

# GSReg: A Package for Gene Set Variability Analysis

Bahman Afsari<sup>1</sup> and Elana J. Fertig<sup>1</sup>

<sup>1</sup>The Sidney Kimmel Comprehensive Cancer Center,  
Johns Hopkins University School of Medicine

Modified: April 8, 2014. Compiled: October 17, 2016

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Input Data</b>	<b>2</b>
2.1	Data structure . . . . .	2
<b>3</b>	<b>Analysis of the pathways</b>	<b>4</b>
3.1	DIRAC Analysis . . . . .	5
3.2	EVA . . . . .	6
3.3	Comparison of DIRAC and EVA . . . . .	7
<b>4</b>	<b>System Information</b>	<b>22</b>
<b>5</b>	<b>Literature Cited</b>	<b>22</b>

## 1 Introduction

The **GSReg** package allows to analyze pathways based on the variability of the expression of sets of genes that are targets of those pathways. Basing this set statistic on variability enables inference of dysregulated pathways in diseases, including notably cancers. The first set statistic for gene variability was in the work of Eddy and his colleagues (see [1]) which used a ranked based methodology called *DIRAC*. *DIRAC* calculates a measure of variability of the ordering of the expression of genes in a pathway for specific phenotype. The basic idea behind *DIRAC* is to generate a template for the pair-wise comparisons of gene expressions

of a pathway within a phenotype. DIRAC calculates a measure of the variability of the ordering within the phenotype, i.e. the expected distance of a sample from the phenotype and the template of the phenotype. In mathematical terms, if we denote two i.i.d. samples from the same phenotypes by  $X$  and  $X'$  and  $D$  Kendall- $\tau$ -distance on the specific pathways, then the EVA statistic is  $E(D(X, X'))$ . It identifies significantly dysregulated pathways by estimating p-values from a permutation test. Eddy et al. found that more pathological phenotypes usually have more pathways with higher variability compared to less pathological phenotypes.

However, the permutation test in DIRAC is computationally intensive and reaching low p-values may be impractical since they require a huge number of permutations. Low p-values are required for multiple hypothesis correction. A similar measure of variability of the orderings of gene sets was proposed in [2]. This method approximates the p-value theoretically, without a permutation test. This method is based on Kendall- $\tau$  distance [3] and the theory of U-Statistics, thus we call this method Gene Set Expression Variation Analysis (or in short EVA). Specifically, Kendall- $\tau$  distance between two expression profiles counts the number of disagreeing pairwise comparisons between two profiles. The EVA measures the variability of the gene expression of pathway genes from a phenotype by calculating the expectation of Kendall- $\tau$  distance between two random samples from the phenotype. EVA then identifies if the variability is significantly different across two phenotypes. To approximate this p-value EVA applies a U-Statistic Theory approach.

The **GReg** package contains two following utilities:

1. Identifying the dysregulated pathways with *DIRAC* measure of variability. The significance is calculated using permutation test. This is the first time that DIRAC analysis has been implemented in *R*. It also is more adaptable to new datasets than the original Matlab code in [1].
2. Identifying the dysregulated pathways with *EVA* measure of variability. The significance is approximated through applying U-statistics theory. This is very time efficient and consistent with both *DIRAC* and applying permutation test on EVA.

## 2 Input Data

### 2.1 Data structure

In short, the **GReg** package requires the following data in the following format:

1. Gene Expression Data

- (a) The expression be in the form of a matrix where rows represent genes (or probes) and columns represent samples.
  - (b) The expression matrix cannot have NAs.
  - (c) The expression matrix rows must have names of genes or the probes.
2. Pathways
- (a) The list of pathways must contain character vectors. Only the elements of the vectors which appear in rownames of the expression matrix are considered for analysis.
  - (b) The list of the pathways must have names for each vectors.
3. Phenotypes
- (a) A factor with binary levels.

We used the data provided in the **GSBenchMark** package to reproduce the results in Eddy et al. [1]. The **GSBenchMark** contains data for the pathways as well as the gene expression and phenotype data from twelve studies. We load the information about the pathways from

**GSBenchMark:**

```
> library(GSBenchMark)
> data(diracpathways)
> class(diracpathways)

[1] "list"

> names(diracpathways)[1:5]

[1] "DEATHPATHWAY"      "TCAPOPTOSISPATHWAY" "CCR3PATHWAY"
[4] "NEUTROPHILPATHWAY" "ALTERNATIVEPATHWAY"

> class(diracpathways[[1]])

[1] "character"
```

AS mentioned **GSReg** package requires the information of the pathways to be as a list of character vectors. Also, **GSReg** requires the pathways to have names. The variable **diracpathways** contains gene pathways. It is a list. Each element represents a pathway with its name. Each elements contains a list of characters which represent the genes in the pathway. e.g. `diracpathways[["DEATHPATHWAY"]]`.

Now, we load the datasets' names:

```

> data(GSBenchMarkDatasets)
> print(GSBenchMark.Dataset.names)

[1] "leukemia_GSEA"          "marfan_GDS2960"          "melanoma_GDS2735"
[4] "parkinsons_GDS2519"    "prostate_GDS2545_m_nf"  "prostate_GDS2545_m_p"
[7] "prostate_GDS2545_p_nf" "sarcoma_data"           "squamous_GDS2520"
[10] "breast_GDS807"         "bipolar_GDS2190"

```

The remaining examples in this vignette rely on one of the datasets, i.e. “squamous GDS2520.” Similar analyses may be reproduced for other datasets by selecting a different element of “GS-BenchMark.Dataset.names.”

```

> DataSetStudy = GSBenchMark.Dataset.names[[9]]
> print(DataSetStudy)

[1] "squamous_GDS2520"

> data(list=DataSetStudy)

```

The data consists of two variables: **exprsdata** and **phenotypes**. **exprsdata** consists of a gene expression matrix where the rows and columns represent genes and the samples respectively. **GSReg** requires the rownames of gene expression variable represent the gene names, *i.e.* they are represented in the pathway information variable.

The **GSReg** does not allow any missing data. To comply with the requirements we remove genes with NAs. The user may use any imputation to resolve this issue:

```

> if(sum(apply(is.nan(exprsdata),1,sum)>0))
  exprsdata = exprsdata[-which(apply(is.nan(exprsdata),1,sum)>0),];

```

One can extract the gene names by:

```

> genenames = rownames(exprsdata);
> genenames[1:10]

[1] "MAPK3"    "TIE1"    "CYP2C19" "CXCR5"   "CXCR5"   "DUSP1"   "MMP10"   "DDR1"
[9] "EIF2AK2" "HINT1"

```

### 3 Analysis of the pathways

Here, we demonstrate how to use the **GSReg** package to compute DIRAC and EVA statistics.

