

sigPathway: Pathway Analysis with Microarray Data

Weil Lai¹, Lu Tian², and Peter Park^{1,3}

April 22, 2010

1. Harvard-Partners Center for Genetics and Genomics, 77 Avenue Louis Pasteur, Boston, MA 02115
2. Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, 680 North Lake Shore Drive, Chicago, IL 60611
3. Children's Hospital Informatics Program, 300 Longwood Avenue, Boston, MA 02115

Contents

| | | |
|----------|---------------------|----------|
| 1 | Introduction | 1 |
| 2 | Data | 1 |
| 3 | Example | 2 |
| 4 | Notes | 6 |

1 Introduction

sigPathway is an R package that performs pathway (gene set) analysis on microarray data. It calculates two gene set statistics, the NT_k (Q1) and NE_k (Q2), by permutation, ranks the pathways based on the magnitudes of the two statistical tests, and estimates q-values for each pathway (Tian et al., 2005). The program permutes the rows and columns of the expression matrix for NT_k and NE_k , respectively. In this vignette, we demonstrate how the user can use this package to identify statistically significant pathways in their data and export the results to HTML for browsing.

2 Data

In Tian et al. (2005), microarray data from patients with diabetes, inflammatory myopathies, and Alzheimers' data sets were analyzed. To save disk space, a small portion of the inflammatory myopathies data set has been included with *sigPathway* as an example data set. Expression values and annotations for this data set are stored in the `MuscleExample` workspace. This workspace contains the following R objects:

tab a filtered numeric matrix containing expression values from 7/13 normal (NORM) and 8/23 inclusion body myositis (IBM) samples. The row and column names of the matrix correspond to Affymetrix probe set IDs and sample IDs, respectively. The 5000 probe sets in this matrix represent the most variable probe sets (by expression value) in the 15 arrays.

phenotype a character vector with 0_NORM to represent NORM and 1_IBM to represent IBM

G a pathway annotation list containing the pathway's source, title, and associated probe set IDs

To load this data set, type 'data(MuscleExample)' after loading the *sigPathway* package.

The pathways annotated in **G** were curated from Gene Ontology, KEGG, BioCarta, BioCyc, and SuperArray. Each element *within* **G** is a list describing a pathway with the following sub-elements:

src a character vector containing either the pathway ID (for Gene Ontology) or the name of the pathway database

title a character vector containing the pathway name

probes a character vector containing probe set IDs that are associated with the pathway (by mapping them to Entrez Gene IDs)

The full inflammatory myopathway data set and pathway annotations for other, selected Affymetrix microarray platforms are available at <http://www.chip.org/~ppark/Supplements/PNAS05.html>. For example, the more comprehensive pathway annotation list for the Affymetrix HG-U133A platform is called *GenesetsU133a*. For arrays not listed on the website (or for scenarios such as linkage analysis), the user can make his/her own pathway annotations and use them in *sigPathway* as long as the pathway annotations are arranged in the above format.

3 Example

In this section, we show the R code necessary to conduct pathway analysis with *sigPathway* on an example data set.

First, we load *sigPathway* and the example data set into memory. If we are dealing with the full data set, we could remove probe sets that have expression values less than the trimmed mean in all of the arrays. We assume that the probe sets with lower expression values across all arrays are not of interest. The trimmed mean was used as the filtering criterion in Tian et al. (2005). The probe sets in the example data set were selected for their variance across 15 arrays (not shown).

```
> library(sigPathway)
> data(MuscleExample)
> ls()
```

```
[1] "G"          "phenotype" "tab"
```

For microarray data, the convention is to use rows and columns to represent probe sets and individual arrays, respectively. To tell the program which column in **tab** belongs to which phenotype, we have created a character vector with 0_NORM to represent NORM and 1_IBM to represent IBM. Because 0_NORM comes before 1_IBM in alphanumeric order, the program internally treats NORM as 0 and IBM as 1. Alternatively, we could have simply used the numerals 0 and 1 to represent NORM and IBM. Note that the row names for **tab** are probe set IDs.

```
> dim(tab)
```

```
[1] 5000  15
```

```
> print(tab[501:504, 1:3])
```

