
| NM_000018 | 9505 | 9432 | 9217 | 9780 | 4432 | 33252 | 15478 |
| NM_000019 | 5609 | 5580 | 5707 | 5622 | 1530 | 16616 | 11988 |
| NM_000020 | 5961 | 5625 | 5774 | 5862 | 98 | 1658 | 497 |
| NM_000021 | 3935 | 4092 | 3931 | 3964 | 12021 | 1570 | 2009 |
| NM_000022 | 335 | 267 | 292 | 286 | 723 | 144 | 75 |
| NM_000023 | 103 | 108 | 97 | 85 | 16 | 1463 | 13083 |
| NM_000024 | 2488 | 2608 | 2726 | 2579 | 3478 | 338 | 325 |
| NM_000025 | 0 | 0 | 0 | 0 | 0 | 8 | 0 |
| NM_000026 | 1182 | 1174 | 1148 | 1168 | 2806 | 1716 | 7589 |
| NM_000027 | 455 | 460 | 417 | 500 | 1475 | 376 | 108 |
| NM_000028 | 1379 | 1564 | 1442 | 1433 | 2683 | 7910 | 6610 |

Table 1: First few rows of the included IlluminaBodymap dataset. See `?IlluminaBodymap` for more details.

| | Sample | Condition | SHORTNAME |
|---|---|---|---|
| 1 | adipose.1 | A | AD1 |
| 2 | adipose.2 | A | AD2 |
| 3 | adipose.3 | A | AD3 |
| 4 | adipose.4 | A | AD4 |
| 5 | blood.1 | C | BL1 |
| 6 | blood.2 | C | BL2 |

Table 2: First few rows of the included IlluminaBodymapMeta dataset. See `?IlluminaBodymapMeta` for more details. These are easy to build with a spreadsheet, saved to csv and R's builtin `?read.csv`

function. Finally, to indicate which conditions will be tested as this dataset includes samples from 'condition' A, B, and C, we use `GSEPD_ChangeConditions`.

```
set.seed(1000) #fixed randomness
isoform_ids <- Name_to_RefSeq(c("GAPDH","HIF1A","EGFR","MYH7","CD33","BRCA2"))
rows_of_interest <- unique( c( isoform_ids ,
                    sample(rownames(IlluminaBodymap),
                    size=1000,replace=FALSE)))
G <- GSEPD_INIT(Output_Folder="OUT",
          finalCounts=round(IlluminaBodymap[rows_of_interest , ]),
          sampleMeta=IlluminaBodymapMeta,
            COLORS=c(blue="#4DA3FF",black="#000000",gold="#FFFF4D"))

## Keeping rows with counts (913 of 1006)

G <- GSEPD_ChangeConditions( G, c("A","B"))
```

This should only take a moment, and create the folder `OUT` which will hold your generated results. If you're familiar with R objects, you can explore the `G` object here and change default parameters, all set by `GSEPD_INIT`. We'll change some parameters now to demonstrate:

```
  G$MAX_Genes_for_Heatmap <- 25
  G$MAX_GOs_for_Heatmap <- 30
  G$MaxGenesInSet <- 12
  G$LIMIT$LFC <- log( 2.50 , 2 )
  G$LIMIT$HARD <- FALSE
```

Here we changed five default settings on the `G` master object. The parameters `MAX_Genes_for_Heatmap` and `MAX_GOs_for_Heatmap` cap how many rows you'll see on your final differential expression heatmap (Figure 6) and the projection HMA file (Figure 4), choosing the most significant rows so your figures are shorter. If you'd like a figure containing everything, make these values large.

The parameter `MaxGenesInSet` controls the size of evaluated GO-Terms. Default is 30, here we reduce it to 12 for speed. Calculating projection significance for large sets can be slow. Also see `MinGenesInSet` for culling niche gene sets. The goal here is to limit our results to gene sets which are not too broad and not too narrow.

The parameter `LIMIT$LFC` is the $\log_2$ minimum foldchange required for significance, here we've set it to require 250% expression up or down (default is 200%, at LFC=1). Finally `LIMIT$HARD` if `TRUE` (the default), means figures and plots will respect the specified p-value limits. Sometimes your comparison won't have any significant genes or GO-terms and later stages of the pipeline will error or quit. To force generation of all stages and plots of less-than-strictly significant sets, we have set the `LIMIT$HARD=FALSE`. You'll see messages during processing if very few genes would be strictly significant, as the system adjusts the threshold automatically. By default, limits are hard at $p = 0.05$.

Now we're ready to run the pipeline:

```
G <- GSEPD_Process( G )
```

```
## converting counts to integer mode
## Generating OUT/DESEQ.counts.Ax4.Bx8.csv
## converting counts to integer mode
## estimating size factors
## estimating dispersions
## gene-wise dispersion estimates
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing
## Generating OUT/DESEQ.Volcano.Ax4.Bx8.png
## Generating OUT/DESEQ.RES.Ax4.Bx8.csv
## dumping rows with raw PVAL>0.990 OR baseMean < 1 OR on excludes list
## Rows remaining:  866
## Converting identifiers with local DB
## Converting identifiers with biomart
## Generating OUT/DESEQ.RES.Ax4.Bx8.Annote.csv
## Too many genes found differentially expressed, changing the  threshold to raw p=0.000000
so we can use 25 genes in the heatmap.
## Generating OUT/HM.Ax4.Bx8.25.pdf
## Loading hg19 length data...
## Fetching GO annotations...
## Loading required package:  AnnotationDbi
## Calculating the p-values...
## 'select()' returned 1:1 mapping between keys and columns
## Loading hg19 length data...
## Fetching GO annotations...
## Calculating the p-values...
## 'select()' returned 1:1 mapping between keys and columns
## Loading hg19 length data...
## Fetching GO annotations...
## Calculating the p-values...
## 'select()' returned 1:1 mapping between keys and columns
## Generating OUT/GOSEQ.RES.Ax4.Bx8.GO.csv
## Written GO categories, now reverse mapping
## Generating OUT/GSEPD.RES.Ax4.Bx8.GO2.csv
## Generating OUT/GSEPD.RES.Ax4.Bx8.MERGE.csv
## Generating OUT/GSEPD.PCA_AG.Ax4.Bx8.pdf
## Generating OUT/GSEPD.PCA_DEG.Ax4.Bx8.pdf
## Generating OUT/SCA.GSEPD.Ax4.Bx8.pdf
```

```
## Calculating Projections and Segregation Significance
## Generating OUT/GSEPD.HMA.Ax4.Bx8.pdf
## Generating OUT/GSEPD.HMA.Ax4.Bx8.csv
## All Done!
```

