How to use the `pathVar` package

April 28, 2020

1 Introduction

This package studies the variability of a dataset related to different pathways. For each pathway, how does the variability change compared to the whole set of genes from our dataset? Do we have an unusually high number of low variability genes in a particular pathway? These are some of the questions our package will answer. A summary of the pipeline of the package may be found in Figure 1.

The eight main functions are:

- `diagnosticsVarPlot` gives you 3 plots one for the standard deviation (sd), one for median absolute deviation (mad) and one for coefficient of variation (cv) against the mean to help you decide which one would be the best with your dataset when you have one group of samples. It also return the correlation between each variability statistics and the mean.
- `diagnosticsVarPlotsTwoSample` gives you 3 plots one for the standard deviation (sd), one for median absolute deviation (mad) and one for coefficient of variation (cv) against the mean to help you decide which one would be the best with your dataset when you are comparing two groups of samples to each other. It also return the correlation between each variability statistics and the mean.
- `makeDBList` puts your own list of pathways and genes related to them into a list in a good format.
- `pathVarOneSample` classifies your genes into one to four clusters with respect to sd, mad, cv or mean. Then, it compares the counts of genes in each class from your dataset in one pathway with the counts of the genes in each class from the whole dataset. For that, it uses a Chi-square or an exact test. You can give your own list of pathways (using the output of `makeDBList`) or use Reactome and KEGG pathways that are already included.
- `pathVarTwoSamplesCont` It splits the samples into two groups that you define. It compares the density of the variability (sd, mad, cv) or of the mean of the genes in a pathway from group 1 with the density from group 2. For that, it uses the bootstrap Kolmogorov-smirnov test. You can give your own list of pathways (using the output of `makeDBList`) or use Reactome and KEGG pathways that are already included.
- `pathVarTwoSamplesDisc` It splits the samples into two groups that you define. It classifies your genes into three clusters with respect to sd, mad, cv or mean for each group. It compares the counts of genes in each class in a pathway from group 1 the counts of genes in each class from group 2 in the same pathway. For that, it uses a Chi-square or an exact test. You can give your own list of pathways (using the output of `makeDBList`) or use Reactome and KEGG pathways that are already included.
- `sigPway` takes the output of `pathVarOneSample` or `pathVarTwoSamples` and will tell you which pathways are significant. For the one sample case, it will also tell you which categories are significant.

- `plotPway` plots the result of `pathVarOneSample` or `pathVarTwoSamples` for a chosen pathway. In the one sample case, the figure will contain the reference counts along with the plot of the chosen pathway. In the continuous two samples case, it will plot the two densities (one for each group) of the statistics (sd, mad, cv or mean) of the chosen pathway. In the discrete two samples case, the figure will contain the counts from group 1 along with the counts of group 2 of the chosen pathway.

- `plotAllTwoSampleDistributionCounts` It splits the samples into two groups that you define. It classifies your genes into clusters with respect to sd, mad, cv or mean for each group. It compares the counts of all genes in your data set from group 1 to the counts of all genes in your data set from group 2. For that, it uses a Chi-square or an exact test.

- `saveAsPDF` saves as a pdf the plots for the one or two samples case of the significant pathways or a chosen list of pathways.

- `getGenes` take the result of `pathVarTwoSamplesCont` and returns one list of genes for group 1 and one for group 2 of a chosen pathway having their statistics (sd, mad, cv or mean) inside a chosen interval. It also returns the set of all the genes from your dataset that belong to the chosen pathway.

