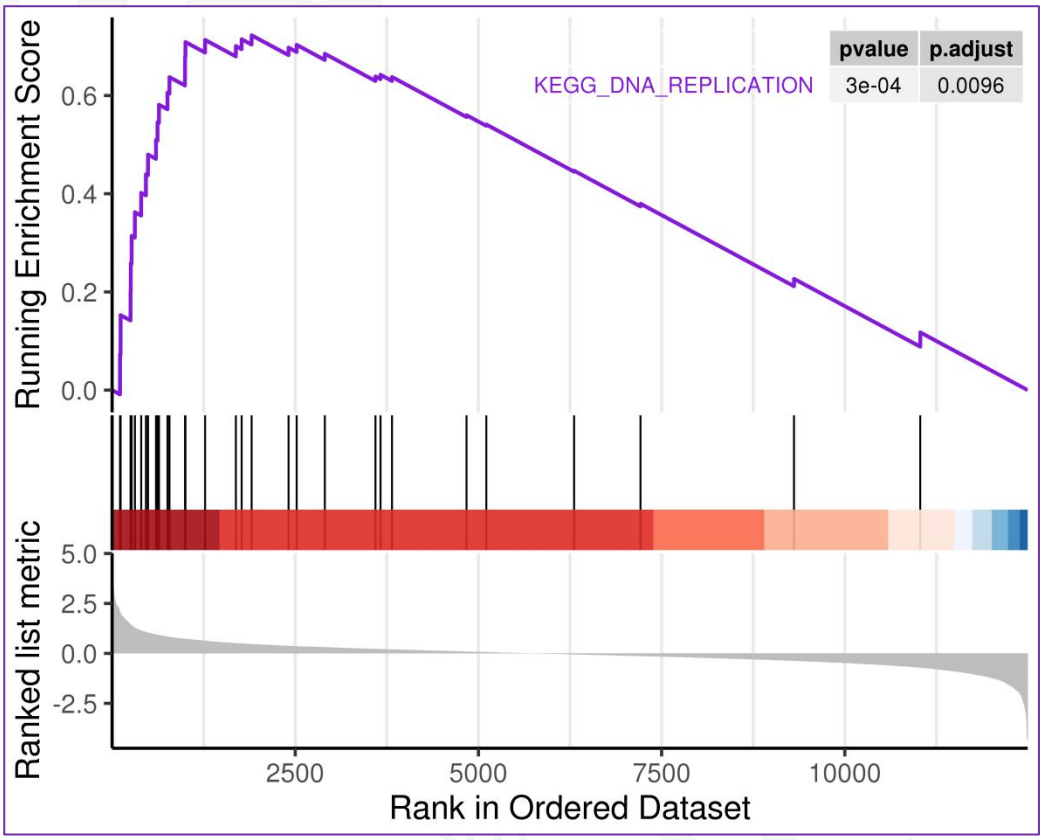
三分钟绘制一张优美的 GSEA 图

云平台官网：http://www.lc-bio.cn/
绘制韦恩图：http://www.lc-bio.cn/overview/37?tools=GSEA_V1.5
撰稿人　吕枫烨　（联系方式：0571-87662413-8059）



GSEA 作为一种富集方法可以有效弥补“根据差异倍数阈值筛选”这种方法的疏漏，**能反映一批基因的微量变化的累积所造成的显著功能差异**，在表达差异不那么显著的情况下，帮我们找到最值得关注的基因。

本文主要为您迅速绘制一张优美的 GSEA 提供步骤指导（3-8 要点）；另附背景知识了解（1-2）和细节优化方法（9 及之后的要点），可后期慢慢研究调整；同时提供相关分析方法说明（见工具中的“分析方法”页），为您深入了解提供支持。

1.引子：了解 GSEA（知识补充，已经熟悉的可跳过）

举个简单的例子，我们有一个肿瘤药物**敏感组** vs **不敏感**对照组。我们可以拿到它的表达谱，我们可以根据差异表达的基因得到一个**基因列表 L（gene list）**。拿到这个基因列表 L，你会怎么处理呢?