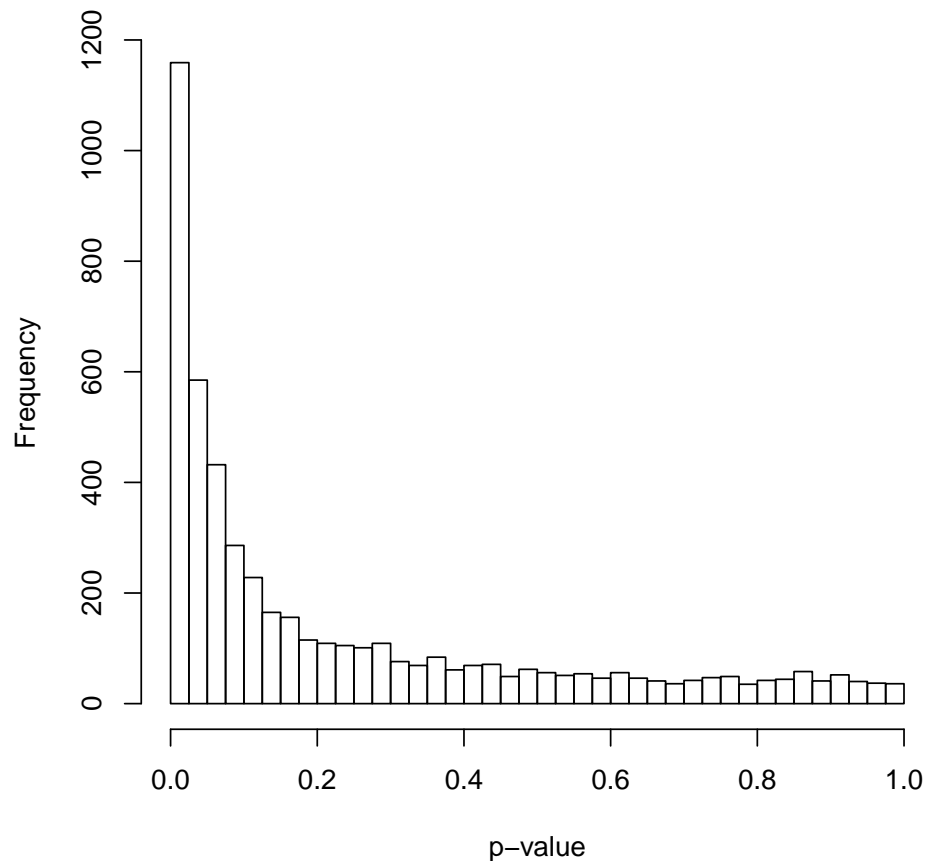
| | GEIM1.IBM.S | GEIM7.IBM.S | GEIM20.IBM.S |
|-------------|-------------|-------------|--------------|
| 217466_x_at | 3203 | 4085 | 23736 |
| 211939_x_at | 28250 | 32293 | 36890 |
| 203932_at | 6452 | 3596 | 13392 |
| 200715_x_at | 20792 | 12647 | 18865 |

```
> table(phenotype)
```

| phenotype | |
|-----------|---|
| 0_NORM | 7 |
| 1_IBM | 8 |

How much do IBM and NORM samples differ? Let us plot the unadjusted p-values for each probe set from the 2 group (sample) t-test, assuming unequal variances and using the Welch approximation to estimate the appropriate degrees of freedom.

```
> statList <- calcTStatFast(tab, phenotype, ngroups = 2)
> hist(statList$pval, breaks = seq(0, 1, 0.025), xlab = "p-value",
+      ylab = "Frequency", main = "")
```



