文献背景[^Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges]

pathway analysis已经成为差异表达基因和蛋白探究潜在生物学意义的首选方式，该方法减少了复杂性的同时增加了解释能力。

在功能水平对多高通量的molecular measurements的分析具有以下作用，首先，将数千个基因，蛋白或着分子根据它们所有涉及到pathways群分，这样减少到数百个pathways能够减少分析的复杂性；其次，在不同的实验条件下识别具有差异的pathways相比简单的一列差异基因或则蛋白，更有解释意义。

pathway analysis已经应用于Gene Ontology(GO, also referred to as a 'gene' set), physical interaction networks(protein-protein interactions), kinetic simulation of pathways, steady-state pathway analysis(flux-balance analysis), and in the inference of pathways from expression and sequence data.

这里介绍的pathway analysis具体指在当前公共的数据库的基础上例如GO, KEGG探索pathway信息。也就是，knowledge base-driven pathway analysis。

Firste Generation: Over-Representation Analysis (ORA) Approaches

使用统计学知识评估差异表达基因中的涉及到特殊pathway的一部分基因。也值得是"2 x 2" table method。首先，对输入基因列表使用明确的阈值或标准，例如选择FDR为0.05时的过表达或低表达的基因；然后，计算差异基因所设计的pathway(例如，计算microarray上所有的基因)；最后，在输入基因列表基础上，计算每一个pathway的过表达或低表达。

最常用的统计方法为，hypergeometric，chi-square，或binomial distribution。

ORA tools		
Onto-Express	Web (http://vortex.cs.wayne.edu)	[4,5]
GenMAPP	Standalone (http://www.genmapp.org)	[11,71]
GoMiner	Standalone, Web (http://discover.nci.nih.gov/gominer)	[72,73]
FatiGO	Web (http://babelomics.bioinfo.cipf.es)	[74]
GOstat	Web (http://gostat.wehi.edu.au)	[7]
FuncAssociate	Web (http://llama.mshri.on.ca/funcassociate/)	[6]
GOToolBox	Web (http://genome.crg.es/GOToolBox/)	[10]
GeneMerge	Standalone, Web (http://genemerge.cbcb.umd.edu/)	[9]
GOEAST	Web (http://omicslab.genetics.ac.cn/GOEAST/)	[75]
ClueGO	Standalone (http://www.ici.upmc.fr/cluego/)	[76]
FunSpec	Web (http://funspec.med.utoronto.ca/)	[77]
GARBAN	Web	[78]
GO:TermFinder	Standalone (http://search.cpan.org/dist/GO-TermFinder/)	[8]
WebGestalt	Web (http://bioinfo.vanderbilt.edu/webgestalt/)	[79]
agriGO	Web (http://bioinfo.cau.edu.cn/agriGO/)	[80]
GOFFA	Standalone, Web (http://edkb.fda.gov/webstart/arraytrack/)	[81]
WEGO	Web (http://wego.genomics.org.cn/cgi-bin/wego/index.pl)	[82]

具有以下局限，首先，ORA使用的不同统计方法(hypergeometic distribution, binomial distribution, chi-square distribution, etc)，都是独立的检测改变，仅仅考虑的基因的数目而没有考虑与之相关的值，例如探针的密度，因此ORA平等对待每一个基因。但是regulation的范围(fold-change, significance of a change, etc)在分配输入基因比重方面是有用的，同样还有所涉及的pathway；其次，ORA检测过程中仅仅使用了最显著的基因而忽略了其他基因，可能哪些稍微高于fold阈值或者PDR阈值的基因就会丢弃，这就可能导致信息丢失。Breitling et al. addressed this problem by proposing an ORA method for avoiding thresholds. It uses an interactive approach that adds one gene at a time

to find a set of genes for which a pathway is most significant；再次，ORA假设每个基因都是独立的，平等看待每一基因，但是生物是一个复杂的设计到基因产物交互作用的过程，基因表达分析的一个目的可能就在于获得基因产物之间如何地相互作用；最后，ORA假定每个pathway都是独立存在，这也忽略了pathways之间的相互作用问题。