按照一般的分析套路来说，我们肯定会关注 **top 基因**，看看表达倍数差异大的基因（上调或者是下调的 top 基因），然后拿到这些基因进行**分析（通路富集分析，画热图等等）**。

做 GO,KEGG 通路富集分析最后得到的结果就是值得关注的 top 基因以及通路，但是**如果表达差异不显著但是可能对生物通路很重要的基因则会漏掉。**

- **取 top 基因分析有什么缺陷吗?**

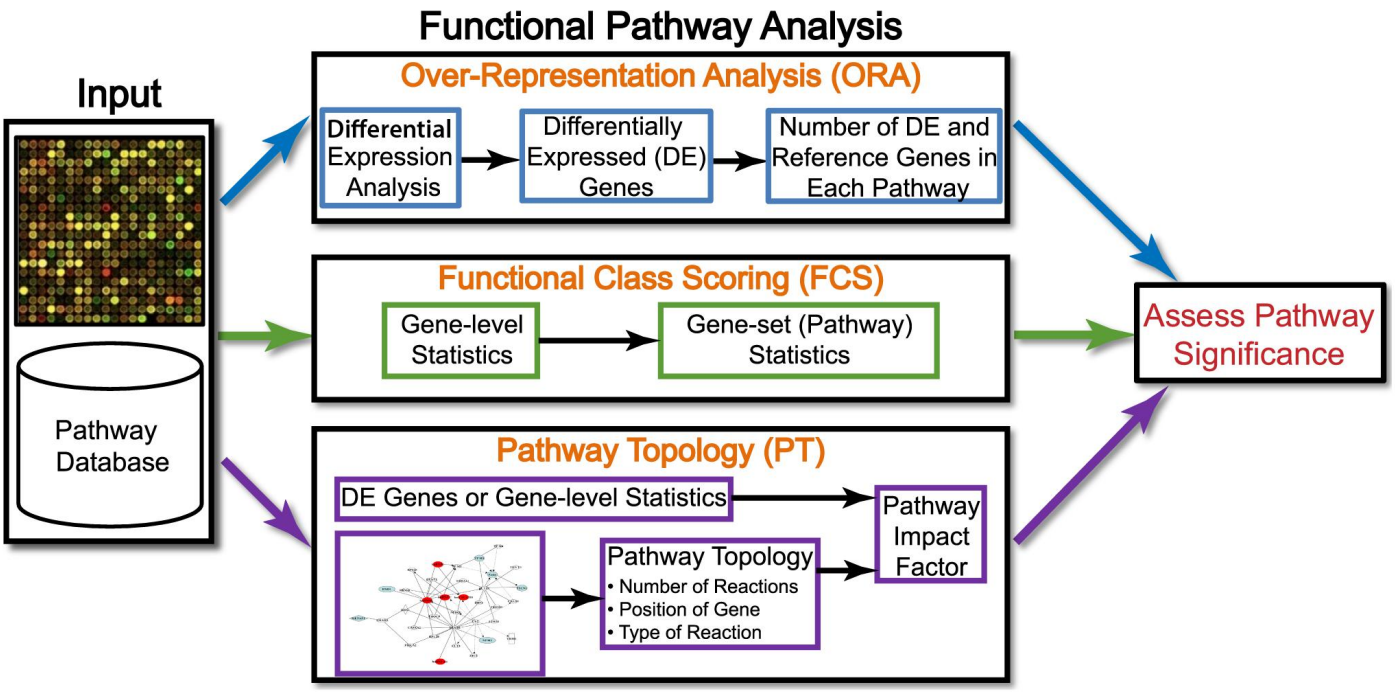
1、经过多重假设检验校正后，**单个基因可能达不到统计学意义的阈值**，因为可能存在背景噪声，相关的生物学差异不大。

2、有些显著性差异表达的基因（具有统计学意义）但是**没有什么生物学意义。**

3、一系列协同作用的基因会影响细胞的生命活动。（与第一条联系起来）**单基因分析可能会遗漏对通路的重要影响。**例如编码代谢途径重要成员的基因上调 20%，可能会显著改变该途径的通量，而且可能比单个基因增加 20 倍更为重要。

4、当研究相同的生物作用通路时，这两项研究中具有统计学意义的基因列表可能会**很少有重叠。**（如肿瘤药物敏感组 VS 不敏感组）

2.了解三种数据分析方法：（知识补充，已经熟悉的可跳过）



参考文献：Khatri P , Sirota M , Butte A J , et al. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges[J]. PLoS Computational Biology, 2012, 8(2):e1002375.