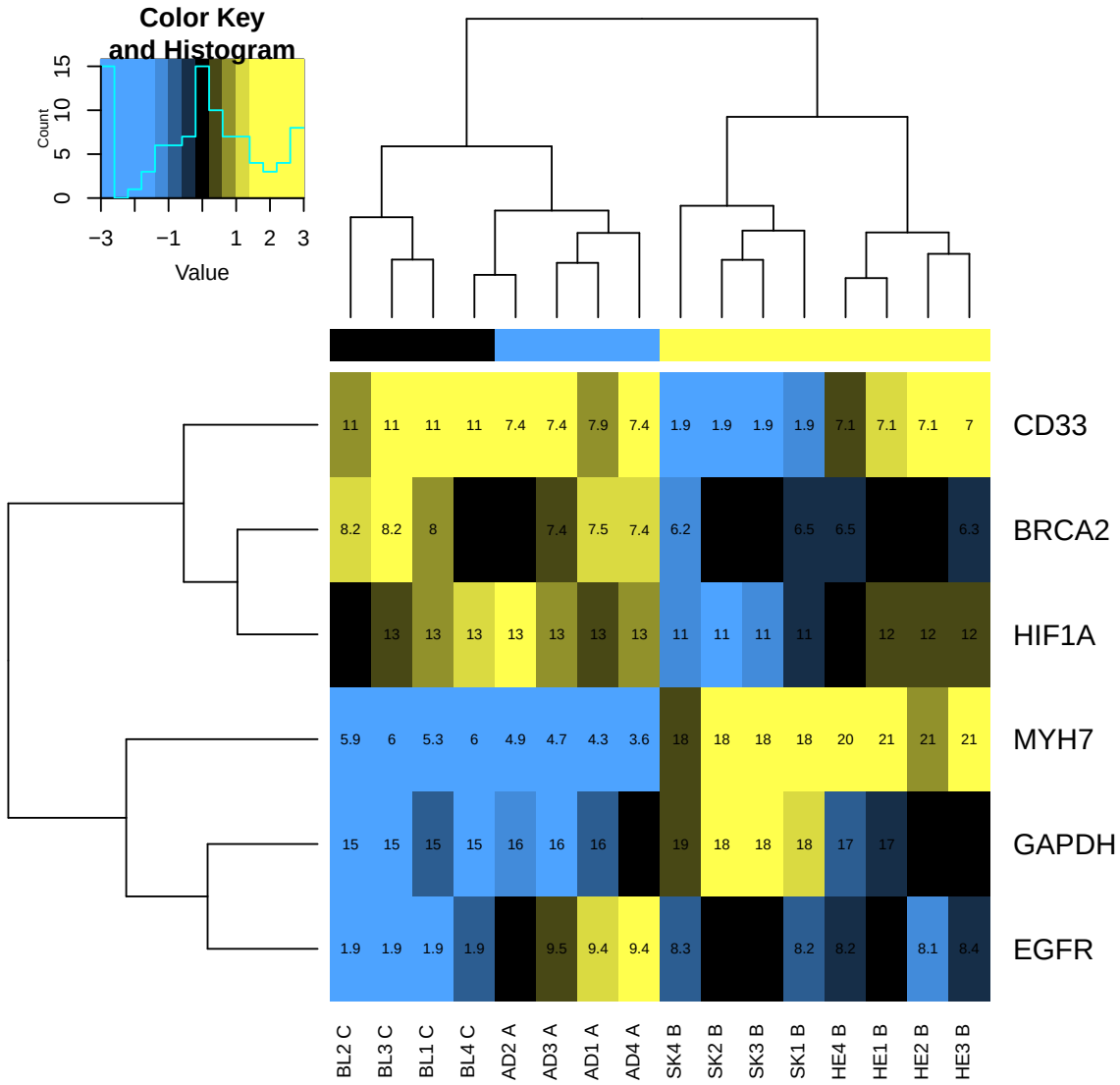
This step can take an hour on a full genome-wide dataset. If you change something and re-run GSEPD will reuse any files it finds with the same filename, so you don't have to wait for each step again, unless the filenames change. GSEPD's automation steps will convert gene identifiers, and GOSeq can take a few minutes as it runs on three differential expression sets (the upregulated, the downregulated, and the combined). All of these results are saved as CSV files under the `OUT` folder.

We save the G object from a `GSEPD_Process` routine, to retain information such as the normalized counts (wherein DESeq adjusted for library sequencing depths). This object will be passed to further visualizations, such as heatmaps and PCA.

```
  print(isoform_ids)

##       GAPDH       HIF1A        EGFR        MYH7        CD33       BRCA2
## "NM_002046" "NM_181054" "NM_201284" "NM_000257" "NM_001772" "NM_000059"

  GSEPD_Heatmap(G,isoform_ids)
```




```
  GSEPD_PCA_Plot(G)
```

```
## Generating OUT/GSEPD.PCA_AG.Ax4.Bx8.pdf, overwriting previous existing version.
## Generating OUT/GSEPD.PCA_DEG.Ax4.Bx8.pdf, overwriting previous existing version.
```

```
## pdf
##   2
```

# 3 Results

Several files are generated from each run. When `GSEPD_Process`is invoked the pre-specified conditions are compared, or when `GSEPD_ProcessAll`is invoked, each condition is tested against all others. For each comparison, files with the comparison listed in their filename are generated. For a condition named "A" with $N$ sam-

ples, versus "B" with $M$ samples, your normalized counts file will be written to `OUT\DESEQ.counts.AxN.BxM.csv`for example. If one of your conditions has the letter `x` in it, please change the delimiter with something like `G$C2T_Delimiter <- 'z'` or other unused character.

### 3.1 Heatmap Organizational Clustering

Each generated heatmap figure organizes the rows and columns such that similar profiles are adjacent. For this we use the default methods of `gplots::heatmap.2` which calls `hclust` on the supplied data. For gene heatmaps where you see numbers in each cell, representing the gene's expression value calculated by `DESeq2::varianceStabilizingTransformation` , the magnitude of a gene's expression might dominate the heirarchical clustering, so scaling is warranted. GSEPD will scale gene expression values within each row(gene) as a normal Gaussian by subtracting the row's mean and dividing out the standard deviation. Therefore it is dependent on the samples used in the figure, and heirarchical clustering with complete linkage is not guaranteed to be stable. To ensure some values of each specified color, the normalized color data is capped to a specifiable minimum and maximum (default 3) before heatmap.2's clustering is performed.

```
Annotated_Filtered <- read.csv("OUT/DESEQ.RES.Ax4.Bx8.Annote_Filter.csv",
                       header=TRUE,as.is=TRUE)
```

| REFSEQ | baseMean | Ax4 | Bx8 | X.X.Y. | LOG2.X.Y. | lfcSE | PVAL | PADJ | HGNC | ENTREZ |
|---|---|---|---|---|---|---|---|---|---|---|
| NM_000039 | 437.24 | 4.22 | 7.43 | 0.02 | -5.80 | 1.08 | 0.00 | 0.00 | APOA1 | 335 |
| NM_000257 | 434081.74 | 4.35 | 19.22 | 0.00 | -15.58 | 0.61 | 0.00 | 0.00 | MYH7 | 4625 |
| NM_000299 | 1797.05 | 12.76 | 7.09 | 51.94 | 5.70 | 0.18 | 0.00 | 0.00 | PKP1 | 5317 |
| NM_000362 | 53403.28 | 17.20 | 14.81 | 4.66 | 2.22 | 0.31 | 0.00 | 0.00 | TIMP3 | 7078 |
| NM_000465 | 236.19 | 6.75 | 8.07 | 0.33 | -1.62 | 0.42 | 0.00 | 0.00 | BARD1 | 580 |
| NM_000517 | 3945.15 | 13.69 | 10.10 | 10.53 | 3.40 | 0.34 | 0.00 | 0.00 | HBA2 | 3040 |
| NM_000526 | 149.09 | 9.23 | 2.11 | 907.21 | 9.83 | 0.68 | 0.00 | 0.00 | KRT14 | 3861 |
| NM_000723 | 3656.43 | 7.10 | 10.66 | 0.02 | -5.54 | 0.98 | 0.00 | 0.00 | CACNB1 | 782 |
| NM_000827 | 136.23 | 8.90 | 5.26 | 11.99 | 3.58 | 0.49 | 0.00 | 0.00 | GRIA1 | 2890 |
| NM_001003806 | 365.99 | 5.22 | 2.64 | 12.16 | 3.60 | 0.54 | 0.00 | 0.00 | TARP | 445347 |

Table 3: First few rows of OUT/DESEQ.RES.Ax4.Bx8.Annote_Filter.csv which contains the DESeq results, cropped for significant results, and annotated with gene names (the HGNC Symbol).