Second Generation: Functional Class Scoring(FCS) Approaches

Functional class scoring(FCS) 假设基因大的改变会对pathways产生显著效果，同时较弱且功能相关的一套基因步调一致的改变也会产生显著效果。首先，计算来molecular measurements的gene-level的统计分析，这涉及到计算差异表达的个体基因和蛋白，当前用于gene-level的统计包含具有phenotype，ANOVA，Q-statistic， signal-to-noise ratio的molecular measurements的相关性；其次，所有基因gene-level的统计分析汇聚成单个pathway-level的统计分析，该统计分析可以是多变量，计算基因间的独立性，也可以是单变量的，不考虑基因间的的相互依靠性。不考虑使用的不同统计方式，pathway-level统计能力依赖于一个pathway中差异表达基因的部分，也就是pathway的大小。Although, multivariate statistics are expected to have higher statistical power, univariate statistics show more power at stringent cutoffs when applied to real biological data(p~0.001), and equal power as multivariate statistics at less stringent cutoffs(p~0.05)；最后，评估pathway-level的统计显著性。当在计算统计显著性时，当前使用的null假设可分为两类：1）competitive null hypothesis 和 2）self-contained null hypothesis。self-contained null hypothesis 变换每个样本的类别标签(i.e., phenotypes)，检测给定pathway中一套基因本身的显著性，不考虑不在其中的基因；competitive null hypothesis针对每个pathway变换基因标签，比较一套存在于pathway中的基因和一套不在pathway中的基因。

FCS tools		
GSEA	Standalone (http://www.broadinstitute.org/gsea/)	[21,29]
sigPathway	Standalone (BioConductor)	[22]
Category	Standalone (BioConductor)	[24]
SAFE	Standalone (BioConductor)	[30]
GlobalTest	Standalone (BioConductor)	[15]
PCOT2	Standalone (BioConductor)	[17]
SAM-GS	Standalone (http://www.ualberta.ca/~yyasui/software.html)	[83]
Catmap	Standalone (http://bioinfo.thep.lu.se/catmap.html)	[84]
T-profiler	Web (http://www.t-profiler.org)	[85]
FunCluster	Standalone (http://corneliu.henegar.info/FunCluster.htm)	[86]
GeneTrail	Web (http://genetrail.bioinf.uni-sb.de)	[87]
GAzer	Web	[88]

FCS方法解决了ORA三个局限性，首先，不需要一个阈值将表达数据任意的分成显著性和非显著性两类，而是使用所有可以得到的molecular measurements来分析pathway；其次，ORA在识别显著性pathway时完成忽略了molecular measurements，FCS使用该信息来检测同一个pathway内一致性改变的基因；最后，通过基因表达一致性的改变，FCS能够实现pathway内基因间的依赖性，而ORA不行。

FCS局限性，首先，FCS类似ORA，独立地分析每一个pathway，pathway同基因一样，可以互相作用重叠的。当使用GO分析时，一个pathway的可能受到影响是因为与其他显著影响pathway存在基因重叠；其次，许多FCS方式针对给定pathway，使用基因表达的改变来对基因求秩序，同时对与进一步的分析不考虑其基因表达的改变信息。例如，2和20倍的改变可以是相同的秩序，虽然20倍的改变应该赋予基因更大的比重。Importantly, however, considering only the ranks of genes is also advantageous, as it is more robust to outliers. A notable exception to this scenario is approaches that use gene-level statistics (e.g., t-statistic) to compute pathway-level scores. For example, an FCS method that computes a pathway-level statistic as a sum or mean of the gene-level statistic accounts for a relative difference in measurements (e.g, Category, SAFE)。

Third Generation: Pathway Topology (PT)-Based Approaches

基于大量可用的公开pathway的信息，而不是每个pathway简单的基因信息，进行pathway分析。不同于GO和Molecular Signatures Database (MSigDB)，这些数据库也提供了给定pathway内基因产物相互作用的信息(e.g., activation, inhibition, etc.)，以及作用位置(e.g., cytoplasm, nucleus, etc.)。它们包括KEGG, MetaCyc, Reactome, RegulonDB, STKE, BioCarta和pantherDB。

PT-based tools		
ScorePAGE	No implementation available	[37]
Pathway-Express	Web (http://vortex.cs.wayne.edu)	[38,39]
SPIA	Standalone (BioConductor)	[40]
NetGSA	No implementation available	[43]

