PGSEA Example Workflow Using Expression Data from GEO

Karl Dykema

April 30, 2008

Laboratory of Computational Biology, Van Andel Research Institute

1 Get Your Data

To test for enrichment of gene sets using our package, you only need two basic items: a gene expression data set and a list of gene sets. Your expression data must either already be in ratio form, or include reference samples by which to generate a ratio on the fly. As you will see later in this example, the "reference" argument will be used because the data is not already in the form of a ratio ("experiment"/"reference").

Online gene expression data repositories represent a great wealth of knowledge and this example workflow will help guide your own analysis by using simple Bioconductor tools to exploit data from such a source, like GEO. To start off we must load the GEOquery library, and then download and parse the data with the getGEO command.

```
> library(PGSEA)
> library(GEOquery)
> library(GSEABase)
> gse <- getGEO("GSE7023", GSEMatrix = TRUE)
Found 1 file(s)
GSE7023_series_matrix.txt.gz
File stored at:
/tmp/RtmpptsBNU/GPL4866.soft
```

Next we process the raw data from GEO to generate the phenoData for the ExpressionSet, an object that holds gene expression data within Bioconductor. See the GEOquery vignette for more information.

```
> subtype <- gsub("\\.", "_", gsub("subtype: ", "", phenoData(gse[[1]])$characteristics_ch1)
> pheno <- new("AnnotatedDataFrame", data = data.frame(subtype),
+ varMetadata = data.frame(labelDescription = "subtype"))
> rownames(pheno@data) <- colnames(exprs(gse[[1]]))
> eset <- new("ExpressionSet", exprs = exprs(gse[[1]]), phenoData = pheno)</pre>
```

Next we must load the gene sets to be used, in this case we will use some included example gene sets that we were able to curate from various publications. This list of gene sets is by no means an extensive or comprehensive list; it is merely included for example purposes. Information on the creation of each gene set has not been included, but may be available upon request.

```
> data(VAIgsc)
> details(VAIgsc[[1]])
setName: CMYC.1 down
geneIds: 27, 101, ..., 377582 (total: 89)
geneIdType: EntrezId
collectionType: ExpressionSet
setIdentifier: 185b7a75-038d-4a5e-c498-d25297d24c9d
description:
organism:
pubMedIds: 16273092
urls:
contributor: Karl Dykema <karl.dykema@vai.org>
setVersion: 0.0.1
creationDate: Fri Sep 14 12:00:24 2007
```

2 Run PGSEA

Finally, we are to the point where we can run PGSEA.

```
> pg <- PGSEA(eset, VAIgsc, ref = which(subtype == "NO"))</pre>
```

The results come back in the form of a matrix. Here is a look at a portion of it. A result of "NA" means that the test did not pass the significance threshold.

> pg[5:8, 5:8]

	GSM162152	GSM162153	GSM162154	GSM162155
HRAS.1 down	NA	NA	NA	NA
HRAS.1 up	NA	3.814831	3.045554	NA
SRC.1 down	NA	NA	NA	NA
SRC.1 up	NA	NA	NA	NA

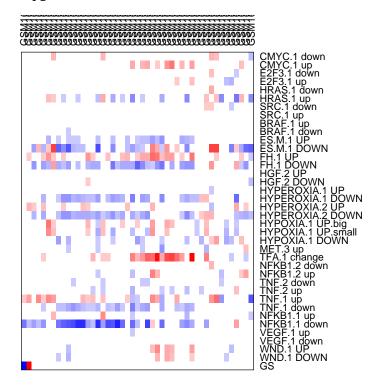
3 Visualize Results

Next we will want to create an attractive plot so that we can visually interpret our results. Looking at the range of the results helps us pick a reasonable scale for the color gradient.

> range(pg, finite = TRUE)

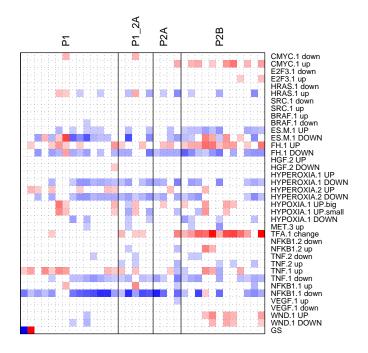
[1] -12.64043 17.89061

> smcPlot(pg, col = .rwb, scale = c(-15, 15))



Next with the addition of a few different arguments we can make this plot look much nicer.

```
> smcPlot(pg, factor(subtype), col = .rwb, scale = c(-15, 15),
+ margins = c(1, 1, 6, 9), show.grid = TRUE, r.cex = 0.75)
```



4 Analyze Your Results

One might visually attempt to determine significant differences between our subtypes, but a computational approach is more robust. One simple way to further mine these results is to return an unfiltered matrix from PGSEA and then use a linear modeling approach. In this example we will highlight the differences between Papillary RCC Type 2b and normal renal tissue.