The two different types of samples are certainly very different by the probe set level, but what pathways are driving the differences? With our pathway annotations, we calculate the NT_k and NE_k statistics for each gene set, and rank the top pathways based on the magnitude of the two statistics. The result is stored in a list (`res.muscle`), of which we will later use to write results to HTML.

```
> set.seed(1234)
> res.muscle <- runSigPathway(G, 20, 500, tab, phenotype, nsim = 1000,
+   weightType = "constant", ngroups = 2, npath = 25, verbose = FALSE,
+   allpathways = FALSE, annotpkg = "hgu133a.db", alwaysUseRandomPerm = FALSE)
```

Selecting the gene sets

Calculating NT_k statistics for each selected gene set

Calculating NE_k statistics for each selected gene set

Summarizing the top 25 pathways from each statistic

Done! Use the `writeSigPathway()` function to write results to HTML

The `set.seed` function is used here only for the purpose of getting the exact results when regenerating this vignette from its source files.

Because there can be many thousands of pathways represented in the pathway annotations, we have chosen to analyze pathways that contain at least 20 probe sets as represented in `tab`. We also exclude pathways represented by more than 500 probe sets because larger pathways tend to be non-specific. These two values were the ones used in Tian et al. (2005). To save space, our pathway annotation list has already been filtered with the above criteria. So, all of the 626 pathways in `G` will be considered in the calculations.

The run time of the NT_k and NE_k is approximately linearly proportional to `nsim`, or the maximum number of permutations. When `alwaysUseRandomPerm` is set to `FALSE` (the default value), the program will use a smaller `nsim` for the NE_k calculations and switch to using complete permutation if the total number of unique permutations for the phenotype is less than `nsim`.

We are setting `weightType` to 'constant' because of the additional time required to calculate variable weights for NE_k . If the histogram of unadjusted p-values (of the probe sets) is nearly horizontal, and we later observe high q-values (i.e., approaching 1) for the top ranked pathways, then setting `weightType` to 'variable' would help lower some of the NE_k q-values.

To rank the pathways, the program adds up the ranks corresponding to the magnitudes of NT_k and NE_k . When `npath` is set to 25 and `allpathways` to `FALSE`, the program considers the top 25 pathways for each gene set statistic before summing the individual ranks. If `allpathways` is set to `TRUE`, then all pathways are ranked for each gene set statistic before summing the individual ranks. Here, `allpathways` is set to `FALSE` because we are interested in observing pathways that are consistently highly ranked for each gene set statistic.

Also, please note that out of the numerous input parameters to `runSigPathway`, `annotpkg` is optional because it refers to a Bioconductor metadata package that may not already be present on your installation of R. In our example, 'hgu133a.db' refers to the BioConductor metadata package of the Affymetrix HG-U133A platform. By specifying 'hgu133a.db' for `annotpkg`, `runSigPathway` will include the accession number, Entrez Gene ID, gene symbol, and gene name of probe sets associated with each pathway in the list of top pathways.

Printed below is a table of the top 10 pathways, the set size, the NT_k and NE_k statistics, and the statistics' ranks and q-values. This table is accessible through the following command:

```
> print(res.muscle$df.pathways[1:10, ])
```

| | Index | Gene Set | Category |
|----|-------|----------|------------|
| 1 | 234 | | GO:0019883 |
| 2 | 292 | | GO:0042611 |
| 3 | 293 | | GO:0042612 |
| 4 | 233 | | GO:0019882 |
| 5 | 84 | | GO:0030333 |
| 6 | 237 | | GO:0019885 |
| 7 | 117 | | GO:0030106 |
| 8 | 92 | | GO:0001772 |
| 9 | 613 | | humanpaths |
| 10 | 601 | | humanpaths |