| X | category | over_represented_pvalue.x | Term.x | GOSEQ_DEG_Type | HGNC | LOG2.X.Y. | REFSEQ | Ax4 | Bx8 | PVAL | PADJ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | GO:0000266 | 0.05 | mitochondrial fission | Down | OPA1 | -1.51 | NM_130834 | 11.60 | 13.10 | 0.00 | 0.00 |
| 21 | GO:0000266 | 0.05 | mitochondrial fission | Down | ERBB4 | -4.30 | NM_001042599 | 5.74 | 9.86 | 0.00 | 0.00 |
| 22 | GO:0000271 | 0.05 | polysaccharide biosynthetic process | Up | GFPT1 | 1.79 | NM_001244710 | 13.70 | 11.90 | 0.00 | 0.00 |
| 23 | GO:0000271 | 0.05 | polysaccharide biosynthetic process | Up | GYG2 | 8.44 | NM_001079855 | 14.20 | 5.48 | 0.00 | 0.00 |
| 24 | GO:0000271 | 0.05 | polysaccharide biosynthetic process | Up | EXT2 | 1.01 | NM_000401 | 11.90 | 10.90 | 0.00 | 0.00 |
| 289 | GO:0001726 | 0.04 | ruffle | Up | WASF2 | 1.24 | NM_006990 | 13.00 | 11.70 | 0.00 | 0.00 |
| 290 | GO:0001726 | 0.04 | ruffle | Up | FGD3 | 1.35 | NM_033086 | 7.27 | 5.89 | 0.00 | 0.00 |
| 291 | GO:0001726 | 0.04 | ruffle | Up | PDLIM7 | -0.17 | NM_203352 | 9.78 | 9.61 | 0.70 | 0.77 |
| 292 | GO:0001726 | 0.04 | ruffle | Up | PLCG1 | 1.06 | NM_182811 | 12.20 | 11.10 | 0.00 | 0.00 |
| 293 | GO:0001726 | 0.04 | ruffle | Up | CLASP2 | -1.43 | NM_001207044 | 10.50 | 11.80 | 0.00 | 0.00 |
| 294 | GO:0001726 | 0.04 | ruffle | Up | DBNL | 1.42 | NM_001014436 | 11.70 | 10.30 | 0.00 | 0.00 |
| 295 | GO:0001726 | 0.04 | ruffle | Up | FGR | 2.46 | NM_001042747 | 10.50 | 8.02 | 0.00 | 0.00 |
| 296 | GO:0001726 | 0.04 | ruffle | Up | S100B | 3.94 | NM_006272 | 10.60 | 6.72 | 0.00 | 0.00 |
| 297 | GO:0001726 | 0.04 | ruffle | Up | SPRY2 | 0.67 | NM_005842 | 11.90 | 11.20 | 0.00 | 0.00 |
| 298 | GO:0001726 | 0.04 | ruffle | Up | ITGAV | -1.40 | NM_001145000 | 12.30 | 13.20 | 0.01 | 0.03 |

Table 4: First few rows of OUT/GSEPD.RES.Ax4.Bx8.MERGE.csv showing enriched GO Terms, and each terms' underlying gene expression averages per group. This data is central to the rgsepd package, defining the group centroids per GO-Term. It consists of the cross-product of the GO enrichment statistics and the DESeq differential expression and summarization.

## References

The citation for GSEPD is not available, as the article is under revision.

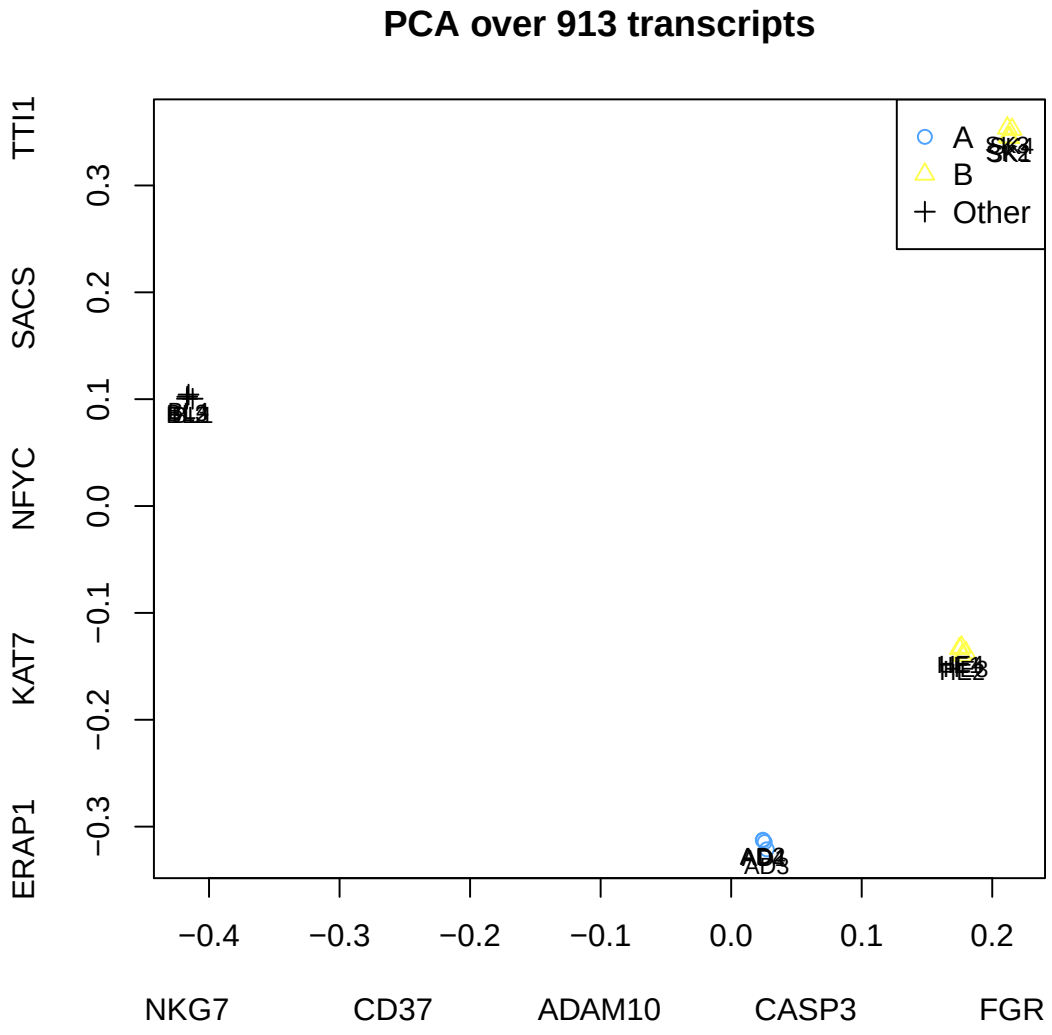



Figure 1: OUT\GSEPD.PCA_AG.Ax4.Bx8.pdf is the Principle Components Analysis of All Genes, from the comparison Ax4 vs Bx8 under run OUT. This function annotates the top four major genes underlying each principle component dimension along each axis (by maximum absolute weight). Samples from the tested condition are marked in the comparison colors, here blue and gold. All genes and all samples are included. A true outlier sample can direct the principle components away from the differentiating genes, and all tested samples can show as a single cluster.