### 3.1 DIRAC Analysis

First, we load the library:

```
> library(GSReg)
```

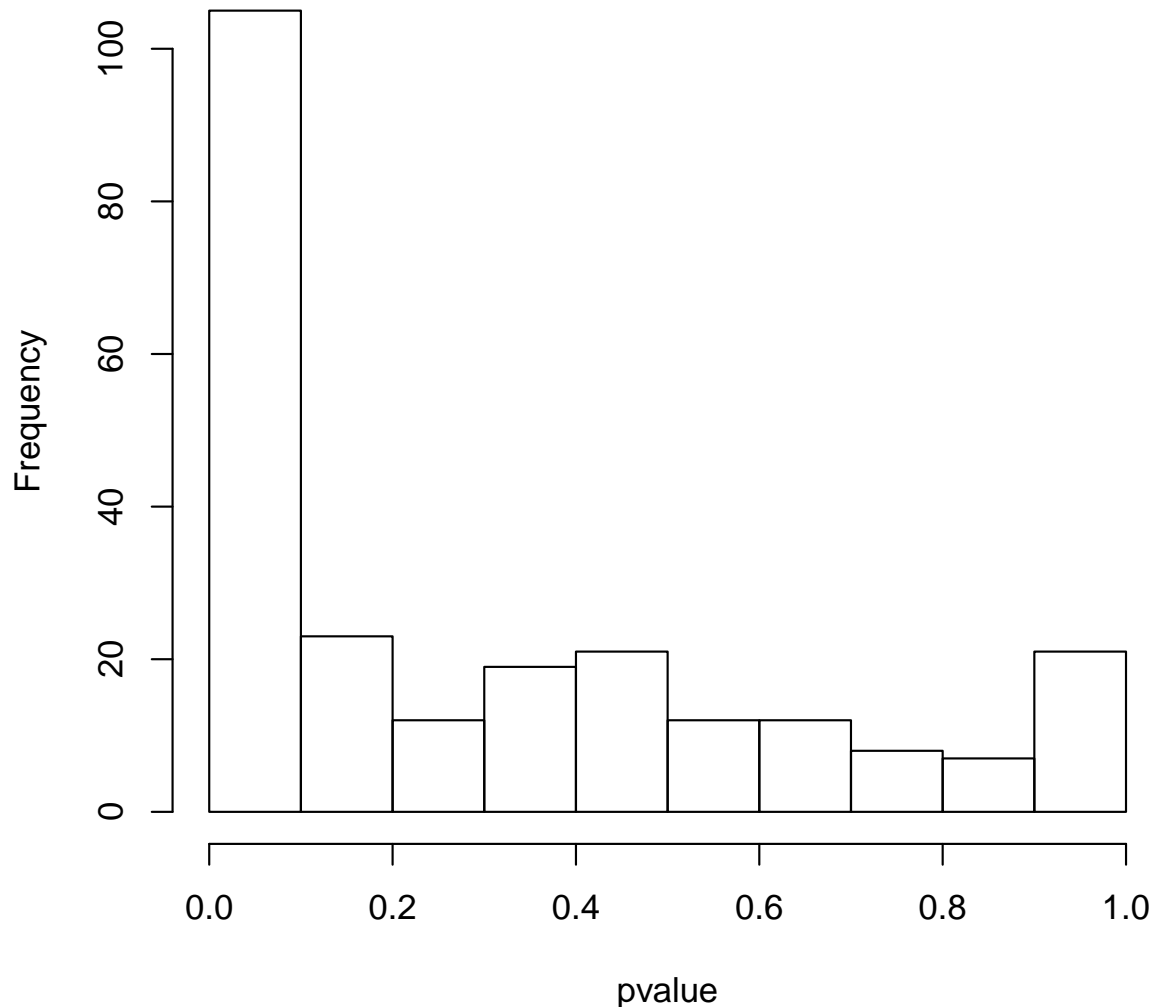
The package also implements the alternative EVA statistic in the function `GSReg.GeneSets.DIRAC`. This function receives gene expression as `geneexpres`, the pathway information as `pathways` and phenotypes of samples as a factor with two levels and length equal to column number of `geneexpres`. *DIRAC* uses a permutation test for p-value calculation; so, `GSReg.GeneSets.DIRAC` receives the number of permutations through (`Nperm`) with default value equal to 1000.

```
> Nperm = 10
> system.time({DIRACan =GSReg.GeneSets.DIRAC(exprsdata,diracpathways,phenotypes,Nperm=Nperm)})
   user  system elapsed 
  2.80    0.00    2.79
```

Here is the histogram of the DIRAC p-values:

```
> hist(DIRACan$pvalues,xlab="pvalue",main="Hist of pvalues applying DIRAC Analysis.")
```

## Hist of pvalues applying DIRAC Analysis.



### 3.2 EVA

The package also implements the alternative EVA statistic in the function `GSReg.GeneSets.EVA`. The function requires the similar inputs as `GSReg.GeneSets.DIRAC` (i.e. `geneexpres`, `pathways`, `phenotypes`) except it does not need `Nperm` since the p-value is not calculated through permutation test but through the mentioned U-statistic theory approach.

```
> #Calculating the variance for the pathways  
> #Calculate how much it takes to calculate the statistics and their p-value for all pathways
```

```

>
> system.time({VarAnKendallV = GSReg.GeneSets.EVA(geneexpres=exprsdata,
  pathways=diracpathways, phenotypes=as.factor(phenotypes)) })
      user  system elapsed
    0.33    0.00    0.33

> names(VarAnKendallV)[[1]]
[1] "DEATHPATHWAY"

> VarAnKendallV[[1]]

$E1
[1] 0.09441352

$E2
[1] 0.1310383

$zscore
[1] -3.711215

$VarEta1
[1] 3.480369e-05

$VarEta2
[1] 6.248694e-05

$sdttotal
[1] 0.009863601

$pvalue
[1] 0.0002062669

```

The output consists of a list. Each element of the list corresponds to a pathway. The element itself is a list. *E1* and *E2* are two fields which contain the measure of variability for phenotype levels(`phenotypes`) [1] and levels(`phenotypes`) [2] respectively. Other list elements are `pvalue` and `zscore` which are calculated through the theory of U-statistics and indicate the statistical significance of the difference between *E1* and *E2*.

### 3.3 Comparison of DIRAC and EVA

We ran the following code to compare statistics from DIRAC and from EVA.

```

> Nperm = 10;
> VarAnPerm = vector(mode="list",length=Nperm)
> for( i in seq_len(Nperm))
{

```

```

    VarAnPerm[[i]] = GSReg.GeneSets.EVA(geneexpres=exprsdata, pathways=diracpathways,
                                         phenotypes=sample(phenotypes))
  }
> pvaluesperm = vector(mode="numeric",length=length(VarAnPerm[[1]]))
> for( i in seq_along(VarAnPerm[[1]]))
{
  z = sapply(VarAnPerm,function(x) x[[i]]$E1 - x[[i]]$E2)
  pvaluesperm[i] = mean(abs(VarAnKendallV[[i]]$E1-VarAnKendallV[[i]]$E2)<abs(z))
}
> zscore = sapply(VarAnKendallV,function(x) x$zscore);
> pvalustat = sapply(VarAnKendallV,function(x) x$pvalue);

```

The figure represents that the theoretical p-value and p-value calculated from permutation test in EVA are very similar and we can use the theoretical p-value as a surrogate for p-value. Here is the histogram.

```

> hist(x=pvalustat,breaks=20,main="P-value Hist of U-Stat",xlim=c(0,1))

```



```

> plot(x=abs(zscore),y=pvaluesperm,xlab="|Z-score|",
      ylab="p-value",col="red1",main="p-value comparisons")
> zscorelin = seq(0,6,0.1);
> pvaltheoretic = (1-pnorm(zscorelin))*2
> lines(x=zscorelin,y=pvaltheoretic,type="l",pch=50,lty=5,col="darkblue")
> legend("topright",legend=c("permutation test","U-Stat Estimation"),
      col=c("red","blue"),text.col=c("red","blue"),
      lty=c(NA,1),lwd=c(NA,2.5),pch=c(21,NA))