| | | Pathway | Set | Size | Percent Up | |
|----|--|--|----------|----------|-------------|----------|
| 1 | | antigen presentation, endogenous antigen | | 22 | 0.00 | |
| 2 | | MHC protein complex | | 20 | 0.00 | |
| 3 | | MHC class I protein complex | | 20 | 0.00 | |
| 4 | | antigen presentation | | 45 | 0.00 | |
| 5 | | antigen processing | | 44 | 0.00 | |
| 6 | antigen processing, endogenous antigen via MHC class I | | | 23 | 0.00 | |
| 7 | | MHC class I receptor activity | | 22 | 0.00 | |
| 8 | | immunological synapse | | 26 | 0.00 | |
| 9 | | Interferon a,b Response | | 71 | 12.68 | |
| 10 | Dendritic / Antigen Presenting Cell | | | 105 | 5.71 | |
| | NTk Stat | NTk q-value | NTk Rank | NEk Stat | NEk q-value | NEk Rank |
| 1 | 18.97 | 0 | 3 | 9.33 | 0 | 2 |
| 2 | 17.83 | 0 | 6 | 9.36 | 0 | 1 |
| 3 | 17.83 | 0 | 6 | 9.36 | 0 | 1 |
| 4 | 19.41 | 0 | 1 | 7.24 | 0 | 7 |
| 5 | 19.03 | 0 | 2 | 7.26 | 0 | 6 |
| 6 | 18.44 | 0 | 4 | 9.11 | 0 | 4 |
| 7 | 18.37 | 0 | 5 | 9.28 | 0 | 3 |
| 8 | 16.95 | 0 | 7 | 8.27 | 0 | 5 |
| 9 | 10.79 | 0 | 8 | 4.83 | 0 | 9 |
| 10 | 10.66 | 0 | 9 | 3.62 | 0 | 11 |

The positive signs on the gene set statistics indicate that the corresponding pathways are more highly expressed in IBM compared to NORM. Had we defined 1 for NORM and 0 for IBM, the interpretation would remain the same, but we would expect the signs for the gene set statistics to be flipped.

Detailed information about each probe set in each pathway on the list of top pathways are stored in the `list.gPS`, an element within `res.muscle`. `list.gPS` is a list containing data frames describing the probe sets for each top pathway. For example, let us view the annotations and test statistics for 10 probe sets in the *MHC class I receptor activity* pathway.

```
> print(res.muscle$list.gPS[[7]][1:10, ])
```

| | Probes | AccNum | GeneID | Symbol |
|-------------|-------------|-----------|--------|--------|
| 201891_s_at | 201891_s_at | NM_004048 | 567 | B2M |
| 216231_s_at | 216231_s_at | AW188940 | 567 | B2M |

| | | | | |
|-------------|-------------|-----------|------|-------|
| 218831_s_at | 218831_s_at | NM_004107 | 2217 | FCGRT |
| 213932_x_at | 213932_x_at | AI923492 | 3105 | HLA-A |
| 215313_x_at | 215313_x_at | AA573862 | 3105 | HLA-A |
| 208729_x_at | 208729_x_at | D83043 | 3106 | HLA-B |
| 209140_x_at | 209140_x_at | L42024 | 3106 | HLA-B |
| 211911_x_at | 211911_x_at | L07950 | 3106 | HLA-B |
| 208812_x_at | 208812_x_at | BC004489 | 3107 | HLA-C |
| 211799_x_at | 211799_x_at | U62824 | 3107 | HLA-C |

| | Name | Mean_0_NORM |
|-------------|--|-------------|
| 201891_s_at | beta-2-microglobulin | 38735.143 |
| 216231_s_at | beta-2-microglobulin | 43285.857 |
| 218831_s_at | Fc fragment of IgG, receptor, transporter, alpha | 1592.000 |
| 213932_x_at | major histocompatibility complex, class I, A | 23739.857 |
| 215313_x_at | major histocompatibility complex, class I, A | 20685.286 |
| 208729_x_at | major histocompatibility complex, class I, B | 6648.571 |
| 209140_x_at | major histocompatibility complex, class I, B | 12258.857 |
| 211911_x_at | major histocompatibility complex, class I, B | 9150.286 |
| 208812_x_at | major histocompatibility complex, class I, C | 13994.429 |
| 211799_x_at | major histocompatibility complex, class I, C | 2167.571 |