ORA和FCS方法仅考虑pathway中基因数据，或着共表达基因来识别显著性pathway，而不考虑这些数据库额外的可行信息。Pathway topoloty (PT)-based methods，能使用这些额外的信息。PT-based 方法同FCS一样执行相同的三步分析处理。其关键差异在于，PT-based方法使用pathway topology来计算gene-level统计结果。

Rahnenfuhrer et al. 提出了ScorePAGE，计算pathway内每个基因对之间的相似性(e.g., correlation, covariance， etc.)。每个基因对之间的相似性测量类似于FCS方法中gene-level的统计，它平均地检测pathway-level的分值。然而，不同于给予所有成对相似性一样的权重，ScorePAGE根据指定pathway内成对基因发生联系所需的反应数目来划分成对相似性。尽管该过程被设计用于分析代谢pathways，理论上也适用于signaling pathways。

最近impact factor (IF)分析过程被提出分析signaling pathways。 IF通过合并一些重要的biological factors，包括基因表达差异，相互作用类型，pathway中基因的位置，来考虑整个pathway的结构和动态性。IF分析将signaling pathway模拟为图像，nodes代表基因，edges代表基因间的相互作用。进一步，定义gene-level的统计，称为一个基因的pertubation factor(PF)，为表达过程中所测量到的改变之和，和和pathway中所有gene的PF的线性方程(Further, it defines a gene-level statistic, called perturbation factor of a gene, as a sum of its measured change in expression and a linear function of the perturbation factors of all genes in a pathway)。由于每个基因的PF都是被线性公式所定义，因此整个pathway都被定义为一个线性系统。

FCS方式使用基因间的相关性，模糊地假设潜在网络，由于是被结构的相关性所定义，因此并不会随着实验条件改变而改变。

NetGSA, that accounts for the change in correlation as well as the change in network structure as experimental conditions change. Their approach, like IF analysis, models gene expression as a linear function of other genes in the network. However, it differs from IF in two aspects. First, it accounts for a gene's baseline expression by representating it as a latent variable in the model. Second, it requires that the pathways be represented as directed acyclic graphs (DAGs). If a pathway contains cycles, NetGSA requires additional latent variables affecting the nodes in the cycles. In contrast, IF analysis does not impose any constraint on the structure of a pathway.

尽管PT-based 方法难以推广，也存在多个共有不足。首先，真实的pathway topology是建立在细胞类型之上，是由于细胞特异的基因的表达轮廓和被研究的条件。然而，该信息几乎不可能获得，且在当前数据基础上被片段化了；其次，无法模拟系统的动态状态，由于较弱的pathway之间的联系无法考虑它们之间的相互作用(One obvious problem is that true pathway topology is dependent on the type of cell due to cell-specific gene expression profiles and condition being studied; the inability to model dynamic states of a system and the inability to consider interactions between pathways due to weak interpathway links to account for interdependence between pathways)。

Outstanding Challenges in Pathway Analysis

当前pathway anaysis的两个挑战可分为两类：1）annotation challenges；2）methodological challenges。

---

clusterProfiler

1. Terminology

Gene sets：单纯的一列功能相关的基因，不需要指定基因之间的关系

Pathways：可以解释为特殊的一列基因，典型地代表一组在生物功能上共同作用的基因

Gene Ontology：定义了用于描述基因功能的概念或者类别，以及这些概念和类别之间的关系。其定义的功能有以下三个方面：

MF：分子功能，基因产物的分子作用

CC：细胞组成，基因产物作用的地点

BP：生物过程，pathways以及larger processes由多重基因产物的作用所构成

GO terms由带有带有方向的非循环的图形所组成，terms之间的edge代表着parent-child 关系。

KEGG：KEGG是手动绘制的，代表了分子之间的协同作用和互相作用的网络图的集合。这些pathways涵盖了大量的生物过程，可以划分为7个大的范畴：metabolism，genetic and environmental information processing, cellular processes, organismal systems, human diseases, and drug development。

GO和KEGG是最常用的功能分析。由于它们长期的收录支持以及包含了广泛的物种，因此是首要的分析选择。其他相关的gene sets有Disease Ontology(DO)，Disease Gene Network(DisGeNET), wikiPathways, Molecular Signature Database(MSigDb)。

- If your input gene id contains duplicated IDs, those duplicated will be removed.
- Those genes that do not have GO annotation will be removed.