```

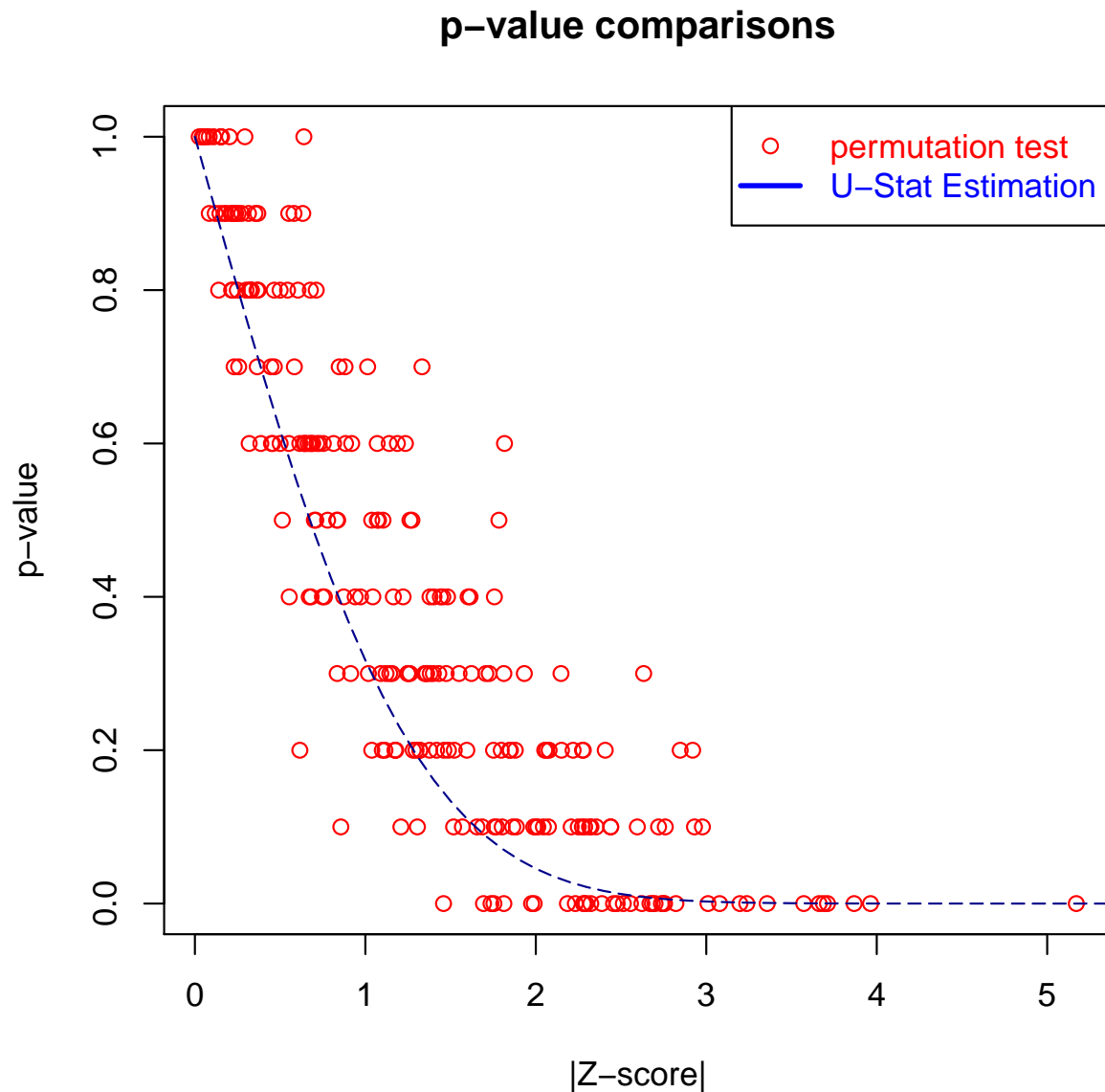


Figure 1: Comparing p-value from permutation test and U-statistic theory with only 10 permutations.

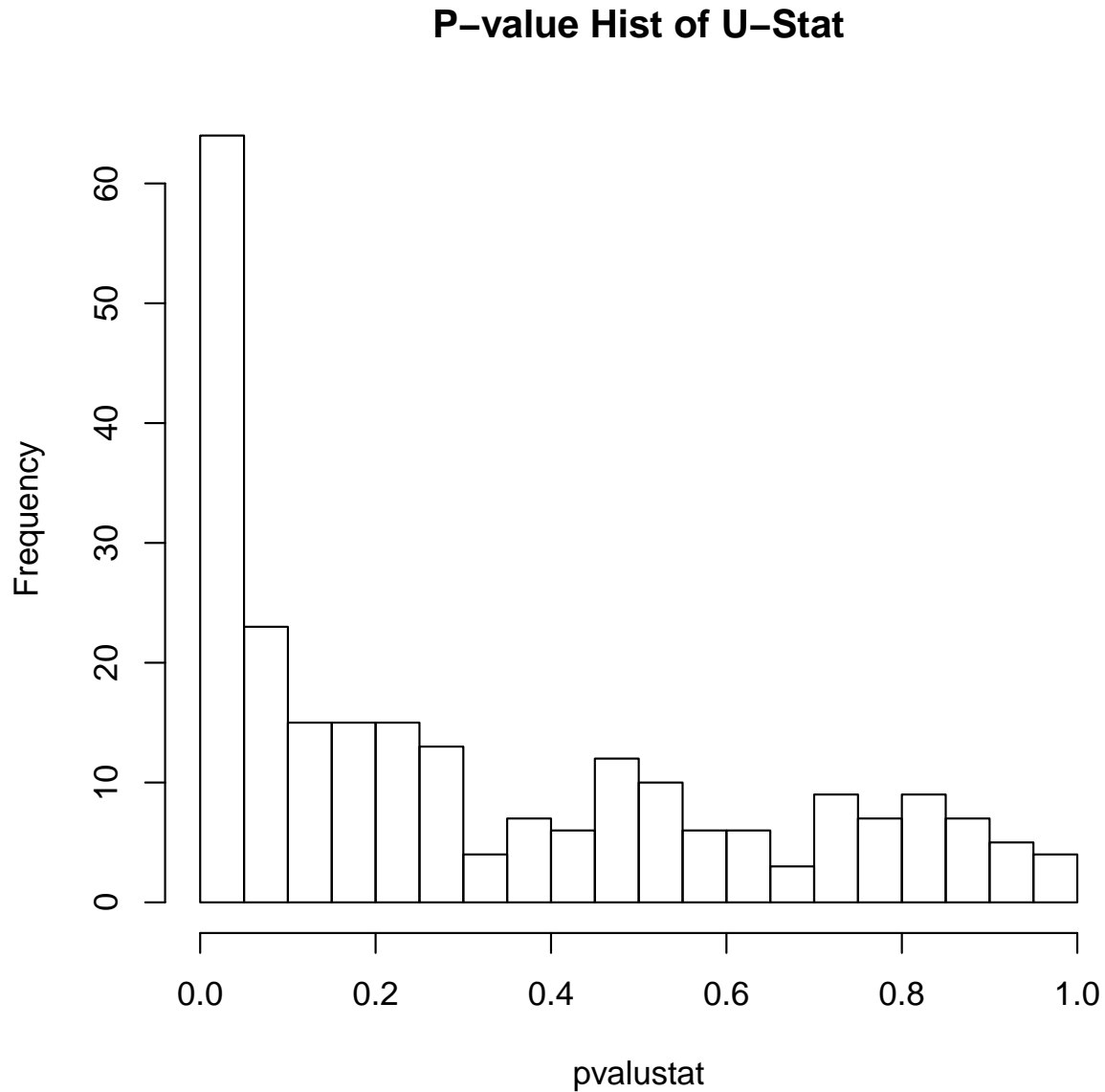


Figure 2 shows the result of comparing p-value EVA computing from 1000 permutation test and approximation using U-statistics theory (offline generated).

To compare with the p-value of the DIRAC analysis, we show the p-values of DIRAC versus U-Statistic methodology:

```
> plot(x=DIRACAn$pvalues,y=pvalustat,xlab="DIRAC",
       ylab="EVA",main=sprintf("P-value Comparison corr=%2.2g",cor(x=DIRACAn$pvalues,y=pvalustat)))
> lmfit = lm(pvalustat~DIRACAn$pvalues-1)
> abline(lmfit)
> cor.test(x=DIRACAn$pvalues,y=pvalustat)
```