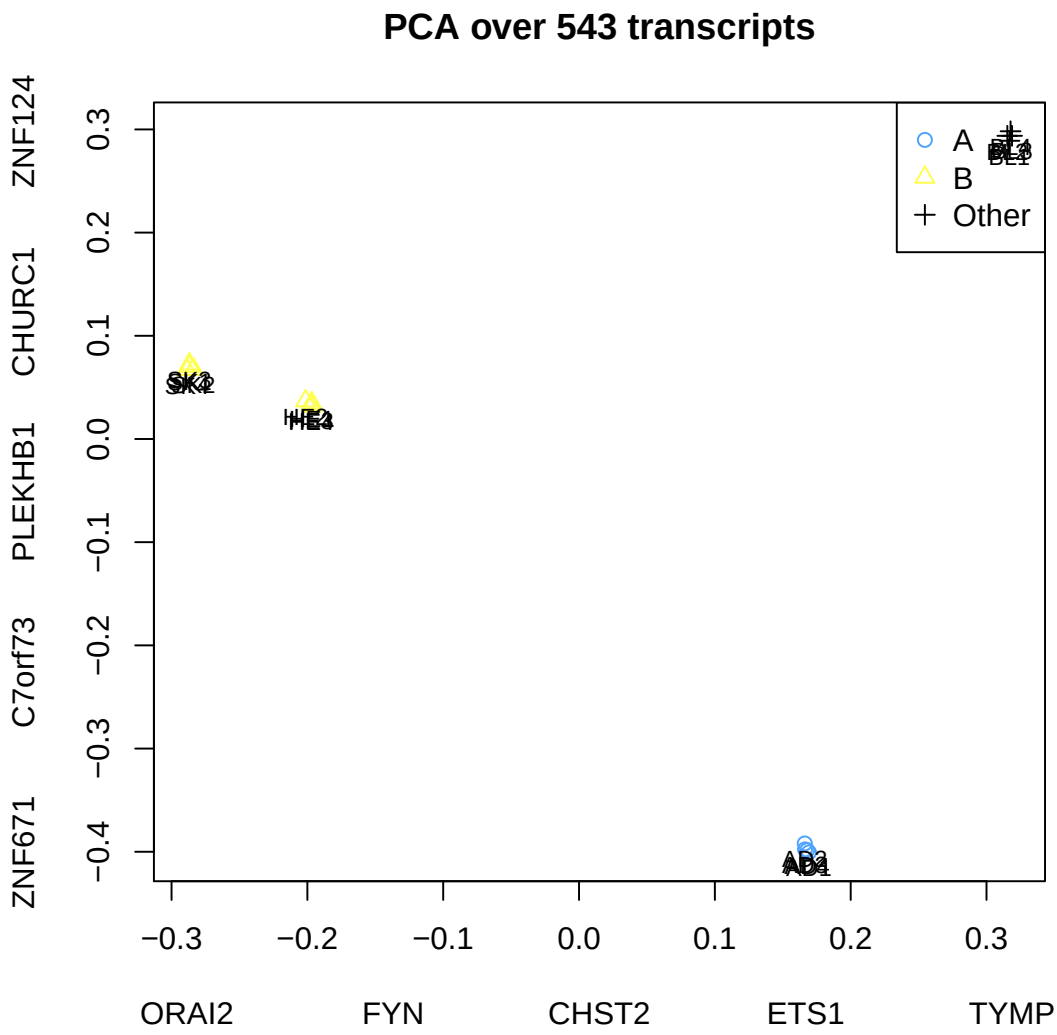



Figure 2: OUT\GSEPD.PCA_DEG.Ax4.Bx8.pdf is the Principle Components Analysis of only Differentially Expressed Genes, from the comparison Ax4 vs Bx8 under run OUT. This function annotates the top four major genes underlying each principle component dimension along each axis (by maximum absolute weight). Samples from the tested condition are marked in the comparison colors, here blue and gold, with the non-comparison samples marked as black 'Other'. **Because we're only reviewing genes found to be differentially expressed, this plot is guaranteed to show separation of your samples, sometimes spuriously.**

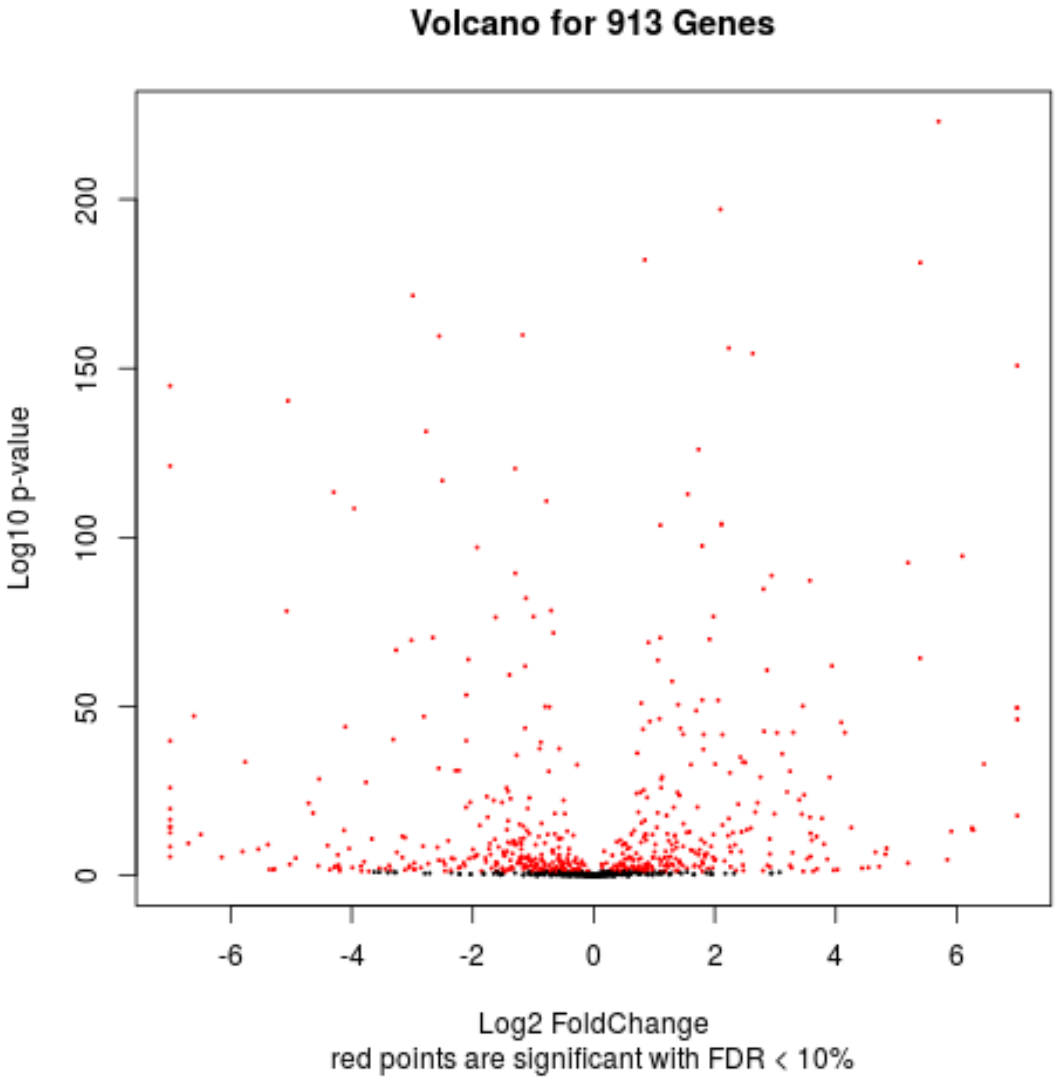



Figure 3: OUT\DESEQ.Volcano.Ax4.Bx8.png is the 'volcano' plot from the comparison Ax4 vs Bx8. Here, the horizontal axis is the relative fold-change between conditions, and the vertical is significance. A left-right balanced figure indicates similar numbers of genes found up and down-regulated. The bundled Illumina-Bodymap dataset does not have biological replicates, causing the Volcano plot to show too many significant genes.

