常规的基于超几何分布的富集分析依赖于显著上调或下调的基因，容易遗漏部分差异表达不显著但有重要生物学意义的基因。基因集富集分析（Gene Set Enrichment Analysis, GSEA）不需要指定明确的差异基因阈值，使用预先定义的基因集（通常来自功能注释或先前实验的结果），将基因按照在两类样本中的差异表达程度排序，然后检验预先设定的基因集合是否在这个排序表的顶端或者底端富集。基因集富集分析检测基因集合而不是单个基因的表达变化，因此可以包含这些细微的表达变化，预期得到更为理想的结果。GSEA 仅是对背景基因集的表达趋势分析，并不涉及功能注释，相应基因的功能注释需要去对应富集结果（eg：GO，KEGG 等富集结果）中查找。

GSEA 分析包括 5 个步骤：

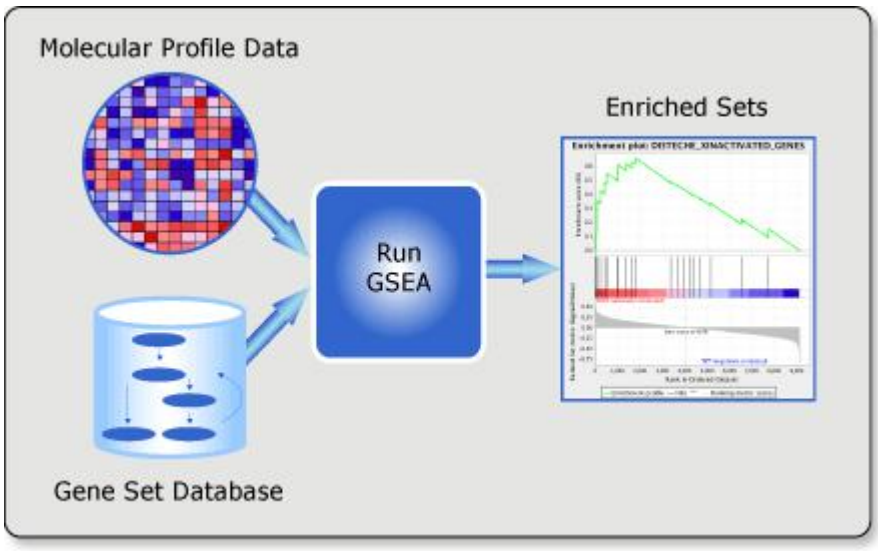
Step1：根据所有基因的表达数据,计算每个基因在两个分组 ClassA、ClassB 中的差异度（signal2 noise），然后按照在两个表型中的差异度从大到小排序，形成一个排好序的基因列表。

Step2：判断基因集 S 中的基因是否在排序表的顶端或者底端富集。

Step3：计算基因集 S 的 ES（Enrichment Score）富集得分。计算方法：从目标基因列表 L 的第一个基因开始，计算一个累计统计值。当遇到一个落在基因集 S 里面的基因，则增加统计值。遇到一个不在基因集 S 里面的基因，则降低统计值，增量的大小取决于基因与表型的相关性。最高峰为富集得分值（ES）。

step4：计算 ES 的显著性水平（名义 P 值）。利用 empirical phenotype-based permutation test 来计算 ES 的名义 P 值，保留了原始表达数据的复杂相关性。