2. Functional Enrichment Analysis Methods

ORA(Over Representation Analysis) 广泛用于判断，在一个已知的实验所驱动的基因列中(a list of differentially expressed genes, DEGs)，已知的生物功能或过程是否over-represented(= enriched)。

P-value 采用超几何分布计算：

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

例如：Microarry study， 11812个基因用于差异表达分析，2个样本中，有202个基因发现DE，在这些DE中，有25个基因注释到了特殊的功能gene set中，该gene set包含了262个基因。则2x2四格表：

```
d <- matrix(c(25, 237, 177, 11373), nrow=2, dimnames=list(c("DE","Not DE"), c("In GS", "Not in GS")))
```

`fisher.test(d, alternative="greater")`

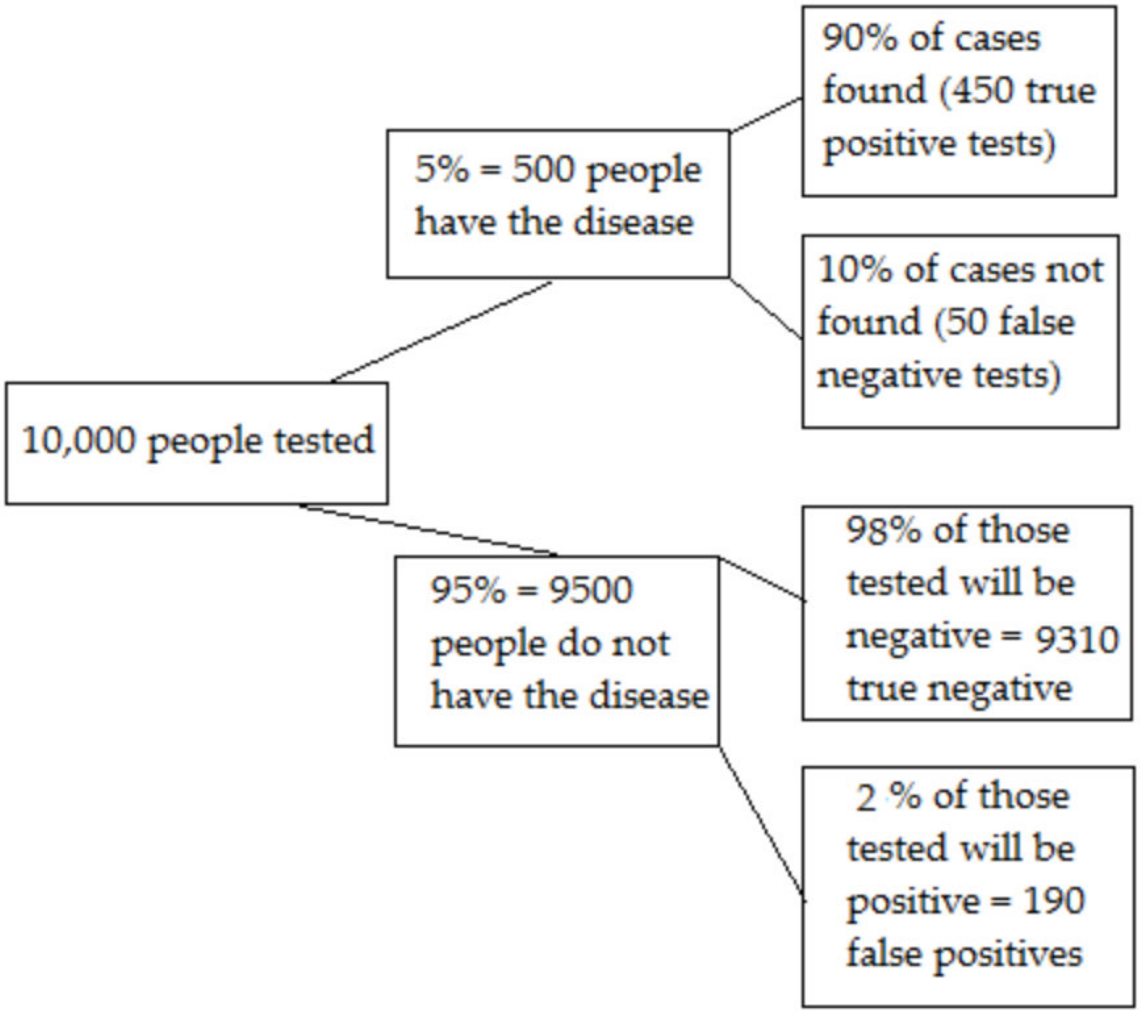
```
> d
       In GS Not in GS
DE        25       177
Not DE   237     11373
> fisher.test(d,alternative="greater")

        Fisher's Exact Test for Count Data

data:  d
p-value = 2.57e-12
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 4.5227    Inf
sample estimates:
odds ratio
  6.774445
```