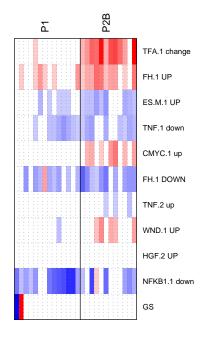
```
> pgNF <- PGSEA(eset, VAIgsc, ref = which(subtype == "NO"), p.value = NA)
> library(limma)
> design <- model.matrix(~-1 + factor(subtype))</pre>
> colnames(design) <- names(table(subtype))</pre>
> fit <- lmFit(pgNF, design)</pre>
> contrast.matrix <- makeContrasts(P2B - NO, levels = design)</pre>
> fit <- contrasts.fit(fit, contrast.matrix)</pre>
> fit <- eBayes(fit)
> topTable(fit, n = 10)[, c("logFC", "t", "adj.P.Val")]
       logFC
                      t
                            adj.P.Val
25
   9.242279 7.797634 2.125059e-08
13 5.206364 6.429828 1.181020e-06
11 -3.417913 -6.138576 2.156243e-06
```

31 -3.310984 -4.910230 1.069416e-04 3.681438 4.821095 1.151535e-04 2 14 -5.053123 -4.352265 4.463288e-04 29 -2.279435 -3.580832 4.228985e-03 36 2.695910 3.466382 5.195338e-03 1.170192 2.676237 3.749419e-02 15 33 -3.784797 -2.671271 3.749419e-02

Now that we have figured out the significantly different gene sets, we can grab the rownames from "topTable" and insert it into our plotting function.

```
> smcPlot(pg[as.numeric(rownames(topTable(fit, n = 10))), ], factor(subtype,
      levels = c("P1", "P2B")), col = .rwb, scale = c(-15, 15),
+
+
```

margins = c(1, 1, 6, 19), show.grid = TRUE, r.cex = 0.75)



Biological interpretation of these results is tricky, and can only end up as "good" as the gene sets and gene expression data used. We have some confidence in most of these signatures, and in this case, a few of the results are worth of note.

The "FH.1" gene sets were created from gene expression data generated from fumarate-hydratase (gene symbol: "FH") deficient uterine fibroids (PMID: 16319128). Using these signatures in PGSEA attempted to detect the same pattern of altered expression as was seen in the fibroids. "FH" is known to be involved with aggressive renal papillary cancers (PMIDs: 17895761, 17392716ÿ 17270241, etc.) so this result is expected. While the involvement of "FH" and renal cancer was already well-established, the association with "c-MYC" was unreported, so it was an interesting research topic for us. For more information see: "Detection of DNA copy number changes and oncogenic signaling abnormalities from gene expression data reveals MYC activation in high-grade papillary renal cell carcinoma." (Cancer Res. 2007 Apr 1;67(7):3171-6. PMID: 17409424)

5 Using Large Databases of GeneSets with PGSEA

Instead of using our included example gene sets, one may wish to go another route and use a large database such as "GO". We have included a helper function to quickly generate gene sets from this database.

```
> gos <- go2smc()
> pg <- PGSEA(eset, gos, ref = which(subtype == "NO"))
> pgNF <- PGSEA(eset, gos, ref = which(subtype == "NO"), p.value = NA)
> design <- model.matrix(~-1 + factor(subtype))
> colnames(design) <- names(table(subtype))
> fit <- lmFit(pgNF, design)
> contrast.matrix <- makeContrasts(P2B - P1, levels = design)
> fit <- contrasts.fit(fit, contrast.matrix)
> fit <- eBayes(fit)</pre>
```