| | Mean_1_IBM | StDev_0_NORM | StDev_1_IBM | T-Statistic | p-value |
|-------------|------------|--------------|-------------|-------------|--------------|
| 201891_s_at | 64165.88 | 5551.0402 | 7325.835 | 7.629433 | 4.155125e-06 |
| 216231_s_at | 78550.75 | 4350.8622 | 9728.212 | 9.250160 | 3.329180e-06 |
| 218831_s_at | 4444.00 | 351.2762 | 2263.444 | 3.515833 | 8.973576e-03 |
| 213932_x_at | 65602.12 | 6463.6639 | 8931.066 | 10.485603 | 1.363431e-07 |
| 215313_x_at | 68365.12 | 5568.5959 | 9316.500 | 12.197747 | 5.615578e-08 |
| 208729_x_at | 46637.62 | 609.5618 | 11082.266 | 10.188448 | 1.804436e-05 |
| 209140_x_at | 65679.25 | 1433.5514 | 5988.982 | 24.441414 | 9.843999e-09 |
| 211911_x_at | 53755.88 | 2499.9600 | 14031.821 | 8.832473 | 3.162204e-05 |
| 208812_x_at | 62945.38 | 1825.2766 | 8910.416 | 15.178761 | 5.386934e-07 |
| 211799_x_at | 23667.38 | 451.3210 | 8454.758 | 7.180791 | 1.749504e-04 |

A much more intuitive method to browse through the results is to write the results to HTML, which can then be read by an Internet browser program (e.g., Mozilla Firefox, Microsoft Internet Explorer). Writing the results can be achieved with the `writeSigPathway` function. Please refer to the help file of `writeSigPathway` for more details on how to save the results to a specific directory. Figures 1 and 2 show examples of the HTML output after running `writeSigPathway` and opening the corresponding HTML file in an Internet browser.

4 Notes

This vignette was compiled with the following settings:

```
> print(sessionInfo())
```

```
R version 2.11.0 (2010-04-22)
i386-pc-mingw32
```

```
locale:
```