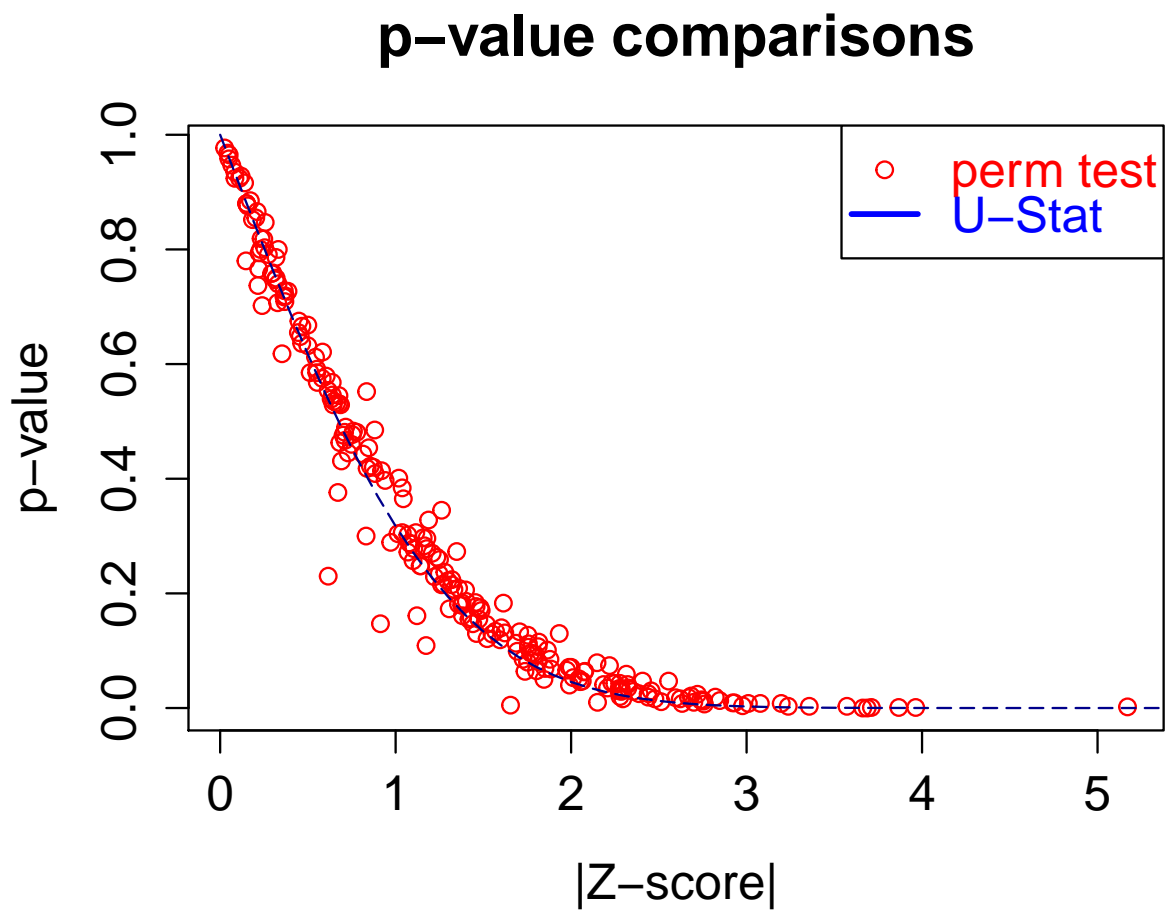
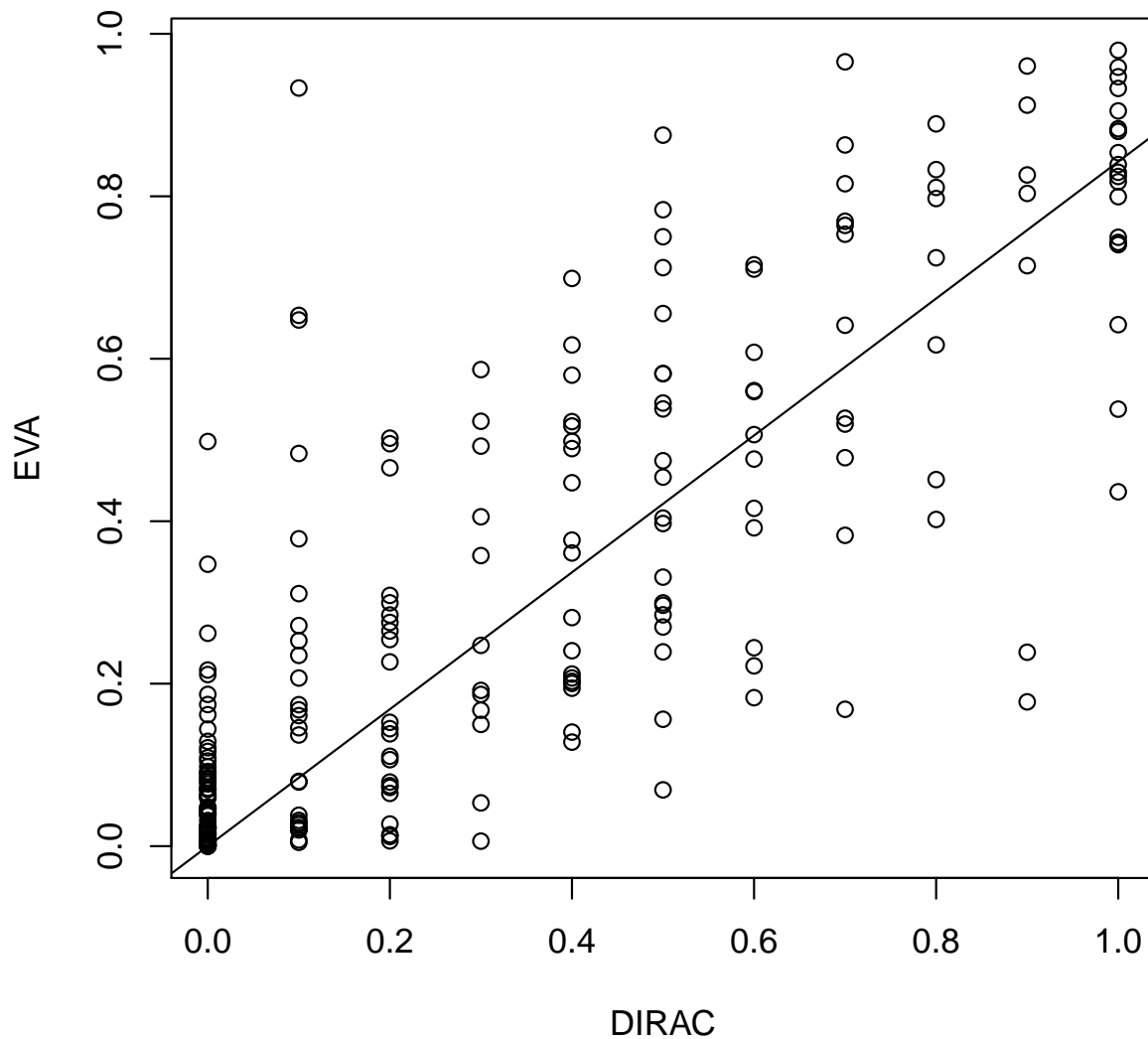


Figure 2: Theoretical p-value versus empirical p-value using 1000 permutations.

Pearson's product-moment correlation

```
data: DIRACn$pvalues and pvalustat
t = 22.599, df = 238, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7809504 0.8623471
sample estimates:
      cor
0.8259051
```

### P-value Comparison corr=0.83



Also, the correlation of the p-values of DIRAC and U-Statistics is very high:

```
> cor(x=DIRACan$pvalues,y=pvalustat)
[1] 0.8259051
```

If we use 1000 permutations instead of 10 permutations, we can see that the correlation is higher (0.88) as seen in Figure (3). The dysregulated pathways identified by *DIRAC* are the following pathways:

[1] "DEATHPATHWAY"	"NEUTROPHILPATHWAY"	"PGC1APATHWAY"
[4] "RARRXRPATHWAY"	"SKP2E2FPATHWAY"	"KERATINOCYTEPATHWAY"
[7] "CHEMICALPATHWAY"	"TGFBPATHWAY"	"PROTEASOMEPATHWAY"
[10] "MAPKPATHWAY"	"PDGFPATHWAY"	"BIOPEPTIDEPATHWAY"
[13] "SPPAPATHWAY"	"PYK2PATHWAY"	"MYOSINPATHWAY"
[16] "BETAOXIDATIONPATHWAY"	"IL7PATHWAY"	"FMLPPATHWAY"
[19] "VITCBPATHWAY"	"CD40PATHWAY"	"CDC25PATHWAY"
[22] "MTORPATHWAY"	"RNAPATHWAY"	"FBW7PATHWAY"
[25] "LYMPHOCYTEPATHWAY"	"LAIRPATHWAY"	"HIVNEFPATHWAY"
[28] "ALKPATHWAY"	"P35ALZHEIMERSPATHWAY"	"MSPPATHWAY"
[31] "GSK3PATHWAY"	"RELAPATHWAY"	"METPATHWAY"
[34] "TNFR2PATHWAY"	"AT1RPATHWAY"	"FREEPATHWAY"
[37] "ARAPPATHWAY"	"MRPPATHWAY"	"P53HYPOXIAPATHWAY"
[40] "IL18PATHWAY"	"STRESSPATHWAY"	"MEF2DPATHWAY"
[43] "STAT3PATHWAY"	"HSP27PATHWAY"	"EPONFKBPATHWAY"
[46] "NKCELLSPATHWAY"	"MONOCYTEPATHWAY"	"CARM_ERPATHWAY"

DIRAC and EVA have been shown mathematically similar. The main advantages of the EVA is efficiency in calculation as well as easier interpretation. Figure 3 a graphical example of such comparison. One can see that the p-values generated by DIRAC and EVA have high correlation, i.e. 0.88. Note that EVA is much faster than DIRAC. For example, in this case, we ran the computations on a Lenovo Thinkpad with Core(TM) i7-3720QM Intel CPU @2.6 GHz. For a thousand permutation, the DIRAC analysis took 207.47 seconds while the latter only took 0.3 seconds. Note that for multiple hypothesis adjustment, a thousand permutations may not be satisfactory and we require hundreds of thousand or a million permutation which may not be feasible.