The linear model approach above will highlight the differences between subtypes P2B and P1 for all go terms. Below, we plot the results and restrict the subtypes shown to those two that we are currently interested in.

```
> smcPlot(pg[as.numeric(rownames(topTable(fit, n = 30, resort.by = "logFC"))),
+ ], factor(subtype, levels = c("P1", "P2B")), col = .rwb,
+ scale = c(-15, 15), margins = c(1, 1, 6, 19), show.grid = TRUE,
+ r.cex = 0.75)
```

G0:0000776 kinetochore G0:0007093 mitotic cell cycle checkpoint G0:0007088 regulation of mitosis G0:000082 G1/S transition of mitotic cell cycle G0:000082 G1/S transition of mitotic cell cycle G0:000083 transcription from RNA polymerase III promoter G0:0007173 epidermal growth factor receptor signaling pathway G0:0001143 (5)=cycle-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004143 (5)=cycle-nucleotide phosphodiesterase activity G0:0004143 (5)=cyclic-nucleotide phosphodiesterase activity G0:0004143 (5)=cyclic-nucleotide phosphodiesterase activity G0:0004143 (5)=cyclic-nucleotide phosphodiesterase activity G0:0004143 negative regulation of transferase activity G0:00044450 microtubule organizing center part G0:0000479 regulation of kinase activity G0:0003528 steroid hormone receptor binding G0:0003213 regulation of Rab GTPase activity G0:0003213 regulation of Rab GTPase activity G0:0003213 regulation of Rab GTPase activity G0:000384 daminent G0:0003313 embryonic appendage morphogenesis G0:0003313 embryonic appendage morphogenesis G0:0003856 epithelial cell differentiation G0:0003856 pithelial cell differentiation G0:0001864 protein autoprocessing G0:0001764 neuron migration GS	Б	P2B	
G0:0007088 regulation of mitosis G0:0006270 DNA replication initiation G0:0006270 DNA replication initiation G0:0006280 G1/S transition of mitotic cell cycle G0:0006383 transcription from RNA polymerase III promoter G0:0007173 epidermal growth factor receptor signaling pathway G0:000114 3',5'-cyclic-nucleotide phosphodiesterase activity G0:0004114 2',5'-cyclic-nucleotide phosphodiesterase activity G0:0004114 2',5'-cyclic-nucleotide phosphodiesterase activity G0:00051348 negative regulation of transferase activity G0:00051348 negative regulation of protein kinase activity G0:0003673 negative regulation of protein kinase activity G0:0003528 steroid hormone receptor binding G0:00005131 regulation of Asb GTPase activity G0:00031570 DNA integrity checkpoint G0:0000584 actin filament G0:0000584 actin filament G0:0003231 regulation of Asb GTPase activity G0:0003236 embryonic limb morphogenesis G0:0030326 dynein complex G0:0030326 dynein complex G0:0030385 epithelial cell differentiation G0:0003548 phospholipid transporter activity G0:0001864 protein autoprocessing G0:0001864 protein autoprocessing G0:0001864 protein autoprocessing G0:0001764 neuron migration			
G0:0006270 DŇA replication initiation G0:000082 G1/S transition of mitotic cell cycle G0:000633 transcription from RNA polymerase III promoter G0:000114 3;5-cyclic-nucleotide phosphodiesterase activity G0:000114 3;5-cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:000414 3;5-cyclic-nucleotide phosphodiesterase activity G0:0005884 damaged DNA binding G0:000573 negative regulation of transferase activity G0:0005484 negative regulation of transferase activity G0:0005473 negative regulation of rotein kinase activity G0:0003528 steroid hormone receptor binding G0:0003770 DNA integrity checkpoint G0:0003770 DNA integrity checkpoint G0:0003137 regulation of Rab GTPase activity G0:00035170 DNA integrity checkpoint G0:0003236 embryonic limb morphogenesis G0:003513 empronic appendage morphogenesis G0:003516 dynein complex G0:000356 dynein complex G0:0003564 phospholipid transporter activity G0:0003564 phospholipid transporter activity G0:0001864 protein autoprocessing G0:0001864 protein autoprocessing G0:0001764 neuron migration			