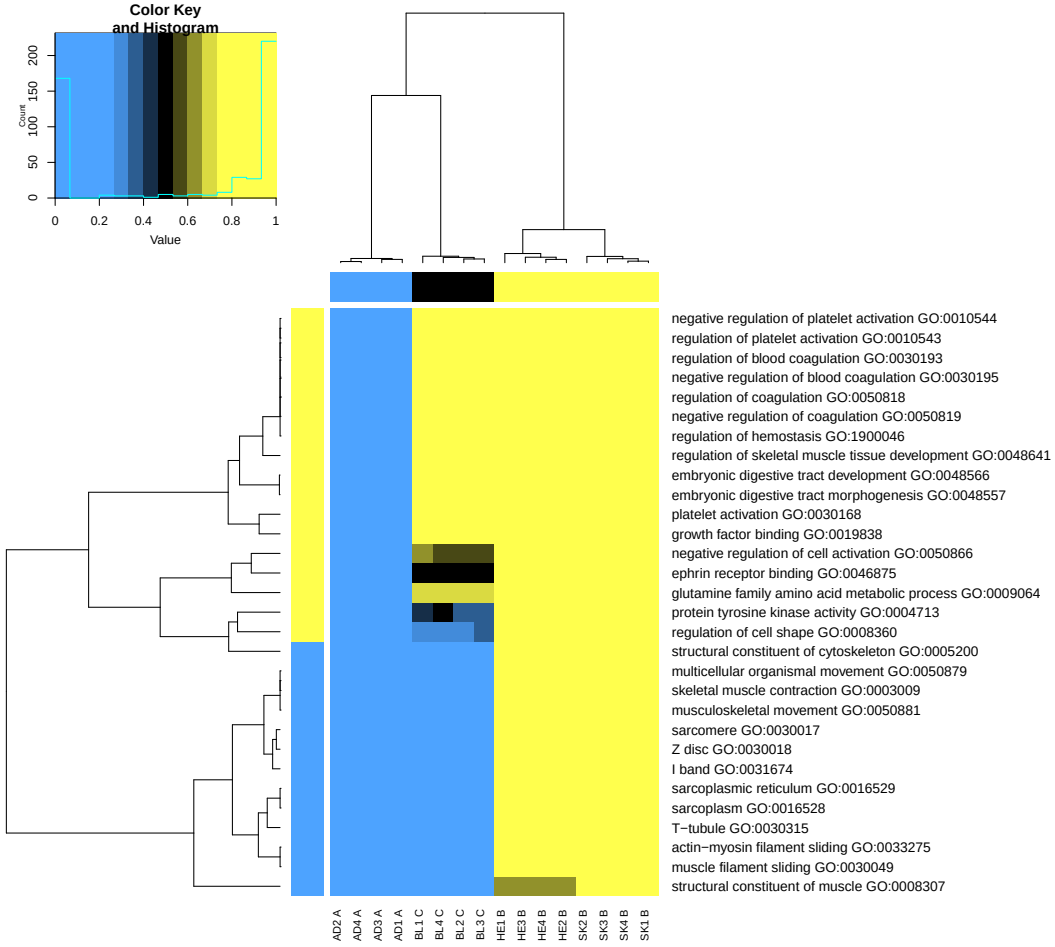



Figure 4: OUT\GSEPD.HMA.Ax4.Bx8.pdf is the projection display summary heatmap from the comparison Ax4 vs Bx8. Here, as in a normal heatmap, your samples are columns, and rows are GO Terms with significant segregation ability. All rows and columns are arranged to cluster. The color in each cell represents the sample's Alpha score for that GO Term, with blue indicating similarity to class 'A', and the gold indicating similarity to class 'B'. The unlabeled top row indicates the comparison categories, also seen on the sample labels along the bottom (A, B, or C). Any white dots indicate the sample was distant from the axis between conditions, so the color should be interpreted with caution or investigated further. The most interesting results from GSEPD are here, when unclassified samples (Blood/C) are scored as similar to either of the tested conditions on a geneset-by-geneset basis. Both the Alpha table and Beta table are summarized in the HMA figure.

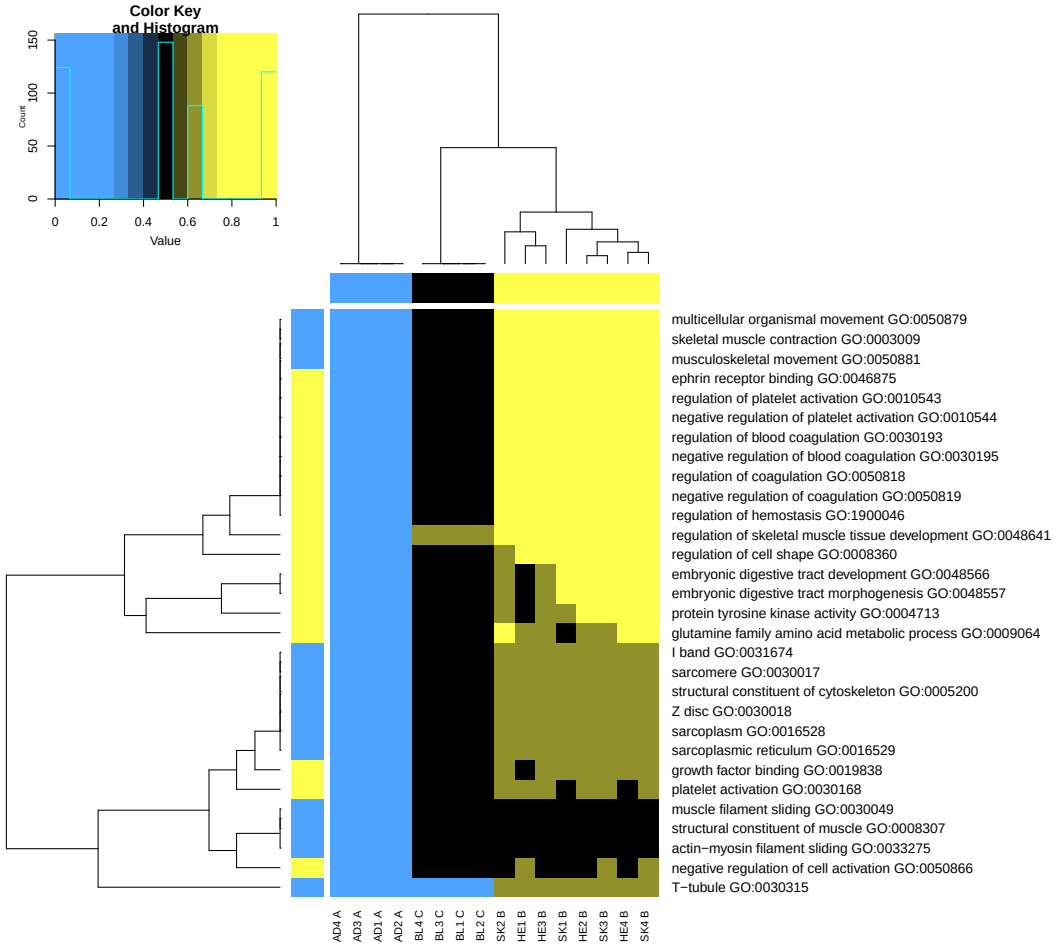



Figure 5: OUT\GSEPD.HMG.Ax4.Bx8.pdf is a simplified projection display summary heatmap from the comparison Ax4 vs Bx8. Here, as in a normal heatmap, your samples are columns, and rows are GO Terms with significant segregation ability. All rows and columns are arranged to cluster. The color in each cell represents the sample's Gamma(1/2) score for that GO Term, with blue indicating similarity to class 'A', and the gold indicating similarity to class 'B'. The unlabeled top row indicates the comparison categories, also seen on the sample labels along the bottom (A, B, or C). Black areas indicate the sample was distant from both conditions. The most interesting results from GSEPD are here, when unclassified samples (Blood/C) are scored as similar to either of the tested conditions on a geneset-by-geneset basis. Both the Gamma1 table and Gamma2 table are summarized in this HMG figure.