95%置信区间为4.5，而算得的率为6.774，远大于4.5，接受DE和Not DE之间存在差异。

Adjustment for Multiple Hypothesis Testing

当整个gene set都被评估后，DOSE针对多重假设检验调整评估的显著性水平，同时针对FDR control 计算q-value。

The false discovery rate (FDR)指的是type I 型错误的期待比率。Type I型错误就是错误拒绝null假设；换言之，就是得到了假阳性。FDR就是假阳性的数目在所有拒绝假设中的比率(拒绝 null hypothesis)。医学上，就是当你得到一个阳性检测结果但实际上并没有得病的比率，其对应面就是阳性预测值(PPV)，就是阳性结果准确性的比率。



p-value告诉我们单次检测假阳性的概率；当针对小样本进行大量数目的检测时(genomics或protoemics), 就应使用q-value了：p-value为5%意味着5%的检测将会导致加假阳性结果，q-value为5%意味着5%的显著性结果将会为假阳性。使用q-value来控制FDR的过程称为Benjamini-Hochberg procedure。

3. Universal enrichment analysis

clusterProfiler支持超几何检验和ontology/pathway的gene set enrichment分析。clusterProfiler针对超几何检验提供enricher函数，针对gene set enrichment分析提供GSEA函数，同时接受自定义的注释。这两个函数接受2个额外的参数，TERM2GENE，TERM2NAME。TERM2GENE为一个数据框，第一列为term ID，第二列为对应的比对的基因；TERM2NAME同样为一数据框，第一列为term ID，第二列为对应的term 名称，且TERM2NAME是可选的。

4. Gene Ontology Analysis

GO分析(groupGO(), enrichGO(), gseGO())支持具有OrgDb对象的物种。Bioconductor 现已经提供了约20种物种饿OrgDb数据，当然也可以使用AnnotationForge创建Orgdb数据，参考GOSemSim。

假如拥有GO注释数据(数据框格式，第一列为gene ID，第二列为GO ID)，可使用enricher(), gseGO()函数执行over-representation test和gene set enrichement analysis。

```
library(AnnotationForge)
makeOrgPackage(go=fGO,
               version="0.1",
               maintainer = "carlos <carlos@google.com>",
               author="carlos <carlos@google.com>",
               outputDir = ".",
               tax_id="58729",
               genus="Taeniopygia",
               species = "guttata",
               goTable = "go",verbose=T)

install.packages("./org.Tguttata.eg.db",repos=NULL,type="source")
```

假如基因是通过直接注释(direction annotation)而来的，那么它们也应该被它们ancestor GO nodes所注释(indirect annotation)。假如只有direct annotaion，可将它们的注释传递给buildGOmap函数，该函数将会推导indirection annotation，生成使用与enricher()和gseGO()的数据框。

Over-representation test

`ego <- enrichGO(gene=gene, universe=names(geneList), OrgDb=org.Hs.eg.db, ont="CC", pAdjustMethod="BH", pvalueCutoff=0.01, qvalueCutoff=0.05, readable=T)`

`ego2 <- enrichGO(gene=gene.df$ENSEMBL, OrgDb=org.Hs.eg.db, keyType="ENSEMBL", ont="CC", pAdjustMethod="BH", pvalueCutoff=0.01, qvalueCutoff=0.05)`

`ego2 <- setReadable(ego2, OrgDb=org.Hs.eg.db)`

Drop specific GO terms or level

dropGO函数可用于去除enrichGO和compareCluster结果中的特殊GOterms或GO levels。

Test GO at sepcific level

enrichGO不包含参数用于限制特殊GO level的test，但是可使用gofilter函数来限制结果到指定的GO level，用于enrichGO和compareCluster的输出的结果。