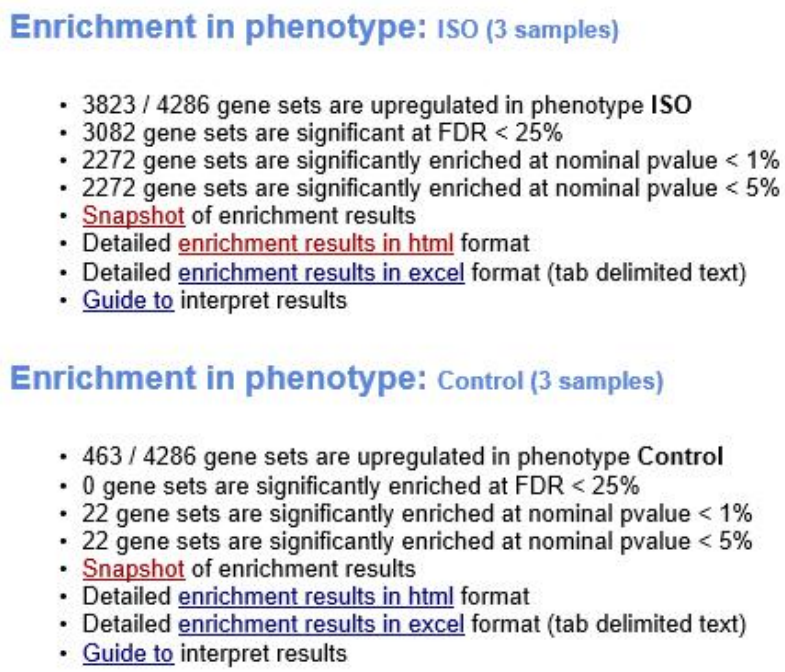
step5：多重假设检验。考虑该基因集的大小，将每个基因集的 ES 标准化，得到标准化的富集分数（NES）;通过计算 false discovery rate (FDR) 值，来控制假阳性率。



3.31 GSEA 富集分析原理图

我们使用本地版 GSEA 分析工具 http://www.broadinstitute.org/gsea/index.jsp，分别对该

物种的 GO、KEGG、Reactome、DO、DisGeNET 等数据集进行 GSEA 分析。。每个比较组合文件中有 index.html 文件，可以看到 GSEA 分析整体情况，列出各组样本基因表达情况，以及显著富集条件，点击超链接会展现详细分析结果。



n/4286 表示：4286 个基因集中，n 个基因集在这一表型中富集；a 个基因集的 FDR 小于 25%； b 个基因集的名义 p 值小于 1%； c 个基因集的名义 p 值小于 5%。