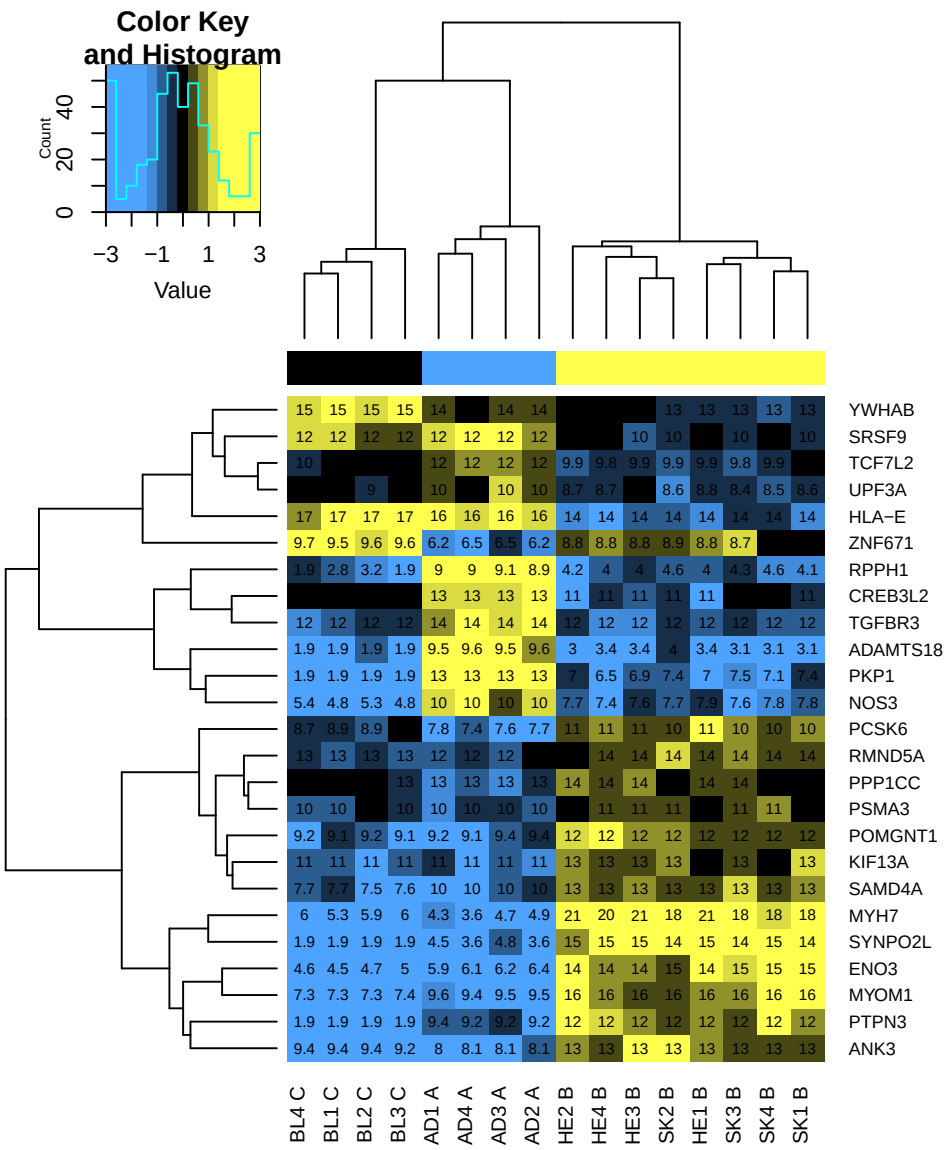



Figure 6: OUT\HM.Ax4.Bx8.25.pdf is a heatmap of your comparisons with full expression details for those genes found significantly expressed. The number of genes is given in the filename. Each row is scaled such that the lowest values are blue and the highest are gold (default colors are changable in GSEPD_INIT). The numbers within each cell are the expression values as variance-stabilized normalized counts, similar to a log transform, provided by DESeq2. Across the top, between the sample clustering dendrogram and the heatmap itself is a colorbar annotating which samples belong to which comparison group, for this differential expression blue vs gold. Black corresponds to extraneous samples, contributing context. Other variations of this plot with less information are generated as HMS files (compared samples only) and HM- files (minimal figure without details).

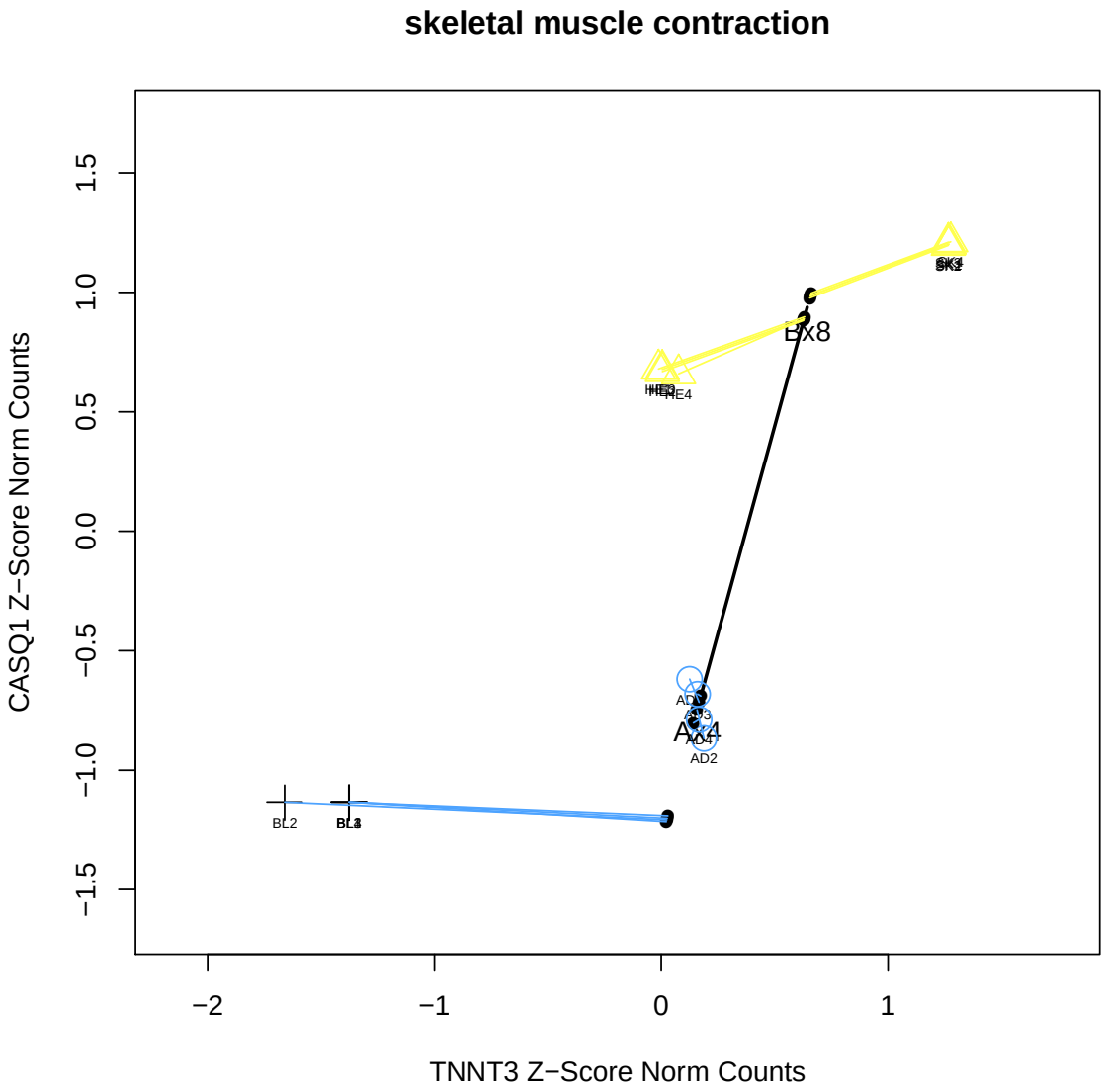



Figure 7: OUT\SCGO\GSEPD.Ax4.Bx8.GO0003009.pdf First page of the projections file for one GO Term. All significant sets have this figure generated displaying the central axis between two groups as a black line (with ends annotated Ax4 and Bx8), and each sample's closest point along the line. These files have half as many pages as genes in the set, so we can display pairs as two-dimensional scatterplots. It's a workaround to the problem of displaying an N-dimensional scatterplot. For sets with fewer than 11 genes, full pairings are viewable with the Pairs file, Figure 8

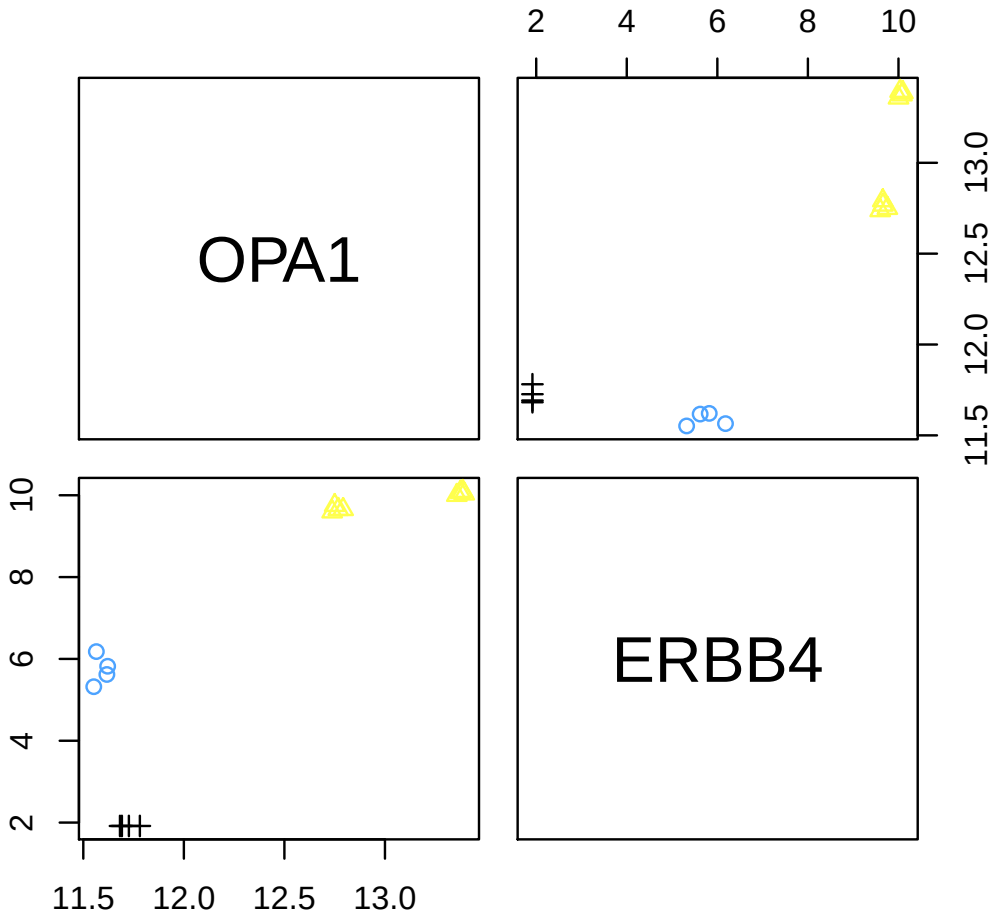



Figure 8: OUT\SCGO\Pairs.Ax4.Bx8.GO0000266.pdf The Pairs file is generated for significant GO terms with between two and ten genes, making it easy to review correlations or subgroups of gene expression among low-level gene sets.