1. ORA：Over-representation analysis

曾经一段时间，micro array 技术的风靡产生了对下游分析的极大需求，over-representation analysis (ORA) 应运而生，从一系列基因里**根据阈值提取出部分基因进行显著性分析**，也就是**统计学常见的“2X2 交叉表格法”，对每个 pathway 进行统计分析**，常见 tests 都建立在 hypergeometric distribution 、chi-square、binomial distribution 等方法上。

- **缺陷：**
  1) 只考虑了差异基因列表，并**不考虑差异基因的表达情况；**
  2) 只检验了通过设定阈值筛选标准的基因，**对差异微弱的基因并未考虑**，但实际上会造成 bias；
  3) 每个基因会作为独立存在事件考虑，**未加入相互影响干扰因素；**
  4) Pathway 也是作为独立存在事件考虑。

我们日常分析中最常见的 GO/KEGG 分析就是基于这种原理，虽然老旧但实用。

2. FCS：Functional Class Scoring Approaches

FCA 的推测设想认为**虽然强烈的单个基因的改变可以影响到 pathways，但是微弱的相互协同的功能相关基因的变化也可以拥有这种影响**，所以这种方法的输入数据是一个**基因水平的统计数据（标准化后食用更佳）**，随后把 gene-level 的数据输入到 pathway-level 进行统计，现有方法包括 Kolmogorov-Smirnov statistic, sum, mean, or median of gene-level statistic, the Wilcoxon rank sum, and the maxmean statistic 等，最后再做一个**显著性检验**。

- **相对于 ORA，FCS 完善了三个缺陷**：

1) **不需要人为的阈值确定差异基因 list;**
2) FCS 使用**所有可用的表达水平**进行分析；
3) FCS 考虑了**基因相互间的变化**，解释了基因变化与 pathway 之间的**依赖性。**

- **缺陷**：

1) 类似于 ORA，pathway 之间的分析依旧是彼此独立的（此种原因可以解释为单个基因同时存有多种功能，在多个 pathway 中发挥作用，**overlap 过多的 pathway 就会相互干扰）；**
2) 使用 rank 的方式纵然有着很多优点，但是**忽略了单个基因的变化幅度**，也就是权重。

本篇所讲述的 GSEA 使用的就是这种原理。

3. PT：Pathway Topology -Based Approaches

因为 ORA 和 FCS 只考虑了基因而未利用额外的数据信息所以天然的存有着分析短板。PT 就是**尝试利用额外的信息进行统筹分析**，但是它其实和 FCS 的分析过程是没有差别的，唯一的区别在于在进行 gene-level statistics 的时候使用 pathway topology 方法。

Rahnenfuhrer et al.推出的 **ScorePAGE**，通过**计算相关和协方差**的方式来得到类似于 FCS 的 pathway-level 的结果，但是又综合考虑了两组 gene list 之间需要 connect 的难度从而进行给分，而不是像 FCS 分配统一权重。

- **缺陷**：
  1) PT-based 的方法千差万别，结论也千差万别，**很难界定结果的准确性；**
  2) 精确的分析结果依赖于数据库的信息准确性，但是**细胞特异性的基因表达数据目前还非常不完善**，这也是卡住方法开发的门槛；
  3) **相关分析无法考虑动态变化**，毕竟生物系统是一个不断协调变化的过程。

综上所述，GSEA 是目前相对较合适的一种寻找值得关注的基因和通路的方法。

参考资料：https://www.jianshu.com/p/be1211dce097

3.了解输入数据格式

点击“示例文件下载”查看（Fig. 1）：

	A	B
1	ID	Expression
2	4312	4.572613
3	8318	4.514594
4	10874	4.418218
5	55143	4.144075
6	55388	3.876258
7	991	3.677857
8	6280	3.501963
9	2305	3.291812
10	9493	3.286223
11	1062	3.219761

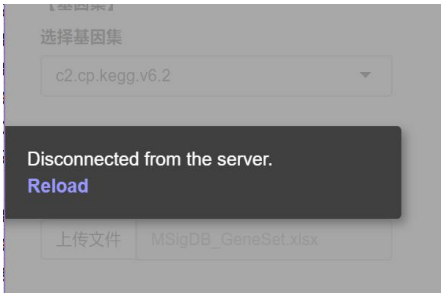


Fig. 1　　Fig. 2　　Fig. 3

- 输入文件必须为 xlsx 格式（txt、xls、csv 等均不行）。
- 第一列为 ID，第二列为表达量（Fig. 2：如前文所述，GSEA 使用的数据是**表达量**，而非 FC 值）。
- 输入数据的 ID **必须**与基因集中的 ID 是同一种类型，否则出现 Fig. 3 的报错。
- 目前只支持两种 ID 类型：**entrez** 和 **symbols**。上述示例数据用的是 entrez ID，它是一串只包含数字的基因编号，在 NCBI 和 KEGG 数据库中均通用。