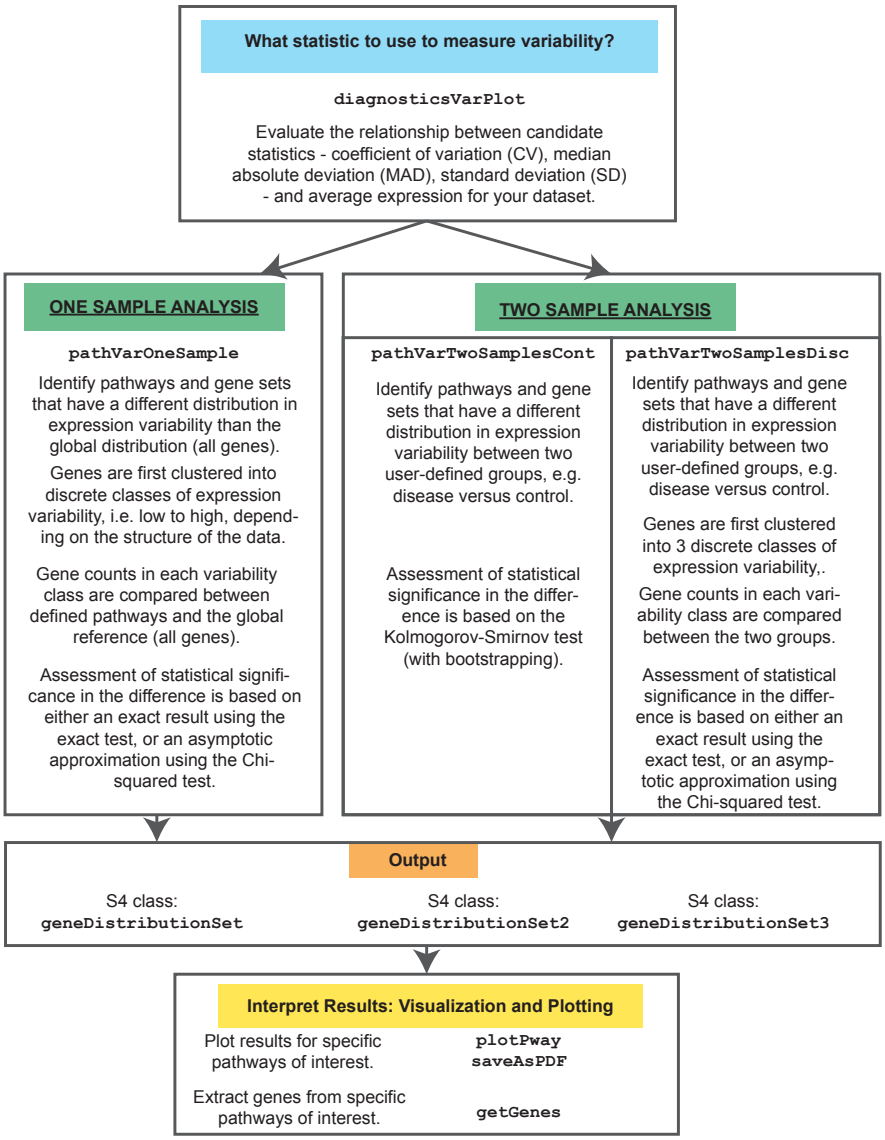

Figure 1: Outline of the pathVar analysis.

Two pathway libraries are included:

- KEGG with 272 pathways (`pways.kegg`).
- Reactome with 946 pathways (`pways.reactome`).

Each one of these two variables contain the pathname, related pathID (if any), the genes and the size of each pathway.

2 Bock dataset

The Bock et al. (2011) data set has 20 human ESC lines, and 12 iPSC lines. We use this data set to illustrate the functionality of our pathVar package that has been created to make inferences on the

functional consequences of expression variability. We downloaded the normalized microarray data sets (`http://www.medical-epigenomics.org/papers/broad_mirror/scorecard/index.html`).