Note that it is possible that some of the genes in a pathway are not represented in the expression data or are too short (e.g. less than 5 genes). Both `GSReg.GeneSets.EVA` and `GSReg.GeneSets.DIRAC` may ignore such pathways through parameter `minGeneNum`. Please see the manual for more details. If we the user wants to compare the results of DIRAC and

EVA, they can run the following code for plot DIRAC diagram of significantly perturbed pathways:

```
> DIRACan =GSReg.GeneSets.DIRAC(exprsdata,diracpathways,phenotypes,Nperm=1000)
```

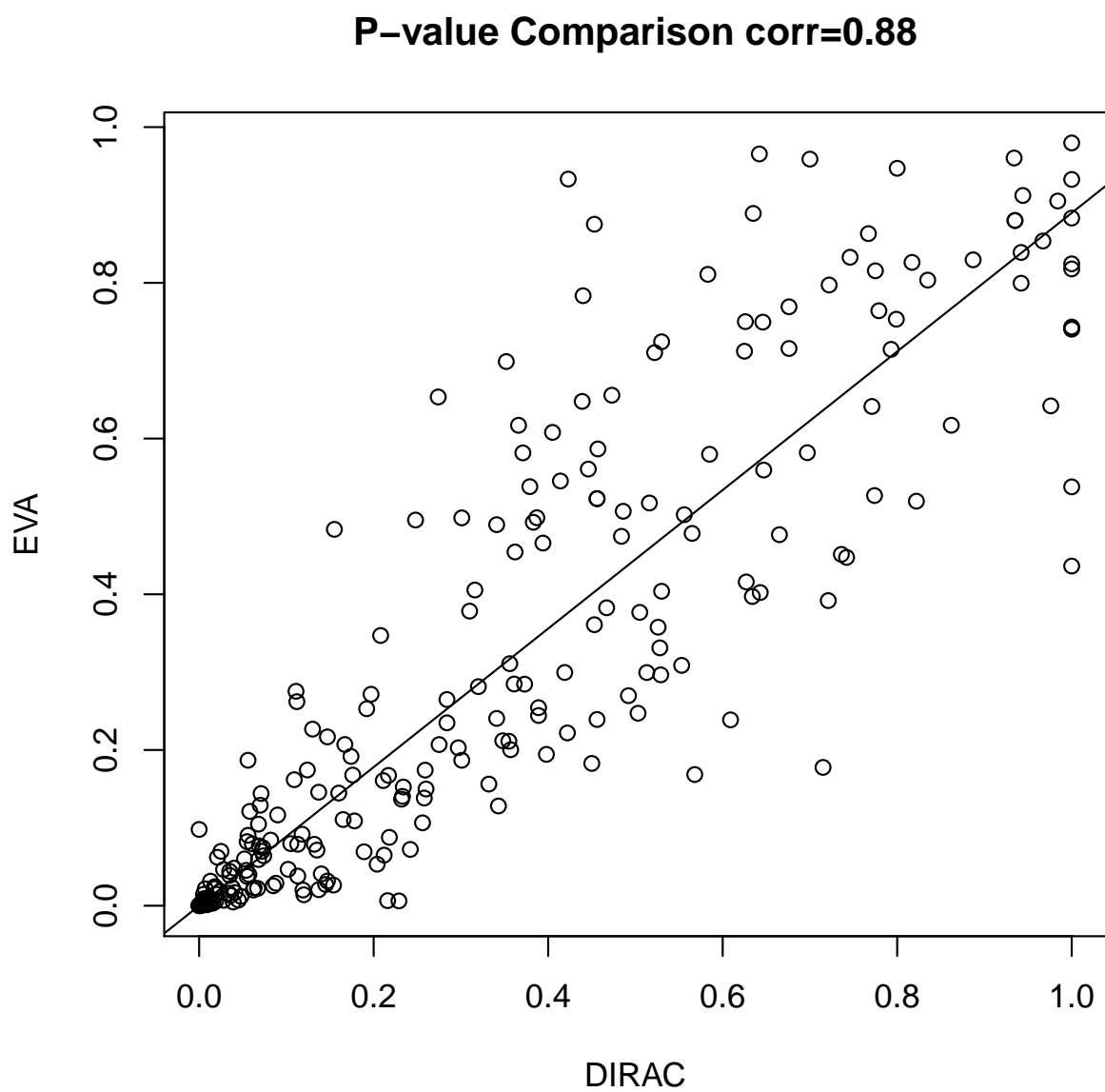


Figure 3: Comparing p-values EVA versus DIRAC. The correlation is 0.88.

```

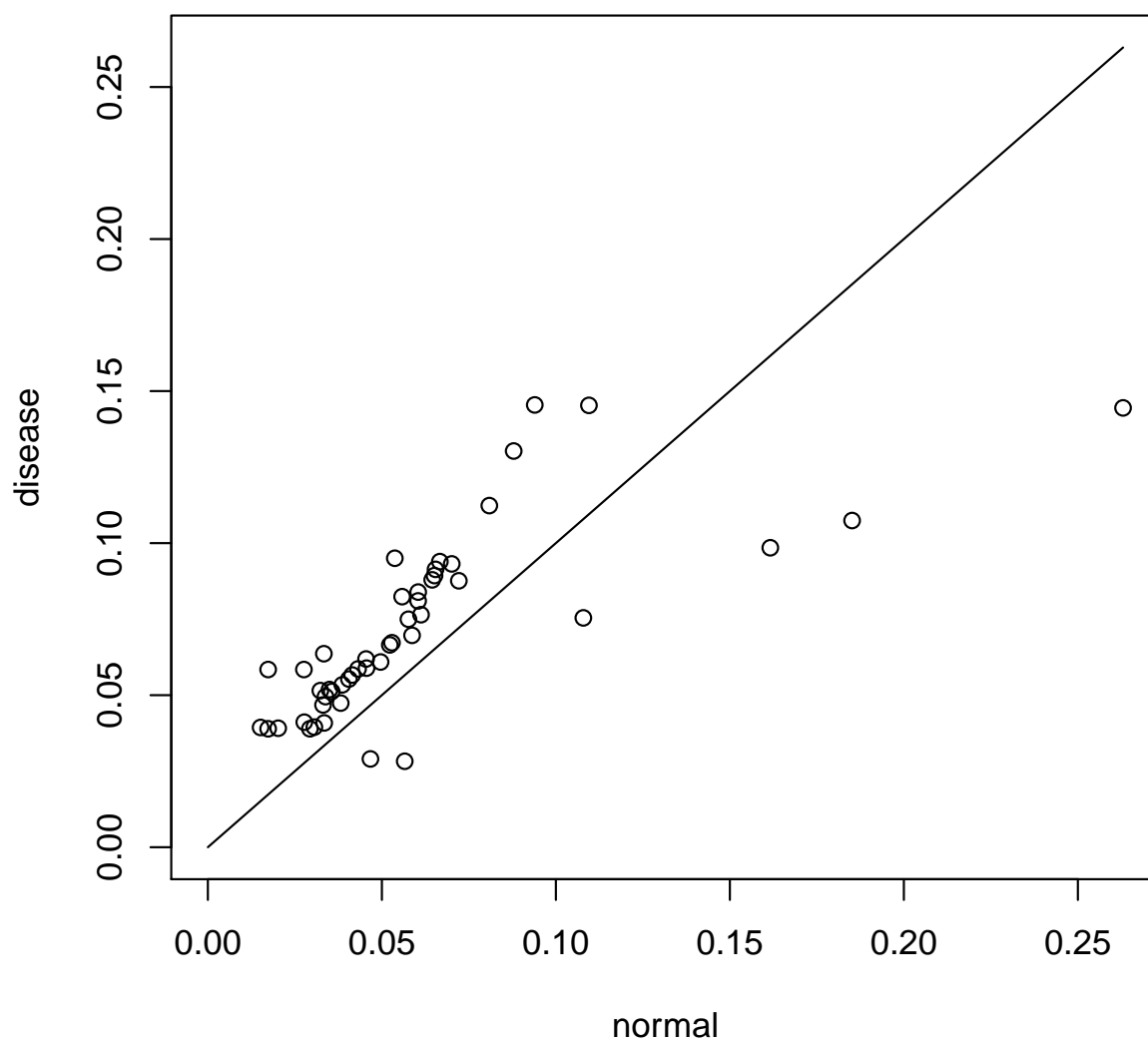
> significantPathwaysDIRAC = names(DIRACAn$mu1)[which(DIRACAn$pvalues<0.05)];
> mu1 = DIRACAn$mu1[significantPathwaysDIRAC];
> mu2 = DIRACAn$mu2[significantPathwaysDIRAC];
> #The dysregulated pathways
> names(mu1)

[1] "DEATHPATHWAY"          "NEUTROPHILPATHWAY"    "PGC1APATHWAY"
[4] "RARRXRPATHWAY"        "SKP2E2FPATHWAY"       "KERATINOCYTEPATHWAY"
[7] "CHEMICALPATHWAY"      "TGFBPATHWAY"          "PROTEASOMEPATHWAY"
[10] "MAPKPATHWAY"          "PDGFPATHWAY"          "BIOPEPTIDEPATHWAY"
[13] "SPPAPATHWAY"          "PYK2PATHWAY"          "MYOSINPATHWAY"
[16] "BETAOXIDATIONPATHWAY" "IL7PATHWAY"           "FMLPPATHWAY"
[19] "VITCBPATHWAY"         "CD40PATHWAY"          "CDC25PATHWAY"
[22] "MTORPATHWAY"          "RNAPATHWAY"           "FBW7PATHWAY"
[25] "LYMPHOCYTEPATHWAY"    "LAIRPATHWAY"          "HIVNEFPATHWAY"
[28] "ALKPATHWAY"           "P35ALZHEIMERSPATHWAY" "MSPPATHWAY"
[31] "GSK3PATHWAY"          "RELAPATHWAY"          "METPATHWAY"
[34] "TNFR2PATHWAY"         "AT1RPATHWAY"          "FREEPATHWAY"
[37] "ARAPPATHWAY"          "MRPPATHWAY"           "P53HYPOXIAPATHWAY"
[40] "IL18PATHWAY"          "STRESSPATHWAY"        "MEF2DPATHWAY"
[43] "STAT3PATHWAY"         "HSP27PATHWAY"         "EPONFKBPATHWAY"
[46] "NKCELLSPATHWAY"       "MONOCYTEPATHWAY"      "CARM_ERPATHWAY"

> plot(x=mu1,y=mu2,
       xlim=c(0,max(mu1,mu2)),ylim=c(0,max(mu1,mu2)),xlab="normal",ylab="disease",
       main="(a) DIRAC significantly dysregulated pathways")
> lines(x=c(0,max(mu1,mu2)),y=c(0,max(mu1,mu2)))

```

### (a) DIRAC significantly dysregulated pathways



Now, if we do the analysis using EVA, we have:

```
> significantPathwaysGSV = names(which(pvalustat<0.05));
```

[1] "DEATHPATHWAY"	"TCAPOPTOSISPATHWAY"	"NEUTROPHILPATHWAY"
[4] "PGC1APATHWAY"	"TERCPATHWAY"	"RARRXRPATHWAY"
[7] "SKP2E2FPATHWAY"	"KERATINOCYTEPATHWAY"	"CHEMICALPATHWAY"
[10] "METHIONINEPATHWAY"	"TGFBPATHWAY"	"PS1PATHWAY"
[13] "PROTEASOMEPATHWAY"	"CDK5PATHWAY"	"MAPKPATHWAY"
[16] "NTHIPATHWAY"	"PDGFPATHWAY"	"BIOPEPTIDEPATHWAY"