Snapshot of enrichment results 中为基因集的富集得分曲线图。默认情况下，只显示前 20 个基因集的富集得分曲线图。

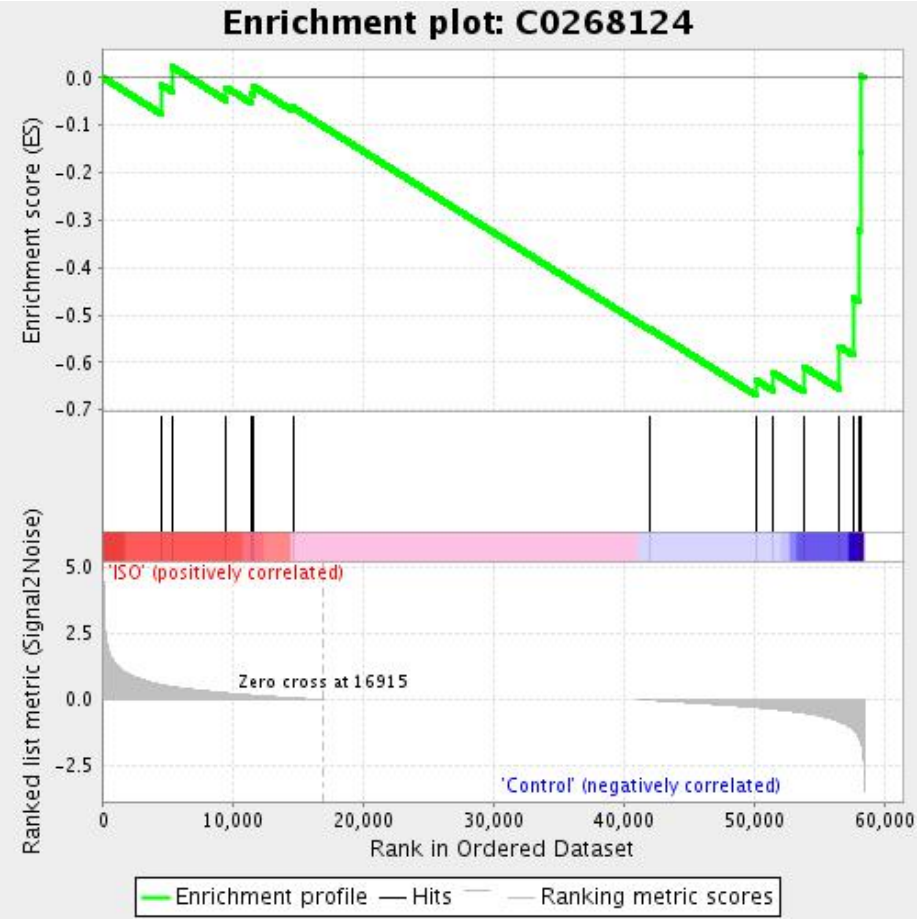
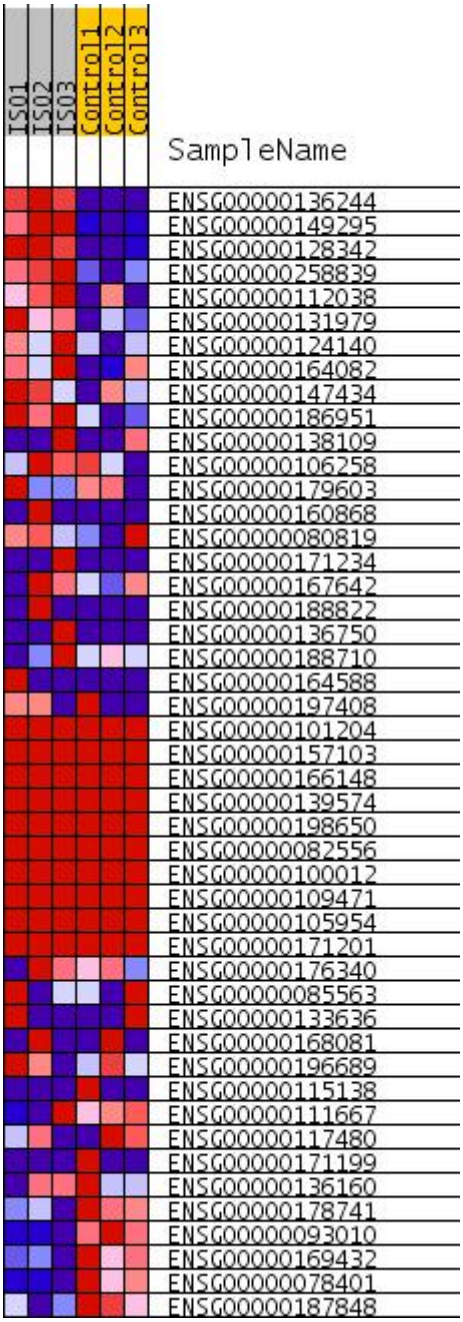


图 3.32 GSEA 富集得分曲线图

上图分成 3 个部分，第一部分为基因 Enrichment Score 的折线图，展示的是 ES 的值计算过程，从左至右每到一个基因，计算出一个 ES 值，连成线。最高峰为富集得分(ES)。在最左侧或最右侧有一个特别明显的峰的基因集通常是感兴趣的基因集。中间部分为 hit，每一条线代表基因集中的一个基因，及其在基因列表中的排序位置。最下面部分为所有基因的 rank 值分布图，展示的是基因与表型关联的矩阵，红色为与第一个表型(ISO)正相关，在 ISO 中表达高，蓝色与第二个表型(Control)正相关，在 Control 中表达高。

另外，还对基因集下的基因在所有样本中表达量的分布以热图的形式进行展示，其中每一列代表一个样本。每一行代表一个基因，基因表达量从低到高，颜色从蓝色过渡到红色。



点击 enrichment results in html，可以在网页查看每个组别所有富集到的基因集，示例如下：

Table: Gene sets enriched in phenotype Control (3 samples) [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	C0268124	Details ...	15	-0.67	-1.39	0.000	1.000	0.957	8342	tags=53%, list=14%, signal=62%
2	C0392607	Details ...	22	-0.50	-1.34	0.000	1.000	1.000	8342	tags=41%, list=14%, signal=48%
3	C0014067	Details ...	15	-0.62	-1.33	0.000	1.000	1.000	9524	tags=40%, list=16%, signal=48%
4	C2931187	Details ...	15	-0.59	-1.32	0.185	1.000	1.000	13450	tags=53%, list=23%, signal=69%
5	C0240479	Details ...	17	-0.51	-1.31	0.000	1.000	1.000	1225	tags=18%, list=2%, signal=18%
6	C1167918	Details ...	30	-0.62	-1.30	0.000	1.000	1.000	13978	tags=60%, list=24%, signal=79%