In the package the dataset may be found under the name `bock`. The column names correspond to the samples with embryonic stem cells (hES) being ESC and induced pluripotent stem cells (hIPS) is iPSC.

```
> samp.id <- colnames(bock)
> cell.id <- character(length(samp.id))
> cell.id[grep("hES", samp.id)] <- "esc"
> cell.id[grep("hiPS", samp.id)] <- "ips"
```

For the one sample case we will use only the ESC and filter the genes with a low signal. We will keep the genes with at least 75% of their samples greater than 1.

```
> qc.esc <- apply(bock[,cell.id == "esc"], 1, function(x,ct){ sum(x >= ct) }, ct=1)
> bock.esc <- bock[qc.esc >= .75*sum(cell.id == "esc"),]
```

For the two samples case, we will keep the genes with at least 75% of their samples greater than 1 for ESC and for iPSC.

```
> qc.ips <- apply(bock[,cell.id == "ips"], 1, function(x,ct){ sum(x >= ct) }, ct=1)
> bock.esc_ips <- bock[qc.esc >= .75*sum(cell.id == "esc") &
+  qc.ips >= .75*sum(cell.id == "ips"),]
```

3 How to choose the variability statistics to use?

This figure helps us visualize the association between a chosen variability statistic and the mean expression. In theory, we would like to choose the variability statistics that has the smallest correlation with the mean. When looking at the ESC cells of the Bock data, we see from Figure 2 that the standard deviation has the least correlation with the mean, therefore our pathVar analysis is based on the standard deviation.

```
> diagnosticsVarPlots(bock.esc)
```

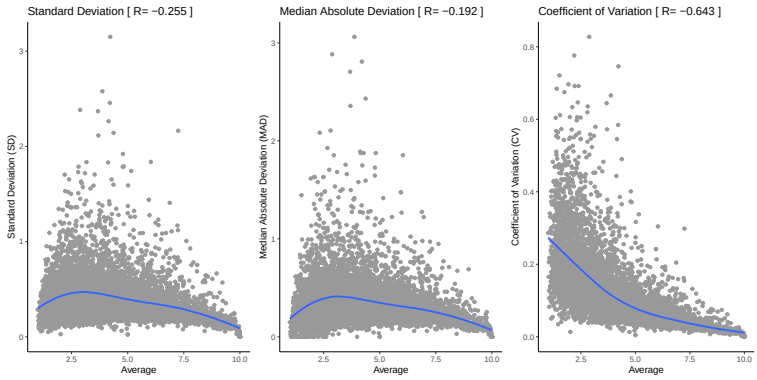

Figure 2: Relationships between SD, MAD, CV with the mean for the one sample case.

We can also create this figure for the two sample case, comparing the ESCs and iPSCs. We will only run this on the first five thousand genes of the Bock data to save time.

```
> diagnosticsVarPlotsTwoSample(bock.esc_ips[1:5000,], groups=as.factor(c(rep(1,10),rep(2,10))))
```

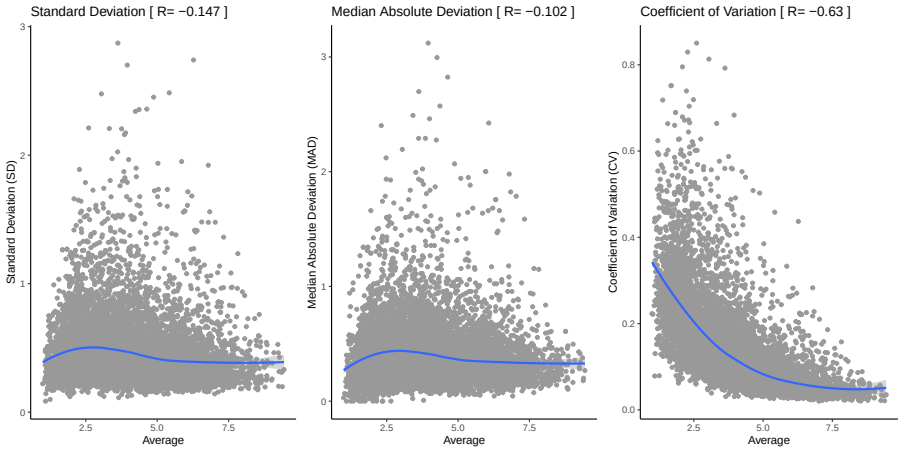

Figure 3: Relationships between SD, MAD, CV with the mean for the two sample case.

4 Finding pathways with a significant difference in expression variability for the whole dataset.

The steps of the function for the one sample case are as follow:

1. Compute a variability statistic (sd, mad, cv) for each gene. If you want to compare the result you obtain with the variability statistics to a regular analysis based on the mean, you have the option to choose mean as the statistic.