[19] "SPPAPATHWAY"	"PYK2PATHWAY"	"CDC42RACPATHWAY"



```

[22] "MYOSINPATHWAY"      "BETAOXIDATIONPATHWAY" "IL7PATHWAY"
[25] "FMLPPATHWAY"        "FASPATHWAY"           "VITCBPATHWAY"
[28] "CD40PATHWAY"        "IGF1PATHWAY"          "CDC25PATHWAY"
[31] "MTORPATHWAY"        "RNAPATHWAY"           "FBW7PATHWAY"
[34] "LYMPHOCYTEPATHWAY"  "LAIRPATHWAY"          "HIVNEFPATHWAY"
[37] "ALKPATHWAY"         "PEPIPATHWAY"          "MSPPATHWAY"
[40] "EDG1PATHWAY"        "GSK3PATHWAY"          "RELAPATHWAY"
[43] "METPATHWAY"         "TNFR2PATHWAY"         "AT1RPATHWAY"
[46] "ATRBRCPATHWAY"      "GLYCOLYSISPATHWAY"    "TIDPATHWAY"
[49] "EPOPATHWAY"         "WNTPATHWAY"           "ARAPPATHWAY"
[52] "MRPPATHWAY"         "P53HYPOXIAPATHWAY"    "PITX2PATHWAY"
[55] "IL18PATHWAY"        "STRESSPATHWAY"        "MEF2DPATHWAY"
[58] "MITOCHONDRIAPATHWAY" "STAT3PATHWAY"         "EPONFKBPATHWAY"
[61] "NKCELLSPATHWAY"     "MONOCYTEPATHWAY"      "CARM_ERPATHWAY"
[64] "HCMVPATHWAY"

```

```
> eta1 = sapply(VarAnKendallV,function(x) x$E1)[significantPathwaysGSV];
```

DEATHPATHWAY	TCAPOPTOSISPATHWAY	NEUTROPHILPATHWAY	PGC1APATHWAY
0.09441352	0.08600289	0.35559678	0.04477053
TERCPATHWAY	RARRXRPATHWAY	SKP2E2FPATHWAY	KERATINOCYTEPATHWAY
0.07316017	0.06914038	0.03367003	0.07404055
CHEMICALPATHWAY	METHIONINEPATHWAY	TGFBPATHWAY	PS1PATHWAY
0.08521303	0.09913420	0.05028305	0.08080808
PROTEASOMEPATHWAY	CDK5PATHWAY	MAPKPATHWAY	NTHIPATHWAY
0.09617180	0.07975863	0.08504987	0.08091115
PDGFPATHWAY	BIOPEPTIDESPATHWAY	SPPAPATHWAY	PYK2PATHWAY
0.08698709	0.06767807	0.09690598	0.04711514
CDC42RACPATHWAY	MYOSINPATHWAY	BETAOXIDATIONPATHWAY	IL7PATHWAY
0.08948195	0.09005280	0.02683983	0.10609668
FMLPPATHWAY	FASPATHWAY	VITCBPATHWAY	CD40PATHWAY
0.05736961	0.08676830	0.07888408	0.05294705
IGF1PATHWAY	CDC25PATHWAY	MTORPATHWAY	RNAPATHWAY
0.09115972	0.06265031	0.04188827	0.03009689
FBW7PATHWAY	LYMPHOCYTEPATHWAY	LAIRPATHWAY	HIVNEFPATHWAY
0.03174603	0.22390572	0.15222872	0.09020600
ALKPATHWAY	PEPIPATHWAY	MSPPATHWAY	EDG1PATHWAY
0.09130361	0.05194805	0.12150072	0.08801738
GSK3PATHWAY	RELAPATHWAY	METPATHWAY	TNFR2PATHWAY
0.11946928	0.06302309	0.10561315	0.05051566
AT1RPATHWAY	ATRBRCPATHWAY	GLYCOLYSISPATHWAY	TIDPATHWAY
0.05943314	0.07166907	0.02308802	0.07331013
EPOPATHWAY	WNTPATHWAY	ARAPPATHWAY	MRPPATHWAY
0.07319696	0.10526414	0.05988456	0.04877345
P53HYPOXIAPATHWAY	PITX2PATHWAY	IL18PATHWAY	STRESSPATHWAY
0.07708666	0.09375387	0.13015873	0.05505364
MEF2DPATHWAY	MITOCHONDRIAPATHWAY	STAT3PATHWAY	EPONFKBPATHWAY
0.04399740	0.12572150	0.05179344	0.07910272
NKCELLSPATHWAY	MONOCYTEPATHWAY	CARM_ERPATHWAY	HCMVPATHWAY
0.07718643	0.25171192	0.06462137	0.07781385

```
> eta2 = sapply(VarAnKendallV,function(x) x$E2)[significantPathwaysGSV];
```