sigPathway_results/TopPathwaysTable.html

List of Top Pathways

| | IndexG | Gene Set Category | Pathway | Set Size | Percent Up | NTk Stat | NTk q-value | NTk Rank | NEk Stat | NEk q-value | NEk Rank |
|----|--------|-------------------|--|----------|------------|----------|-------------|----------|----------|-------------|----------|
| 1 | 234 | GO:0019883 | antigen presentation, endogenous antigen | 22 | 100 | 18.97 | 0.0000 | 3.0 | 9.33 | 0.0000 | 2.0 |
| 2 | 292 | GO:0042611 | MHC protein complex | 20 | 100 | 17.83 | 0.0000 | 6.0 | 9.36 | 0.0000 | 1.0 |
| 3 | 293 | GO:0042612 | MHC class I protein complex | 20 | 100 | 17.83 | 0.0000 | 6.0 | 9.36 | 0.0000 | 1.0 |
| 4 | 233 | GO:0019882 | antigen presentation | 45 | 100 | 19.41 | 0.0000 | 1.0 | 7.24 | 0.0000 | 7.0 |
| 5 | 84 | GO:0030333 | antigen processing | 44 | 100 | 19.03 | 0.0000 | 2.0 | 7.26 | 0.0000 | 6.0 |
| 6 | 237 | GO:0019885 | antigen processing, endogenous antigen via MHC class I | 23 | 100 | 18.44 | 0.0000 | 4.0 | 9.11 | 0.0000 | 4.0 |
| 7 | 117 | GO:0030106 | MHC class I receptor activity | 22 | 100 | 18.37 | 0.0000 | 5.0 | 9.28 | 0.0000 | 3.0 |
| 8 | 92 | GO:0001772 | immunological synapse | 26 | 100 | 16.95 | 0.0000 | 7.0 | 8.27 | 0.0000 | 5.0 |
| 9 | 613 | humanpaths | Interferon a,b Response | 71 | 87 | 10.79 | 0.0000 | 8.0 | 4.83 | 0.0000 | 9.0 |
| 10 | 601 | humanpaths | Dendritic / Antigen Presenting Cell | 105 | 94 | 10.66 | 0.0000 | 9.0 | 3.62 | 0.0000 | 11.0 |
| 11 | 19 | GO:0045012 | MHC class II receptor activity | 21 | 100 | 8.45 | 0.0000 | 21.0 | 4.91 | 0.0000 | 8.0 |
| 12 | 236 | GO:0019884 | antigen presentation, exogenous antigen | 21 | 100 | 8.45 | 0.0000 | 21.0 | 4.91 | 0.0000 | 8.0 |
| 13 | 238 | GO:0019886 | antigen processing, exogenous antigen via MHC class II | 21 | 100 | 8.45 | 0.0000 | 21.0 | 4.91 | 0.0000 | 8.0 |
| 14 | 481 | GO:0009615 | response to virus | 31 | 87 | 5.10 | 0.0000 | 75.0 | 3.84 | 0.0000 | 10.0 |
| 15 | 576 | KEGG | Jak-STAT signaling pathway | 38 | 87 | 4.90 | 0.0000 | 84.0 | 3.29 | 0.0000 | 18.0 |
| 16 | 40 | GO:0006968 | cellular defense response | 35 | 89 | 4.73 | 0.0000 | 93.0 | 3.54 | 0.0000 | 13.0 |
| 17 | 42 | GO:0006959 | humoral immune response | 46 | 91 | 4.81 | 0.0000 | 86.0 | 3.19 | 0.0000 | 21.0 |
| 18 | 612 | humanpaths | Th1-Th2-Th3 | 34 | 88 | 4.21 | 0.0000 | 114.0 | 3.31 | 0.0000 | 17.0 |
| 19 | 575 | KEGG | Toll-like receptor signaling pathway | 40 | 85 | 4.14 | 0.0000 | 117.0 | 3.27 | 0.0000 | 19.0 |
| 20 | 625 | humanpaths | Asthma | 20 | 100 | 3.69 | 0.0000 | 137.0 | 3.19 | 0.0000 | 22.0 |
| 21 | 470 | GO:0043085 | positive regulation of enzyme activity | 29 | 86 | 3.45 | 0.0000 | 147.0 | 3.36 | 0.0000 | 15.0 |
| 22 | 89 | GO:0045333 | cellular respiration | 40 | 18 | -7.82 | 0.0000 | 22.0 | -2.01 | 0.0285 | 174.0 |
| 23 | 526 | BioCarta | p38 MAPK Signaling Pathway | 24 | 92 | 2.88 | 0.0042 | 191.0 | 3.33 | 0.0000 | 16.0 |
| 24 | 18 | GO:0005884 | actin filament | 26 | 73 | 2.51 | 0.0107 | 222.5 | 3.56 | 0.0000 | 12.0 |
| 25 | 529 | BioCarta | Activation of Csk by cAMP-dependent Protein Kinase | 27 | 74 | 2.37 | 0.0154 | 236.0 | 3.26 | 0.0000 | 20.0 |

Figure 1: List of Top Pathways in Inclusion Body Myositis versus Normal

sigPathway_results/pathways/pathway_117.html

[Back to Table of Top Pathways](#)