2. Classify the genes with respect to the discrete levels of high, medium and low variability (at most 4 clusters).

3. For each pathway, we extract the gene in our dataset and in which cluster they belong.

4. For each pathway, we look at how the gene counts are distributed in each category and compare it to these counts derived from the whole dataset. The two possibilities to test this difference are the Chi-squared test or the multinomial exact test.

5. We build a data table with the results step 4.

Two pathway libraries are already included in the package: `pways.reactome` and `pways.kegg`.It is possible through `makeDBList` to load any other database that is in a txt file.

We will look at the difference in standard deviation using the Kegg pathways and the Chi-squared test.

```
> resOneSam=pathVarOneSample(bock.esc,pways.kegg,test="chisq",varStat="sd")
```

The first three significant pathways from the table obtained with the Chi-squared test are:

```
> resOneSam@tablePway[1:3]
```

```
                          PwayName        PwayID        APval
1:         ECM-receptor interaction path:hsa04512 1.916368e-31
2: Protein digestion and absorption path:hsa04974 7.165506e-29
3:                  Focal adhesion path:hsa04510 4.967186e-16
   PercOfGenesInPway NumOfGenesFromDataSetInPathway PathwaySize
```

```
1:          48.27586                                42          87
2:          40.44944                                36          89
3:          60.38647                               125         207
```

For the Chi-squared test, if the pathway has less than 10 genes in our datasets, the analysis is not performed on the basis that the pathway is too small to determine if there is an enrichment in any category of expression variability. The list of these pathways is stored in a the slot named `NAPways`.

```
> resOneSam@NAPways[1:3]

[1] "Ascorbate and aldarate metabolism" "Fatty acid biosynthesis"
[3] "Primary bile acid biosynthesis"
```

For this dataset, the genes were classified in 4 clusters representing different levels of expression variability.

```
> resOneSam@numOfClus

[1] 4
```

We can also get the intersection of genes from our dataset belonging to two specified pathways "Staphylococcus aureus infection" and "Asthma":

```
> intersect(names(resOneSam@genesInPway[["Staphylococcus aureus infection"]]),
+ names(resOneSam@genesInPway[["Asthma"]]))

[1] "HLA-DRB3" "HLA-DRB5" "HLA-DRB1" "HLA-DPB1" "HLA-DMA"  "HLA-DPA1"
```

Now, we can use this result to look only at the significant pathways (with a p-value less than 0.05). As we are working with counts, the function `sigPway` will also use a binomial test to see which of the four categories were significant for each pathway.

```
> sigOneSam=sigPway(resOneSam,0.05)
```

The first pathway that is significant is the "ECM-receptor interactionn" and contains the following genes:

```
> sigOneSam@genesInSigPways1[1]

$`ECM-receptor interaction`
 COL3A1  COL6A3  COL2A1  COL5A1   THBS2    SPP1  COL5A2 COL11A1  COL1A2    CD44
      4       4       4       4       4       4       4       4       4       3
    TNC  COL6A2   LAMA2   ITGAV   ITGA5   THBS4  COL4A1  COL4A2  COL4A5  COL1A1
      3       3       3       3       3       3       3       3       3       3
  ITGA7     FN1  COL4A6   THBS1   LAMA5   ITGB5   LAMC3   LAMC1    SDC1    SDC4
      3       3       3       2       3       2       2       2       2       2
  LAMB2   LAMB1   THBS3    HMMR  COL6A1    DAG1    CD47    AGRN   HSPG2   ITGA6
      3       3       2       2       2       2       2       2       2       2
   SV2A   ITGB1
      2       2
```