Reduce redundancy of enriched GO terms

GO是一个parent-child的结构形式，因此parent term可能会重叠它所有child terms的很大一部分，这会导致冗杂的发现输出。clusterProfiler提供了simplify函数用于减少来自enrichGO和gseGO的输出中的冗杂GO terms。通过计算GO terms之间的相似度，然后去除哪些高度相似的terms，仅保留一个代表性的term。

GO analsysi for non-model organisms

enrichGO和gseGO函数均需要OrgDb数据，若AnnotatonHub没有该物种的OrgDb数据，可以从其他地方获得OrgDb数据，例如，biomaRt和Blast2GO。然后使用enricher或GSEA函数分析。或者，使用AnnotationForge来创建OrgDb数据。

5. KEGG analysis

KEGG.db自从2012年后就没更新了，在clusterProfiler包中，enrichKEGG(for KEGG pathway)和enrichMKEGG(for KEGG module)支持下载最新的KEGG在线版本用于富集分析。

使用search_kegg_organism函数搜索支持的物种

```
library(clusterProfiler)
```

```
kpc <- search_kegg_organism("Klebsiella pneumoniae", by="scientific_name")
```

KEGG over-representation test

```
data(geneList, package="DOSE")
```

```
gene <- names(geneList)[abs[geneList) > 2]
```

```
kk <- enrichKEGG(gene=gene, organism='hsa', pvalueCutoff=0.05)
```

输入的ID类型，可以是kegg，ncbi-geneid，ncbi-proteinid或uniprot。

KEGG Module over-representation test

KEGG Module是人工审核定义的功能单元，在一些情况下，KEGG Module具有明确直接的解释

```
mkk <- enrichMKEE(gene=gene, organism='hsa')
```

6. Visualization of Functional Enrichment Result

enrichplot包提供了多种可视化模型来帮助解释富集结果，它支持来自DOSE, clusterProfiler, ReactomePA和meshes包的富集结果，ORA和GSEA分析结果都支持。

Bar Plot

条形图是最广泛用于查看enriched terms的方式，它能表现出富集值(p value)，gene count或ratio作为bar的高度和颜色。

```
de <- names(geneList)[abs(geneList) > 2]
```

```
edo <- enrichDGN(de)
```

```
library(enrichplot)
```

```
barplot(edo, showCategory=20)
```

Dot plot

点图类似条形图，能够其他score作为点的大小

```
edo2 <- gseNCG(geneList, nPerm=100000)
```

```
p1 <- dotplot(edo, showCategory=30) + ggtitile("dotplot for ORA")
```

```
p2 <- dotplot(edo2, showCategory=30) + ggtitle("dotplot for GSEA")
```

```
plot_grid(p1,p2,ncol=2)
```

Gene-Concept Network

barplot和dotplot只能显示最显著enriched terms，但时用户可能想要看到哪些基因参与到了这些显著terms。为了考虑到潜在的生物学复杂性，哪些基因可能属于多个注释类别，同时提供多个改变的信息。cnetplot函数可以提取这些基因之间的的复杂关系。cnetplot函数通过网络描绘了基因和生物概念(GO terms或KEGG pathway)之间的联系。

```
edox <- setReadable(edo, "org.Hs.eg.db", "ENTREZID")
```

```
cnetplot(edox, foldChange=geneList)
```

使用pvalue或geneNum对categorySize进行scale

```
cnetplot(edox, categorySize="pvalue", foldChange=geneList)
```

```
cnetplot(edox, foldChange=geneList, circular=TRUE, colorEdge=TRUE)
```

Heatmap-like functional classification

heatplot类似于cnetplot，以heatmap形式展示相关性。

```
heatplot(edox)
```

```
heatplot(edox, foldChange=geneList)
```

Enrichment Map

Enrichment map将每一个term组织到网络结构中，每个edge都连接着重叠的gene sets。共同重叠的gene sets就倾向于聚到一起，容易识别fuctional mudule。

```
emapplot(edo)
```

UpSet Plot

upsetplot是cnetplot的另一种形式，用于展示genes和gene sets之间的复杂关系，它强调不同gene sets之间的gene重叠。