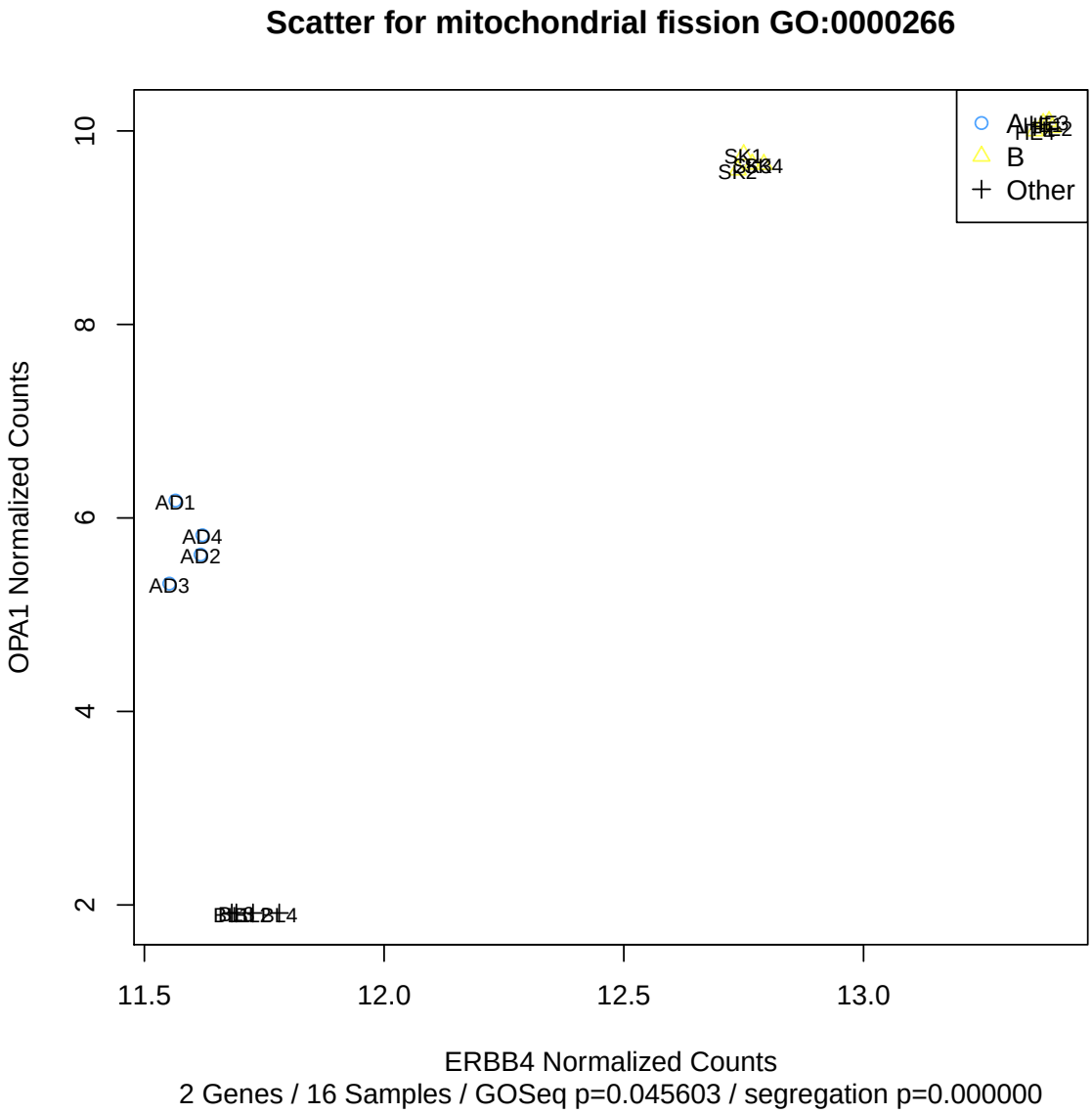



Figure 9: OUT\SCGO\Scatter.Ax4.Bx8.GO0000266.pdf The Scatter file is similar to the PCA files, but the gene set is restricted to those within a given, significant, GO Term. As in the PCA file, genes driving the principle components are annotated along the axes. At the bottom of the figure, statistics are given pertaining to this group. All such data is available in tables, but this Scatter plot helps the user see which samples behave like which groups. Depending on the number of genes relative to the number of samples, different PCA formulas may be used. For a set with only two genes, this file can directly display the expression values as normalized counts, the same as found on the expression heatmap (Fig. 6.)

| | adipose.1 | adipose.2 | adipose.3 | adipose.4 | blood.1 | heart.1 | skeletal_muscle.1 |
|---|---|---|---|---|---|---|---|
| GO:0000266 | 0.02 | 0.01 | -0.05 | 0.02 | -0.21 | 1.16 | 0.84 |
| GO:0001726 | 0.02 | 0.01 | -0.03 | -0.01 | 0.40 | 1.07 | 0.92 |
| GO:0002011 | 0.02 | -0.01 | -0.02 | 0.00 | 0.62 | 1.20 | 0.81 |
| GO:0002053 | -0.07 | 0.01 | 0.04 | 0.02 | 1.56 | 1.72 | 0.59 |
| GO:0002377 | 0.01 | -0.00 | -0.00 | -0.01 | 0.04 | 0.90 | 1.11 |
| GO:0002440 | -0.02 | 0.00 | 0.03 | -0.02 | -0.03 | 0.97 | 1.03 |
| GO:0002475 | -0.00 | -0.00 | 0.00 | 0.01 | 0.39 | 0.84 | 1.18 |
| GO:0002483 | -0.01 | -0.02 | 0.02 | 0.02 | 0.55 | 0.74 | 1.27 |
| GO:0002637 | 0.00 | 0.00 | -0.00 | -0.00 | -0.42 | 0.98 | 0.99 |
| GO:0002639 | 0.00 | 0.00 | -0.00 | -0.00 | -0.42 | 0.98 | 0.99 |

Table 5: First ten rows of OUT/GSEPD.Alpha.Ax4.Bx8.csv showing the group projection scores for each sample, these directly correspond to the colors in the HMA file. Where the HMA displays only significant sets, the Alpha table continues for all tested GO Terms. Both the Alpha table and Beta table are summarized in Figure 4.

| | adipose.1 | adipose.2 | adipose.3 | adipose.4 | blood.1 | heart.1 | skeletal_muscle.1 |
|---|---|---|---|---|---|---|---|
| GO:0000266 | 0.06 | 0.02 | 0.04 | 0.00 | 0.51 | 0.08 | 0.10 |
| GO:0001726 | 0.01 | 0.01 | 0.01 | 0.01 | 0.46 | 0.18 | 0.18 |
| GO:0002011 | 0.01 | 0.02 | 0.02 | 0.03 | 1.03 | 0.31 | 0.30 |
| GO:0002053 | 0.09 | 0.00 | 0.05 | 0.04 | 0.35 | 0.53 | 0.05 |
| GO:0002377 | 0.01 | 0.00 | 0.01 | 0.01 | 0.67 | 0.14 | 0.13 |
| GO:0002440 | 0.04 | 0.03 | 0.06 | 0.04 | 0.41 | 0.21 | 0.21 |
| GO:0002475 | 0.01 | 0.04 | 0.02 | 0.02 | 0.69 | 0.22 | 0.28 |
| GO:0002483 | 0.02 | 0.04 | 0.03 | 0.02 | 0.52 | 0.20 | 0.25 |
| GO:0002637 | 0.01 | 0.00 | 0.01 | 0.00 | 0.81 | 0.11 | 0.09 |
| GO:0002639 | 0.01 | 0.00 | 0.01 | 0.00 | 0.81 | 0.11 | 0.09 |

Table 6: First ten rows of OUT/GSEPD.Beta.Ax4.Bx8.csv showing the linear divergence (distance to axis) for each sample, high values here would be annotated with white dots on the HMA file to indicate that a sample is not falling on the axis. Non-tested samples are expected to frequently have high values here, the C group was not part of the A vs B comparison. Where the HMA displays only significant sets, the Beta table continues for all tested GO Terms. Both the Alpha table and Beta table are summarized in Figure 4.