As our genes were clustered into 4 categories, it would be interesting to have the list of pathways that have a difference in the 4th category (super highly variable genes).

```
> names(which(unlist(lapply(sigOneSam@sigCatPerPway,function(x) 4%in% x))))
```

```
 [1] "ECM-receptor interaction"
 [2] "Protein digestion and absorption"
 [3] "Focal adhesion"
 [4] "Amoebiasis"
 [5] "Cytokine-cytokine receptor interaction"
 [6] "Mineral absorption"
 [7] "PI3K-Akt signaling pathway"
 [8] "Platelet activation"
 [9] "Malaria"
[10] "Signaling pathways regulating pluripotency of stem cells"
[11] "Vitamin digestion and absorption"
[12] "Neuroactive ligand-receptor interaction"
[13] "Rheumatoid arthritis"
[14] "African trypanosomiasis"
[15] "Salivary secretion"
```

We could now look at the visualization of one of these pathways, for example the "ECM-receptor interaction". In Figure 3, you can see that the four variability categories were significantly different (p-value<0.05) from the reference counts.

```
> plotPway(resOneSam,"ECM-receptor interaction",sigOneSam)
```

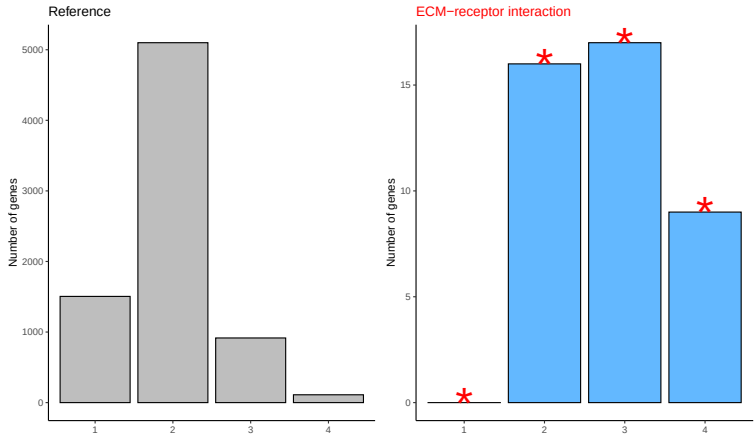

Figure 4: Histogram of the "ECM-receptor interaction" pathway and the reference counts.

5 Finding pathways with a significant difference in variability between two samples using the density.

In situations where we want to identify variability changes between two contrasting phenotypes, e.g. cancer vs normal or ESC vs IPS, the steps involved in this analysis are similar to the one sample case

The steps of the function for the two samples case using the density are as follow:

1. Compute the variability statistics (sd, mad, cv or mean) for each gene
2. For each pathway, we extract the gene in our dataset.
3. For each pathway, we evaluate how different the distribution of expression variability is between the two samples using a bootstrapped version of the Kolmogorov-Smirnov test.

4. We build a data frame with the results of step 3.

For our example, we will look at the difference in standard deviation using the Reactome pathways and the two groups: ESC and IPS.

```
> grp=c(rep(1, sum(cell.id == "esc")), rep(2, sum(cell.id == "ips")))
> set.seed(1)
> resTwoSam=pathVarTwoSamplesCont(bock.esc_ips,pways.kegg,groups=as.factor(grp),varStat="sd")
```

The three most significant pathways are:

```
> resTwoSam@tablePway[1:3]
```

```
                    PwayName        PwayID      APval PercOfGenesInPway
1: Oxidative phosphorylation path:hsa00190 0.09066667          71.42857
2:       Parkinson's disease path:hsa05012 0.09066667          69.23077
3:      Huntington's disease path:hsa05016 0.09066667          67.87565
   NumOfGenesFromDataSetInPathway PathwaySize
1:                             95         133
2:                             99         143
3:                            131         193
```

Let us look now at the significant pathway (p-value<0.1) and the standard deviation of five of the genes belonging to the "Oxidative phosphorylation" (for illustration). On the first line is the standard deviation of ESC (group 1) and the second line is IPS (group 2).

```
> sigTwoSam=sigPway(resTwoSam,0.1)
> rbind(resTwoSam@var1[sigTwoSam@genesInSigPways1[["Oxidative phosphorylation"]]]
+ ,resTwoSam@var2[sigTwoSam@genesInSigPways1[["Oxidative phosphorylation"]]])[,1:5]
```

```
        ATP5G1   ATP6V0E2      SDHC    NDUFV1    TCIRG1
[1,] 0.4207747 0.4706756 0.4206837 0.3501293 0.3676078
[2,] 0.4125640 0.4523365 0.2550743 0.3524234 0.4470001
```