MHC class I receptor activity

| | Probes | AccNum | GeneID | Symbol | Name | Mean_0_NORM | Mean_1_IBM | StDev_0_NORM | StDev_1_IBM | T-Statistic | p-value |
|----|-------------|-----------|--------|---------|--|-------------|------------|--------------|-------------|-------------|---------|
| 1 | 201891_s_at | NM_004048 | 567 | B2M | beta-2-microglobulin | 38735.1 | 64165.9 | 5551.0 | 7325.8 | 7.629 | 0.0000 |
| 2 | 216231_s_at | AW188940 | 567 | B2M | beta-2-microglobulin | 43285.9 | 78550.8 | 4350.9 | 9728.2 | 9.250 | 0.0000 |
| 3 | 218831_s_at | NM_004107 | 2217 | FCGRT | Fc fragment of IgG, receptor, transporter, alpha | 1592.0 | 4444.0 | 351.3 | 2263.4 | 3.516 | 0.0090 |
| 4 | 213932_x_at | AI923492 | 80862 | C6orf12 | chromosome 6 open reading frame 12 | 23739.9 | 65602.1 | 6463.7 | 8931.1 | 10.486 | 0.0000 |
| 5 | 215313_x_at | AA573862 | 3105 | HLA-A | major histocompatibility complex, class I, A | 20685.3 | 68365.1 | 5568.6 | 9316.5 | 12.198 | 0.0000 |
| 6 | 208729_x_at | D83043 | 3106 | HLA-B | major histocompatibility complex, class I, B | 6648.6 | 46637.6 | 609.6 | 11082.3 | 10.188 | 0.0000 |
| 7 | 209140_x_at | L42024 | 3106 | HLA-B | major histocompatibility complex, class I, B | 12258.9 | 65679.3 | 1433.6 | 5989.0 | 24.441 | 0.0000 |
| 8 | 211911_x_at | L07950 | 3106 | HLA-B | major histocompatibility complex, class I, B | 9150.3 | 53755.9 | 2500.0 | 14031.8 | 8.832 | 0.0000 |
| 9 | 208812_x_at | BC004489 | 3107 | HLA-C | major histocompatibility complex, class I, C | 13994.4 | 62945.4 | 1825.3 | 8910.4 | 15.179 | 0.0000 |
| 10 | 211799_x_at | U62824 | 3107 | HLA-C | major histocompatibility complex, class I, C | 2167.6 | 23667.4 | 451.3 | 8454.8 | 7.181 | 0.0002 |
| 11 | 214459_x_at | M12679 | 3107 | HLA-C | major histocompatibility complex, class I, C | 10482.7 | 60946.4 | 2644.2 | 12988.9 | 10.738 | 0.0000 |
| 12 | 216526_x_at | AK024836 | 3107 | HLA-C | major histocompatibility complex, class I, C | 17840.3 | 76157.5 | 4994.1 | 8878.6 | 15.921 | 0.0000 |
| 13 | 200904_at | X56841 | 3133 | HLA-E | major histocompatibility complex, class I, E | 2283.6 | 12515.5 | 314.1 | 4947.8 | 5.836 | 0.0006 |
| 14 | 200905_x_at | NM_005516 | 3133 | HLA-E | major histocompatibility complex, class I, E | 4583.7 | 24874.1 | 721.7 | 7933.4 | 7.200 | 0.0002 |
| 15 | 217456_x_at | M31183 | 3133 | HLA-E | major histocompatibility complex, class I, E | 2692.0 | 9809.9 | 492.5 | 2175.7 | 8.994 | 0.0000 |
| 16 | 204806_x_at | NM_018950 | 3134 | HLA-F | major histocompatibility complex, class I, F | 4062.1 | 29127.1 | 829.9 | 10796.3 | 6.545 | 0.0003 |
| 17 | 221875_x_at | AW514210 | 3134 | HLA-F | major histocompatibility complex, class I, F | 5604.4 | 36141.0 | 937.4 | 9840.3 | 8.732 | 0.0000 |

Figure 2: MHC class I receptor activity


```
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252
```

attached base packages:

```
[1] tools      stats      graphics  grDevices  utils      datasets  methods
[8] base
```

other attached packages:

```
[1] hgu133a.db_2.4.1      org.Hs.eg.db_2.4.1    RSQLite_0.8-4
[4] DBI_0.2-5              AnnotationDbi_1.10.0  Biobase_2.8.0
[7] sigPathway_1.16.0
```

References

Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park.
Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, 102(38):13544–13549, Sep 2005.