# References


[1] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2". In: *bioRxiv* (2014). DOI: 10.1101/002832. URL: http://dx.doi.org/10.1101/002832.

[2] Andrew Rosenberg and Julia Hirschberg. *V-Measure: A conditional entropy-based external cluster evaluation measure*. Department of Computer Science Columbia University New York, NY 10027. 2007. URL: http://www1.cs.columbia.edu/~amaxwell/pubs/v_measure-emnlp07.pdf.

[3] Matthew Young et al. "Gene ontology analysis for RNA-seq: accounting for selection bias". In: *Genome Biology* 11.2 (2010), R14. ISSN: 1465-6906. DOI: 10.1186/gb-2010-11-2-r14. URL: http://genomebiology.com/2010/11/2/R14.


# Session Info

```
  sessionInfo()

## R version 3.3.0 (2016-05-03)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.4 LTS
##
## locale:
```

| | adipose.1 | adipose.2 | adipose.3 | adipose.4 | blood.1 | heart.1 | skeletal_muscle.1 |
|---|---|---|---|---|---|---|---|
| GO:0000266 | 0.06 | 0.02 | 0.06 | 0.02 | 0.49 | 1.16 | 0.84 |
| GO:0001726 | 0.03 | 0.03 | 0.04 | 0.02 | 1.10 | 1.15 | 1.01 |
| GO:0002011 | 0.02 | 0.02 | 0.02 | 0.03 | 1.13 | 1.23 | 0.86 |
| GO:0002053 | 0.11 | 0.01 | 0.06 | 0.04 | 1.60 | 1.79 | 0.59 |
| GO:0002377 | 0.02 | 0.01 | 0.01 | 0.02 | 1.09 | 0.92 | 1.13 |
| GO:0002440 | 0.10 | 0.07 | 0.16 | 0.10 | 0.96 | 1.09 | 1.15 |
| GO:0002475 | 0.01 | 0.06 | 0.03 | 0.03 | 1.16 | 0.91 | 1.27 |
| GO:0002483 | 0.03 | 0.06 | 0.05 | 0.04 | 0.98 | 0.80 | 1.33 |
| GO:0002637 | 0.01 | 0.00 | 0.01 | 0.00 | 0.97 | 0.99 | 0.99 |
| GO:0002639 | 0.01 | 0.00 | 0.01 | 0.00 | 0.97 | 0.99 | 0.99 |

Table 7: First ten rows of OUT/GSEPD.HMG1.Ax4.Bx8.csv showing the z-scaled distance to the Group1 centroid for each sample, these directly correspond to the colors in the HMG file. Where the HMG displays only significant sets, the Gamma table continues for all tested GO Terms. Both the Gamma1 and Gamma2 tables are summarized in Figure 5. Distance is normalized to dimensionality by scaling between the centroids. Thus a score of 0 means the sample resides on the centroid, and a score of 1 means it resides on the opposite class centroid, or equidistant.

| | adipose.1 | adipose.2 | adipose.3 | adipose.4 | blood.1 | heart.1 | skeletal_muscle.1 |
|---|---|---|---|---|---|---|---|
| GO:0000266 | 0.98 | 0.99 | 1.05 | 0.98 | 1.29 | 0.18 | 0.18 |
| GO:0001726 | 0.98 | 0.99 | 1.03 | 1.01 | 1.19 | 0.41 | 0.41 |
| GO:0002011 | 0.98 | 1.01 | 1.02 | 1.00 | 1.02 | 0.34 | 0.33 |
| GO:0002053 | 1.07 | 0.99 | 0.96 | 0.98 | 0.66 | 0.88 | 0.42 |
| GO:0002377 | 0.98 | 1.00 | 1.00 | 1.01 | 1.45 | 0.24 | 0.23 |
| GO:0002440 | 1.02 | 1.00 | 0.98 | 1.02 | 1.40 | 0.49 | 0.50 |
| GO:0002475 | 1.00 | 1.01 | 1.00 | 0.99 | 1.25 | 0.38 | 0.49 |
| GO:0002483 | 1.01 | 1.02 | 0.98 | 0.98 | 0.92 | 0.40 | 0.47 |
| GO:0002637 | 1.00 | 1.00 | 1.00 | 1.00 | 1.67 | 0.12 | 0.09 |
| GO:0002639 | 1.00 | 1.00 | 1.00 | 1.00 | 1.67 | 0.12 | 0.09 |

Table 8: First ten rows of OUT/GSEPD.HMG2.Ax4.Bx8.csv showing the z-scaled distance to the Group2 centroid for each sample, these directly correspond to the colors in the HMG file. Where the HMG displays only significant sets, the Gamma table continues for all tested GO Terms. Both the Gamma1 and Gamma2 tables are summarized in Figure 5. Distance is normalized to dimensionality by scaling between the centroids. Thus a score of 0 means the sample resides on the centroid, and a score of 1 means it resides on the opposite class centroid, or equidistant.

```
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] org.Hs.eg.db_3.3.0         AnnotationDbi_1.34.2
##  [3] rgsepd_1.4.2               goseq_1.24.0
##  [5] geneLenDataBase_1.8.0      BiasedUrn_1.07
##  [7] DESeq2_1.12.2              SummarizedExperiment_1.2.2
##  [9] Biobase_2.32.0             GenomicRanges_1.24.0
## [11] GenomeInfoDb_1.8.2         IRanges_2.6.0
## [13] S4Vectors_0.10.0           BiocGenerics_0.18.0
```

```
## [15] xtable_1.8-2
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5             locfit_1.5-9.1
##  [3] lattice_0.20-33         GO.db_3.3.0
##  [5] gtools_3.5.0            Rsamtools_1.24.0
##  [7] Biostrings_2.40.0       plyr_1.8.3
##  [9] chron_2.3-47            acepack_1.3-3.3
## [11] RSQLite_1.0.0           evaluate_0.9
## [13] highr_0.6               ggplot2_2.1.0
## [15] gplots_3.0.1            zlibbioc_1.18.0
## [17] GenomicFeatures_1.24.2  gdata_2.17.0
## [19] data.table_1.9.6        annotate_1.50.0
## [21] hash_2.2.6              rpart_4.1-10
## [23] Matrix_1.2-6            splines_3.3.0
## [25] BiocParallel_1.6.2      geneplotter_1.50.0
## [27] stringr_1.0.0           foreign_0.8-66
## [29] RCurl_1.95-4.8          biomaRt_2.28.0
## [31] munsell_0.4.3           rtracklayer_1.32.0
## [33] mgcv_1.8-12             nnet_7.3-12
## [35] gridExtra_2.2.1         Hmisc_3.17-4
## [37] XML_3.98-1.4            GenomicAlignments_1.8.0
## [39] bitops_1.0-6            grid_3.3.0
## [41] nlme_3.1-128            gtable_0.2.0
## [43] DBI_0.4-1               magrittr_1.5
## [45] formatR_1.4             scales_0.4.0
## [47] KernSmooth_2.23-15      stringi_1.0-1
## [49] XVector_0.12.0          genefilter_1.54.2
## [51] latticeExtra_0.6-28     Formula_1.2-1
## [53] RColorBrewer_1.1-2      tools_3.3.0
## [55] survival_2.39-4         colorspace_1.2-6
## [57] cluster_2.0.4           caTools_1.17.1
## [59] knitr_1.13
```