We could now look at the visualization of this pathway "Oxidative phosphorylation". Figure 4 shows the density of expression variability for genes in this pathway for Group 1 (ESC) and Group 2 (IPS). We see that overall the IPS sample is shifted towards higher levels of variability compared to the ESC sample.

```
> plotPway(resTwoSam,"Oxidative phosphorylation",sigTwoSam)
```

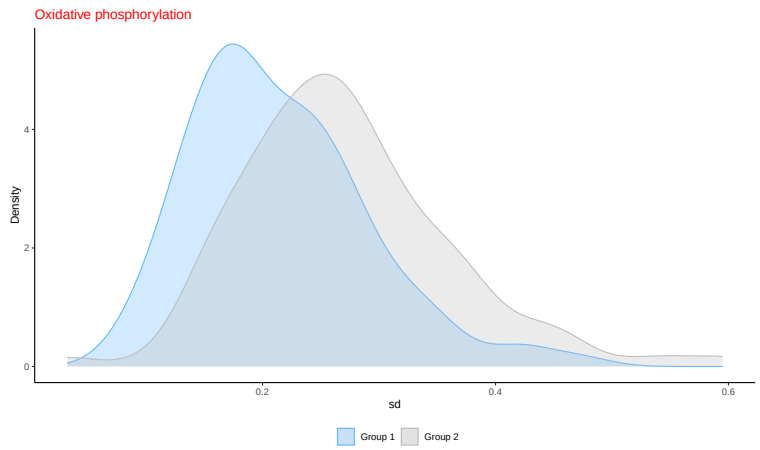

Figure 5: Densities of the standard deviation of ESC (group 1) and IPS (group 2) for the "Oxidative phosphorylation" pathway.

A next step to ask is, based on Figure 4, what genes are the most different between the two groups for this pathway? We can use the `getGenes` function to extract genes falling within a window of interest in the densities above, as specified by the user. For example, we can pick the window (0.25, 0.6) because at sd=0.25 this is where the two densities intersect and appear to deviate between the two groups.

```
> genes=getGenes(resTwoSam,"Oxidative phosphorylation",c(0.25,0.6))
> setdiff(genes@genes1,genes@genes2)

[1] "ATP6V1C1" "ATP5B"    "NDUFB5"
```

6 Finding pathways with a significant difference in variability between two samples using the distribution counts.

In situations where we want to identify variability changes between two contrasting phenotypes, we can also split the genes into several clusters (low, medium, high variability) and compared the distribution counts between the two groups.

The steps of the function for the two samples case using the distribution counts are as follow:

1. Compute the variability statistics (sd, mad, cv or mean) for each gene
2. Classify the genes with respect to the discrete levels of high, medium and low variability for the dataset corresponding to group 1 and the dataset corresponding to group 2 (3 clusters based on 33 and 66 percentile with all the samples).
3. For each pathway, we extract the gene in our dataset and in which cluster they belong.
4. For each pathway, we look at how the gene counts are distributed in each category for group 1 and compare it to these counts derived from group 2. The two possibilities to test this difference are the Chi-squared test or the multinomial exact test.
5. We build a data frame with the results of step 4.