DEATHPATHWAY	TCAPOPTOSISPATHWAY	NEUTROPHILPATHWAY	PGC1APATHWAY
0.13103827	0.04531025	0.23546691	0.05588351
TERCPATHWAY	RARRXRPATHWAY	SKP2E2FPATHWAY	KERATINOCYTEPATHWAY
0.12770563	0.09090909	0.05856181	0.09013983
CHEMICALPATHWAY	METHIONINEPATHWAY	TGFBPATHWAY	PS1PATHWAY
0.11832612	0.15454545	0.07165057	0.12332852
PROTEASOMEPATHWAY	CDK5PATHWAY	MAPKPATHWAY	NTHIPATHWAY
0.13083213	0.10251869	0.10114801	0.09532055
PDGFPATHWAY	BIOPEPTIDESPATHWAY	SPPAPATHWAY	PYK2PATHWAY
0.11066711	0.08559859	0.12604711	0.06023958
CDC42RACPATHWAY	MYOSINPATHWAY	BETAOXIDATIONPATHWAY	IL7PATHWAY
0.11283954	0.11997526	0.06147186	0.13455988
FMLPPATHWAY	FASPATHWAY	VITCBPATHWAY	CD40PATHWAY
0.06835017	0.10725265	0.04252044	0.08041958
IGF1PATHWAY	CDC25PATHWAY	MTORPATHWAY	RNAPATHWAY
0.12087036	0.04413179	0.05994640	0.08719852
FBW7PATHWAY	LYMPHOCYTEPATHWAY	LAIRPATHWAY	HIVNEFPATHWAY
0.06307978	0.16065416	0.11574140	0.11884884
ALKPATHWAY	PEPIPATHWAY	MSPPATHWAY	EDG1PATHWAY
0.11181840	0.13593074	0.17344877	0.11010728
GSK3PATHWAY	RELAPATHWAY	METPATHWAY	TNFR2PATHWAY
0.16320909	0.08203463	0.12858234	0.07308378
AT1RPATHWAY	ATRBRCPATHWAY	GLYCOLYSISPATHWAY	TIDPATHWAY
0.08115533	0.08771185	0.05112348	0.09458733
EPOPATHWAY	WNTPATHWAY	ARAPPATHWAY	MRPPATHWAY
0.09569080	0.12762130	0.07615440	0.08571429
P53HYPOXIAPATHWAY	PITX2PATHWAY	IL18PATHWAY	STRESSPATHWAY
0.09759247	0.11618223	0.18989899	0.07523998
MEF2DPATHWAY	MITOCHONDRIAPATHWAY	STAT3PATHWAY	EPONFKBPATHWAY
0.05508870	0.16778499	0.09647495	0.13372688
NKCELLSPATHWAY	MONOCYTEPATHWAY	CARM_ERPATHWAY	HCMVPATHWAY
0.09574739	0.17662338	0.08397641	0.08899711

> #The dysregulated pathways  
> names(etal)

```
[1] "DEATHPATHWAY"      "TCAPOPTOSISPATHWAY" "NEUTROPHILPATHWAY"
[4] "PGC1APATHWAY"      "TERCPATHWAY"        "RARRXRPATHWAY"
[7] "SKP2E2FPATHWAY"    "KERATINOCYTEPATHWAY" "CHEMICALPATHWAY"
[10] "METHIONINEPATHWAY" "TGFBPATHWAY"        "PS1PATHWAY"
[13] "PROTEASOMEPATHWAY" "CDK5PATHWAY"        "MAPKPATHWAY"
[16] "NTHIPATHWAY"       "PDGFPATHWAY"        "BIOPEPTIDESPATHWAY"
[19] "SPPAPATHWAY"       "PYK2PATHWAY"        "CDC42RACPATHWAY"
[22] "MYOSINPATHWAY"     "BETAOXIDATIONPATHWAY" "IL7PATHWAY"
[25] "FMLPPATHWAY"       "FASPATHWAY"         "VITCBPATHWAY"
[28] "CD40PATHWAY"       "IGF1PATHWAY"        "CDC25PATHWAY"
[31] "MTORPATHWAY"       "RNAPATHWAY"         "FBW7PATHWAY"
[34] "LYMPHOCYTEPATHWAY" "LAIRPATHWAY"        "HIVNEFPATHWAY"
[37] "ALKPATHWAY"        "PEPIPATHWAY"        "MSPPATHWAY"
[40] "EDG1PATHWAY"       "GSK3PATHWAY"        "RELAPATHWAY"
[43] "METPATHWAY"        "TNFR2PATHWAY"       "AT1RPATHWAY"
[46] "ATRBRCPATHWAY"     "GLYCOLYSISPATHWAY"  "TIDPATHWAY"
```

```

[49] "EPOPATHWAY"          "WNTPATHWAY"          "ARAPPATHWAY"
[52] "MRPPATHWAY"          "P53HYPOXIAPATHWAY"  "PITX2PATHWAY"
[55] "IL18PATHWAY"         "STRESSPATHWAY"       "MEF2DPATHWAY"
[58] "MITOCHONDRIAPATHWAY" "STAT3PATHWAY"        "EPONFKBPATHWAY"
[61] "NKCELLSPATHWAY"      "MONOCYTEPATHWAY"     "CARM_ERPATHWAY"
[64] "HCMVPATHWAY"

> plot(x=eta1,y=eta2,xlim=c(0,max(eta1,eta2)),ylim=c(0,max(eta1,eta2)),xlab="normal",ylab="disease",
      main="(b) EVA: Dysregulated pathways")

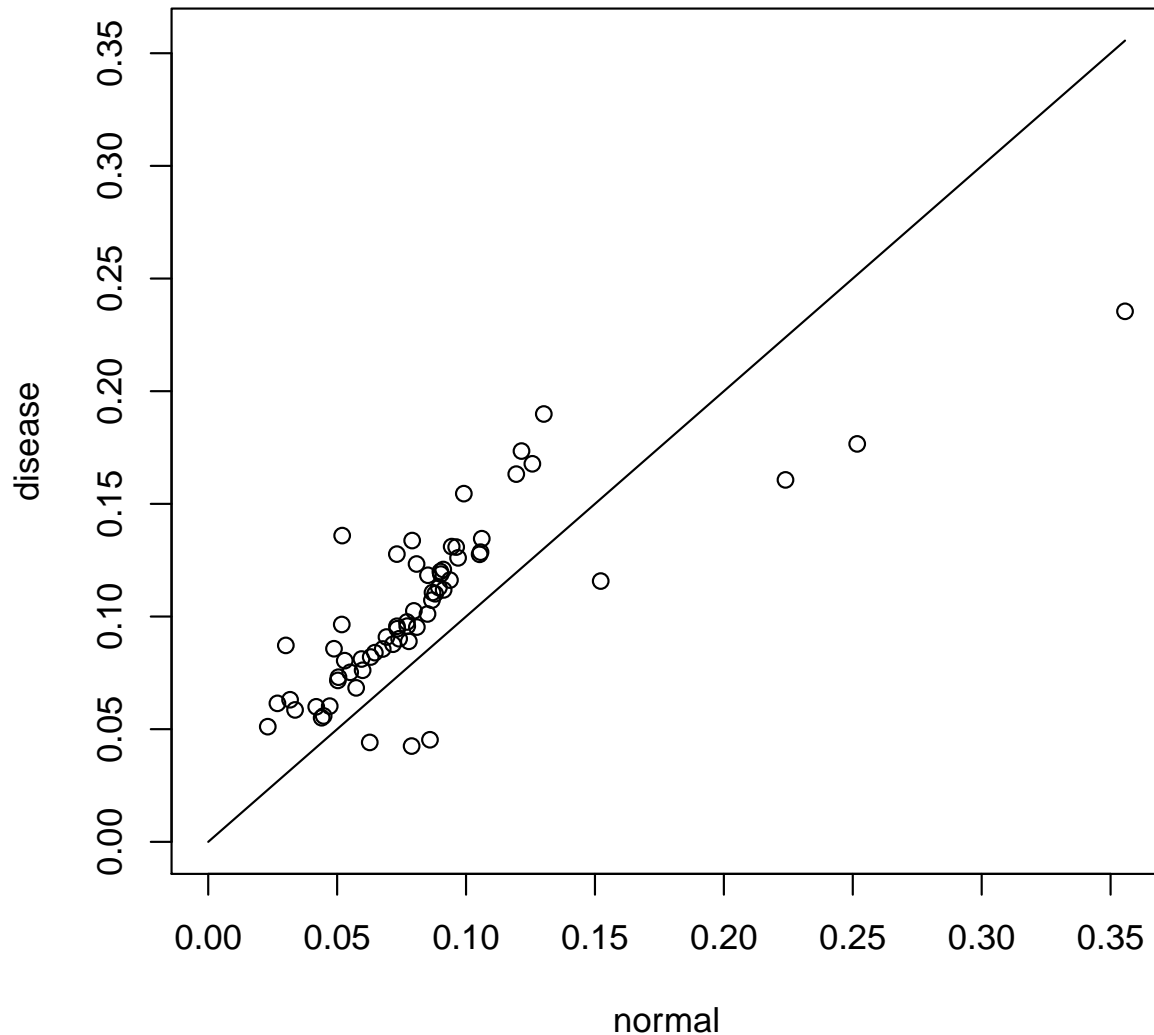
NULL

> lines(x=c(0,max(eta1,eta2)),y=c(0,max(eta1,eta2)))

NULL