G0:000082 G1/S transition of mitotic cell cycle G0:000083 G1/S transcription from RNA polymerase III promoter G0:0007173 epidermal growth factor receptor signaling pathway G0:001173 epidermal growth factor receptor signaling pathway G0:001112 cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0006484 damaged DNA binding G0:00051348 negative regulation of protein kinase activity G0:0003673 negative regulation of protein kinase activity G0:00032528 steroid hormone receptor binding G0:000079 regulation of cyclin-dependent protein kinase activity G0:0003213 regulation of Rab GTPase activity G0:0003264 actin filament G0:0003266 dynein complex G0:0003856 epithelial cell differentiation G0:0003856 epithelial cell differentiation G0:0019861 flagellum G0:0003548 phospholipid transporter activity G0:00019861 flagellum G0:00019640 protein autoprocessing G0:0010764 neuron migration			
G0:0006383 transcription from RNA polymerase III promoter G0:0007173 epidermal growth factor receptor signaling pathway G0:0016383 nuclear matrix G0:0004114 3',5'-cyclic-nucleotide phosphodiesterase activity G0:00051348 damaged DNA binding G0:00051348 negative regulation of transferase activity G0:00051348 negative regulation of protein kinase activity G0:00051348 negative regulation of protein kinase activity G0:0007079 regulation of cyclin-dependent protein kinase activity G0:0005285 steroid hormone receptor binding G0:0003270 regulation of cyclin-dependent protein kinase activity G0:0005285 steroid hormone receptor binding G0:00031570 DNA integrity checkpoint G0:0003261 response to UV G0:0003286 dynein complex G0:0001864 protein autoprocessing G0:00018640 protein autoprocessing			
G0:0007173 epidernal growth factor receptor signaling pathway G0:0016383 nuclear matrix G0:0001141 3;5-cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004148 negative regulation of transferase activity G0:000649 negative regulation of transferase activity G0:0005328 steroid hormone receptor binding G0:0003528 steroid hormone receptor binding G0:0003770 DNA integrity checkpoint G0:0003137 regulation of cyclin-dependent protein kinase activity G0:00032313 regulation of Rab GTPase activity G0:000326 embryonic limb morphogenesis G0:0003286 actin filament G0:0003286 actin filament G0:0003286 dates und G0:0003286 dates und G0:0003286 mergonal complex G0:0003286 mergonal			
G0:0016363 nuclear matrix G0:004114 3',5'-cyclic-nucleotide phosphodiesterase activity G0:004114 3',5'-cyclic-nucleotide phosphodiesterase activity G0:004112 cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0051348 negative regulation of transferase activity G0:00044450 microtubule organizing center part G0:000079 negative regulation of protein kinase activity G0:00036288 steroid hormone receptor binding G0:000379 regulation of cyclin-dependent protein kinase activity G0:00032313 regulation of Rab GTPase activity G0:0003286 actin filament G0:0003286 dynein complex G0:0003286 dynein complex G0:0003865 epithelial cell differentiation G0:0019861 flagellum G0:0019861 flagellum G0:0019864 phospholipid transporter activity G0:0016544 protein autoprocessing G0:0016544 protein autoprocessing G0:001677 protein antipo acid autophosphorylation G0:001764 neuron migration			
G0:0004114 3',5'-cyclic-nucleotide phosphodiesterase activity G0:0004112 cyclic-nucleotide phosphodiesterase activity G0:0004486 damaged DNA binding G0:00051348 negative regulation of transferase activity G0:00051348 negative regulation of protein kinase activity G0:00051348 negative regulation of protein kinase activity G0:0005285 steroid hormone receptor binding G0:000079 regulation of cyclin-dependent protein kinase activity G0:0003673 negative regulation of cyclin-dependent protein kinase activity G0:00031570 DNA integrity checkpoint G0:00005284 actin filament G0:00005284 actin filament G0:000326 embryonic limb morphogenesis G0:003236 embryonic limb morphogenesis G0:003286 dynein complex G0:003286 dynein complex G0:0001864 protein autoprocessing G0:0001864 protein autoprocessing G0:00016540 protein autophosphorylation G0:0001764 neuron migration	2010/00/00		