GS follow link to MSigDB ：该基因集在 GO、KEGG 和 Reactome 数据库中所对应的 ID 号

GS DETAILS：列举了前 20 个基因集的详细信息。

SIZE：基因集中包含的基因个数。

ES：Enrichment Score 富集得分值。

NES：normalized enrichment score，考虑该基因集的大小，将每个基因组的 ES 值标准化，得到标准化的富集分数。

NOM p-val：nominal P value ，名义 p 值，富集分析统计学显著水平

FDR q-val：false discovery rate q-val ，假阳性率 P 值，多重假设检验后得到的富集分析统计学显著水平

FWER p-val：familywise-error rate p-val ，总体错误率 P 值。

RANK AT MAX：最高排名

LEADING EDGE：Tags：在 ES 最大值这个点之前，功能基因集 S 中的成员 s 在目标基因列表中的概率（此时 ES 为正值），这个指标看的是，对 ES 有贡献的基因数目占功能基因集 S 中总数目的百分比。List：在 ES 最大值这个点之前，在排序的目标基因列表 L 中的基因数目（此时 ES 值为正）占 L 总基因数目的百分比,它反映的是，目标基因列表 L 中多少个基因参与了 ES 的计算。Signal 指的是富集信号的强度，它由 Tags 和 List 相加构成。除此之外，在总的 html 页面中，还给出了如下信息

Dataset details

- The dataset has 58395 features (genes)
- No probe set => gene symbol collapsing was requested, so all 58395 features were used

Gene set details

- Gene set size filters (min=15, max=5000) resulted in filtering out 15539 / 19825 gene sets
- The remaining 4286 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the ISO *versus* Control comparison

- The dataset has 58395 features (genes)
- # of markers for phenotype **ISO**: 16915 (29.0%) with correlation area 53.6%
- # of markers for phenotype **Control**: 41480 (71.0%) with correlation area 46.4%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset
- Buttefly plot of significant genes

Global statistics and plots

- Plot of p-values *vs.* NES
- Global ES histogram

Dataset details 给出了基因总数，Gene Set details 给出了基因集的信息。Gene markers 给出了排序之后的基因列表，rank ordered gene list 是富集到的所有基因的具体描述，Heat map and gene list correlation 展示了 100 个与表型相关性高的基因的聚类热图和其与表型的相关性图。Buttefly plot 展示了这 100 个基因与表型的相关性的点图 Global statistics and plots 基因集富集总体数据的点图。Plot of p-values vs. NES 横坐标是矫正后的富集得分，纵坐标分别是假阳性率 P 值和名义 P 值。Global ES histogram 反映所有基因集富集得分情况，横坐标是 ES 富集得分，纵坐标是基因集个数

GSEA 结果解读示例：

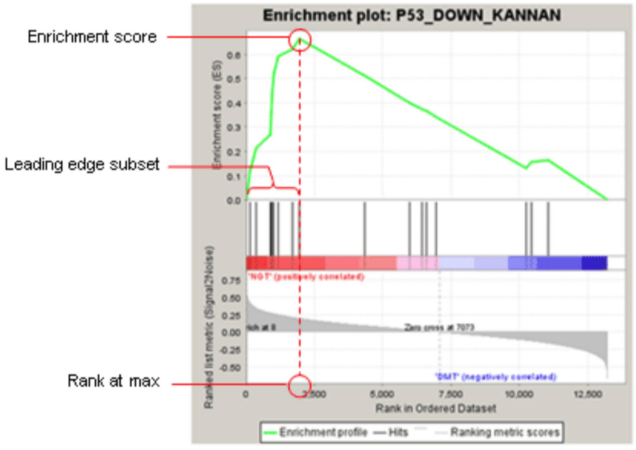


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List