For our example, we will look at the difference in standard deviation using the Reactome pathways and the two groups: ESC and IPS.

```
> resTwoSamDisc=pathVarTwoSamplesDisc(bock.esc_ips,pways.kegg,groups=as.factor(grp),
+ test="exact",varStat="sd")
```

The three most significant pathways are:

```
> resTwoSamDisc@tablePway[1:3]

                    PwayName        PwayID APval PercOfGenesInPway
1:                  Ribosome path:hsa03010     0          72.05882
2: Oxidative phosphorylation path:hsa00190     0          71.42857
3:      Huntington's disease path:hsa05016     0          67.87565
   NumOfGenesFromDataSetInPway PathwaySize
1:                          98         136
2:                          95         133
3:                         131         193
```

Now, we can use this result to look only at the significant pathways (with a p-value less than 0.01).

```
> sigTwoSamDisc=sigPway(resTwoSamDisc,0.01)
```

We could now look at the visualization of the first pathway "Ribosome". In Figure 5, we can see that categories 1 and 3 (low and high SD) were significantly different (p-value<0.05) between the two groups. We see that overall the IPS sample have more highly variable genes and less lowly variable genes than the ESC sample.

```
> plotPway(resTwoSamDisc,"Ribosome",sigTwoSamDisc)
```

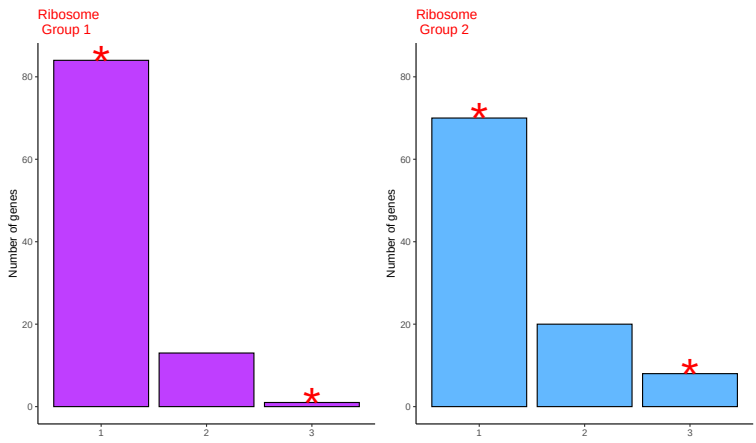

Figure 6: Distribution counts of the standard deviation of ESC (group 1) and IPS (group 2) for the "Ribosome" pathway.

We could also compare the distribution of variability between every gene in our two samples instead of just analyzing the genes in one pathway only.

```
> plotAllTwoSampleDistributionCounts(bock.esc_ips, resTwoSamDisc,
+ perc=c(1/3,2/3), pvalue=0.05, NULL)
```

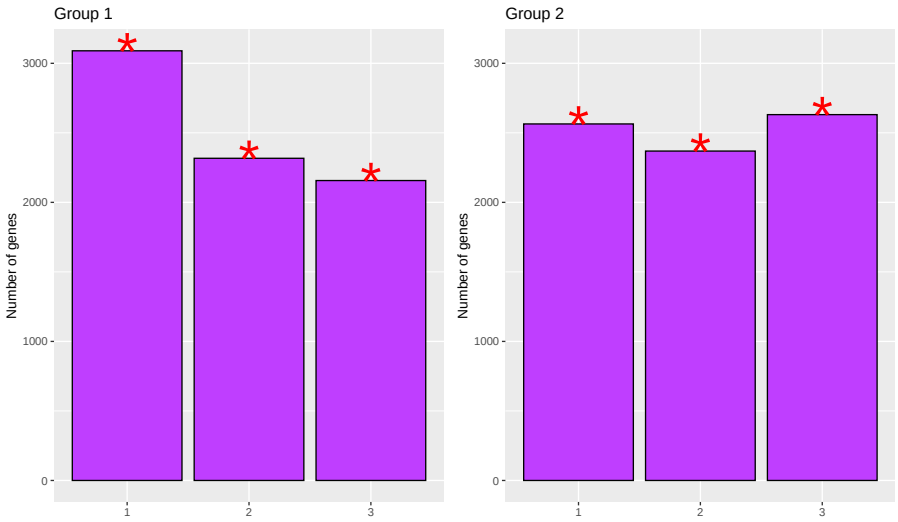

Figure 7: Distribution counts of the standard deviation of ESC (group 1) and IPS (group 2) for the all genes in the Bock data set.