GC:0004112 cyclic=nucleotide phosphodiesterase activity GC:0003684 damaged DNA binding GC:0003684 damaged DNA binding GC:00051348 negative regulation of transferase activity GC:0006469 negative regulation of rotein kinase activity GC:0003528 steroid hormone receptor binding GC:0003528 steroid hormone receptor binding GC:00031570 DNA integrity checkpoint GC:00032313 regulation of Rab GTPase activity GC:00032313 regulation of Rab GTPase activity GC:0003236 embryonic limb morphogenesis GC:0003286 datch liament GC:0003286 datch liament GC:0003286 datch liament GC:0003286 datch liament GC:0003286 dynein complex GC:0003286 dynein complex GC			
GO:0003684 damaged DNA binding GO:0003684 damaged DNA binding GO:00051348 negative regulation of transferase activity GO:0006469 negative regulation of protein kinase activity GO:0003673 negative regulation of kinase activity GO:00035258 steroid hormone receptor binding GO:0000079 regulation of cyclin-dependent protein kinase activity GO:00035270 DNA integrity checkpoint GO:00032313 regulation of Rab GTPase activity GO:00032313 regulation of Rab GTPases activity GO:00032313 regulation of protein kinase activity GO:00032313 regulation of Rab GTPase activity GO:00032366 embryonic limb morphogenesis GO:0030286 dynein complex GO:0030286 flagellum GO:0005848 flagellum GO:0005548 phospholipid transporter activity GO:0016540 protein autoprocessing GO:0016540 protein autophosphorylation GO:0016764 neuron migration			
G0:0051348 negative regulation of transferase activity G0:0051348 negative regulation of transferase activity G0:0051348 negative regulation of protein kinase activity G0:0005285 steroid hormone receptor binding G0:00051341 regound of cyclin-dependent protein kinase activity G0:00051341 regound of cyclin-dependent protein kinase activity G0:00051341 regound of cyclin-dependent protein kinase activity G0:0005141 response to UV G0:0005284 actin filament G0:0005286 dynein complex G0:0005286 dynein complex G0:0005286 filagellum G0:0005884 actin filament G0:000326 embryonic limb morphogenesis G0:0003286 dynein complex G0:0003286 filagellum G0:0005548 phospholipid transporter activity G0:0005548 phospholipid transporter activity G0:0005549 protein autoprocessing G0:0016540 protein autophosphorylation G0:0016740 neuron migration			
G0:0044450 microtubule organizing center part G0:0006469 negative regulation of protein kinase activity G0:0035258 steroid hormone receptor binding G0:000379 regulation of cyclin-dependent protein kinase activity G0:003313 regulation of cyclin-dependent protein kinase activity G0:003313 regulation of Rab GTPase activity G0:0032313 regulation of Rab GTPase activity G0:0009411 response to UV G0:000326 embryonic limb morphogenesis G0:003526 dynein complex G0:003556 dynein complex G0:00305548 phospholipid transporter activity G0:0019861 flagellum G0:0015644 protein autoprocessing G0:001764 neuron migration			
G0:0006469 negative regulation of protein kinase activity G0:0006469 negative regulation of kinase activity G0:0003673 negative regulation of kinase activity G0:0003528 steroid hormone receptor binding G0:000079 regulation of cyclin-dependent protein kinase activity G0:00031570 DNA integrity checkpoint G0:00032313 regulation of Rab GTPase activity G0:000584 actin filament G0:0003026 embryonic appendage morphogenesis G0:0030286 dynein complex G0:0030286 the flagellum G0:0005848 flagellum G0:0005848 plospholipid transporter activity G0:000548 phospholipid transporter activity G0:0003026 dynein autoprocessing G0:0016540 protein autophosphorylation G0:0016740 neuron migration			
G0:0033673 negative regulation of kinase activity G0:0033263 steroid hormone receptor binding G0:0035258 steroid hormone receptor binding G0:00031570 DNA integrity checkpoint G0:00032313 regulation of Asb GTPase activity G0:00032313 regulation of Asb GTPase activity G0:00032313 regulation of Asb GTPase activity G0:0003211 response to UV G0:0003266 embryonic limb morphogenesis G0:0003213 embryonic limb morphogenesis G0:0003286 dynein complex G0:0003286 flagellum G0:0018548 phospholipid transporter activity G0:0018548 phospholipid transporter activity G0:0018548 phospholipid transporter activity G0:0018548 protein autoprocessing G0:0016549 protein autophosphorylation G0:001764 neuron migration			