```

### (b) EVA: Dysregulated pathways



Although there is discrepancy in identified dysregulated pathways ( $p\text{-value} < 0.05$ ), the general trend found in [1] holds still true. The trend is that usually the dysregulated pathways have higher variability measure in more dangerous phenotypes. The figures reveal that both DIRAC and EVA have this property. DIRAC found 48 dysregulated pathways and EVA discovered 64 pathways, 45 pathways showed up in both analysis, and 67 pathways were discovered totally.

```
> print(significantPathwaysGSV)
```

[1] "DEATHPATHWAY"	"TCAPOPTOSISPATHWAY"	"NEUTROPHILPATHWAY"
[4] "PGC1APATHWAY"	"TERCPATHWAY"	"RARRXRPATHWAY"
[7] "SKP2E2FPATHWAY"	"KERATINOCYTEPATHWAY"	"CHEMICALPATHWAY"
[10] "METHIONINEPATHWAY"	"TGFBPATHWAY"	"PS1PATHWAY"
[13] "PROTEASOMEPATHWAY"	"CDK5PATHWAY"	"MAPKPATHWAY"
[16] "NTHIPATHWAY"	"PDGFPATHWAY"	"BIOPEPTIDESPATHWAY"
[19] "SPPAPATHWAY"	"PYK2PATHWAY"	"CDC42RACPATHWAY"
[22] "MYOSINPATHWAY"	"BETAOXIDATIONPATHWAY"	"IL7PATHWAY"
[25] "FMLPPATHWAY"	"FASPATHWAY"	"VITCBPATHWAY"
[28] "CD40PATHWAY"	"IGF1PATHWAY"	"CDC25PATHWAY"
[31] "MTORPATHWAY"	"RNAPATHWAY"	"FBW7PATHWAY"
[34] "LYMPHOCYTEPATHWAY"	"LAIRPATHWAY"	"HIVNEFPATHWAY"
[37] "ALKPATHWAY"	"PEPIPATHWAY"	"MSPPATHWAY"
[40] "EDG1PATHWAY"	"GSK3PATHWAY"	"RELAPATHWAY"
[43] "METPATHWAY"	"TNFR2PATHWAY"	"AT1RPATHWAY"
[46] "ATRBRCAPATHWAY"	"GLYCOLYSISPATHWAY"	"TIDPATHWAY"
[49] "EOPATHWAY"	"WNTPATHWAY"	"ARAPPATHWAY"
[52] "MRPPATHWAY"	"P53HYPOXIAPATHWAY"	"PITX2PATHWAY"
[55] "IL18PATHWAY"	"STRESSPATHWAY"	"MEF2DPATHWAY"
[58] "MITOCHONDRIAPATHWAY"	"STAT3PATHWAY"	"EPONFKBPATHWAY"
[61] "NKCELLSPATHWAY"	"MONOCYTEPATHWAY"	"CARM_ERPATHWAY"
[64] "HCMVPATHWAY"		

> print(significantPathwaysDIRAC)

[1] "DEATHPATHWAY"	"NEUTROPHILPATHWAY"	"PGC1APATHWAY"
[4] "RARRXRPATHWAY"	"SKP2E2FPATHWAY"	"KERATINOCYTEPATHWAY"
[7] "CHEMICALPATHWAY"	"TGFBPATHWAY"	"PROTEASOMEPATHWAY"
[10] "MAPKPATHWAY"	"PDGFPATHWAY"	"BIOPEPTIDESPATHWAY"
[13] "SPPAPATHWAY"	"PYK2PATHWAY"	"MYOSINPATHWAY"
[16] "BETAOXIDATIONPATHWAY"	"IL7PATHWAY"	"FMLPPATHWAY"
[19] "VITCBPATHWAY"	"CD40PATHWAY"	"CDC25PATHWAY"
[22] "MTORPATHWAY"	"RNAPATHWAY"	"FBW7PATHWAY"
[25] "LYMPHOCYTEPATHWAY"	"LAIRPATHWAY"	"HIVNEFPATHWAY"
[28] "ALKPATHWAY"	"P35ALZHEIMERSPATHWAY"	"MSPPATHWAY"
[31] "GSK3PATHWAY"	"RELAPATHWAY"	"METPATHWAY"
[34] "TNFR2PATHWAY"	"AT1RPATHWAY"	"FREEPATHWAY"
[37] "ARAPPATHWAY"	"MRPPATHWAY"	"P53HYPOXIAPATHWAY"
[40] "IL18PATHWAY"	"STRESSPATHWAY"	"MEF2DPATHWAY"
[43] "STAT3PATHWAY"	"HSP27PATHWAY"	"EPONFKBPATHWAY"
[46] "NKCELLSPATHWAY"	"MONOCYTEPATHWAY"	"CARM_ERPATHWAY"

## 4 System Information

Session information:

```
> toLatex(sessionInfo())
```

- R version 3.3.1 (2016-06-21), x86\_64-w64-mingw32
- Locale: LC\_COLLATE=C, LC\_CTYPE=English\_United States.1252, LC\_MONETARY=English\_United States.1252, LC\_NUMERIC=C, LC\_TIME=English\_United States.1252
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: GSBenchmark 0.107.0, GSReg 1.8.0
- Loaded via a namespace (and not attached): tools 3.3.1

## 5 Literature Cited

### References

- [1] James A Eddy, Leroy Hood, Nathan D Price, and Donald Geman. Identifying tightly regulated and variably expressed networks by differential rank conservation (dirac). *PLoS computational biology*, 6(5):e1000792, 2010.
- [2] Bahman Afsari. *Modeling cancer phenotypes with order statistics of transcript data*. PhD thesis, Johns Hopkins University, 2013.
- [3] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 1938.