G0:0035258 steroid hormone receptor binding G0:0035258 steroid hormone receptor binding G0:000079 regulation of cyclin-dependent protein kinase activity G0:0032313 regulation of Rab GTPase G0:003236 embryonic limb morphogenesis G0:003526 dynein complex G0:0030326 dynein complex G0:0030356 epithelial cell differentiation G0:0030854 apholipid transporter activity G0:0018548 phospholipid transporter activity G0:0018548 protein autoprocessing G0:001764 neuron migration			
G0:000079 regulation of cyclin-dependent protein kinase activity G0:00031570 DNA integrity checkpoint G0:0003131 regulation of Rab GTPase activity G0:0005841 actin filament G0:000584 actin filament G0:0003026 embryonic appendage morphogenesis G0:0030266 dynein complex G0:0030266 dynein complex G0:0030266 the filagellum G0:0005548 phospholipid transporter activity G0:0005548 phospholipid transporter activity G0:0016540 protein autoprocessing G0:0016540 protein autophosphorylation G0:001764 neuron migration			
G0:0031570 DŇA integrity checkpoint G0:0032313 regulation of Rab GTPase activity G0:0003211 response to UV G0:0005884 actin filament G0:0030326 embryonic limb morphogenesis G0:003131 ambryonic appendage morphogenesis G0:0030286 dynein complex G0:0030286 filagellum G0:0018548 phospholipid transporter activity G0:0016540 protein autoprocessing G0:0016540 protein autoprocessing G0:0016740 neuron migration			
G0:0032313 regulation of Rab GTPase activity G0:0009411 response to UV G0:0009484 actin filament G0:0005848 actin filament G0:003326 embryonic limb morphogenesis G0:003326 dynein complex G0:0030326 dynein complex G0:0030855 epithelial cell differentiation G0:0019861 flagellum G0:0019861 flagellum G0:0016544 protein autoprocessing G0:0016544 protein autoprocessing G0:001674 neuron migration		20000000000000	
GO:0009411 response to UV GO:0005884 actin filament GO:0005884 actin filament GO:0030266 embryonic limb morphogenesis GO:0030286 dynein complex GO:0030286 dynein complex GO:00105548 phospholipid transporter activity GO:0016540 protein autoprocessing GO:0046777 protein autoprocessing GO:00407764 neuron migration			
GO:0005884 actin filament GO:00030326 embryonic limb morphogenesis GO:0033113 embryonic appendage morphogenesis GO:0035113 embryonic appendage morphogenesis GO:0030286 dynein complex GO:0030286 flagellum GO:00105648 phospholipid transporter activity GO:0016540 protein autoprocessing GO:0016540 protein autoprocessing GO:001677 protein amino acid autophosphorylation GO:0001764 neuron migration			
GO:0030326 embryonic limb morphogenesis GO:0035113 embryonic appendage morphogenesis GO:0030286 dynein complex GO:0030855 epithelial cell differentiation GO:0019861 flagellum GO:0016540 protein autoprocessing GO:0016540 protein autoprocessing GO:001777 protein amtopolyphorylation GO:0001764 neuron migration			
G0:0035113 embryonic appendage morphogenesis G0:0030286 dynein complex G0:0030285 epithelial cell differentiation G0:0019861 flagellum G0:0016540 protein autoprocessing G0:0016540 protein autoprocessing G0:001677 protein amino acid autophosphorylation G0:001764 neuron migration			
GO:0030286 dynein complex GO:0030286 fynein complex GO:0030855 epithelial cell differentiation GO:0019861 flagellum GO:0016540 protein autoprocessing GO:0016540 protein autoprocessing GO:0046777 protein autoprocessing GO:00407764 neuron migration			
GO:0030855 epithelial cell differentiation GO:0019861 flagellum GO:0005548 phospholipid transporter activity GO:0016540 protein autoprocessing GO:0046777 protein amino acid autophosphorylation GO:001764 neuron migration			
GO:0019861 flagellum GO:0005548 phospholipid transporter activity GO:0016540 protein autoprocessing GO:0046777 protein amino acid autophosphorylation GO:001764 neuron migration			
GO:0005548 phospholipid transporter activity GO:0016540 protein autoprocessing GO:0046777 protein amino acid autophosphorylation GO:0001764 neuron migration			
GO:0016540 protein autoprocessing GO:0046777 protein amino acid autophosphorylation GO:0001764 neuron migration			
GO:0046777 protein amino acid autophosphorylation GO:0001764 neuron migration			
GO:0001764 neuron migration			
GS			
			GS