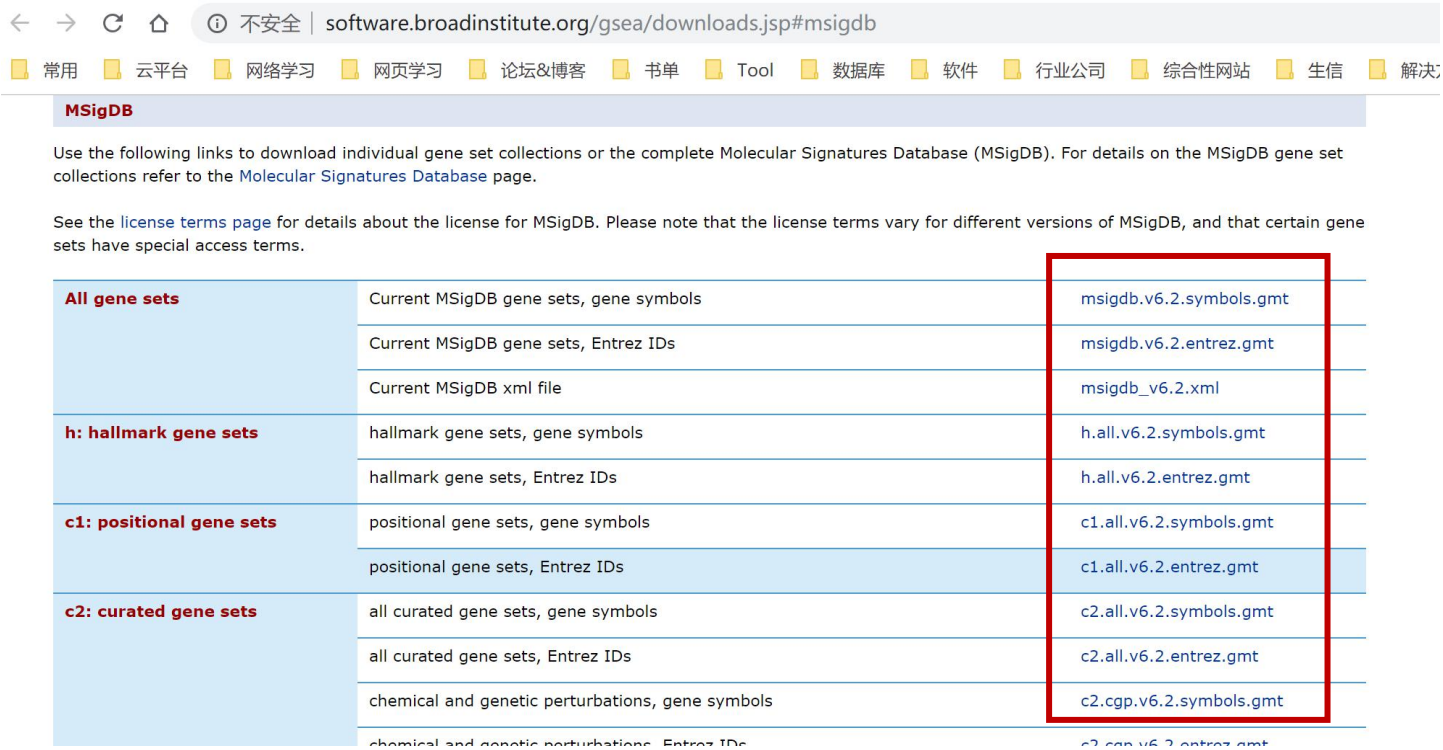
- **上传输入文件前，必须先选择对应的“输入数据的 ID 类型”**，否则分析报错（Fig. 3）。

4.了解基因集

当前分析使用的基因集
当选择另一个基因集时，会立即重新分析

1. 提供的基因集全部来自 GSEA 官方提供的数据库：**MSigDB**。
2. 点击该链接直接跳转至**官方下载路径**（前提是已注册登录该网站，反之，会进入官网的注册登录界面）。
3. 下图即为该链接点开后的界面，点击红框内的条目即可下载对应的数据集。
4. 不同数据集可能是包含关系，如带“all”字样的数据集可能包含该分类下所有其他数据集的内容。**为避免重复分析，使用前请先了解相关数据集的内容。**

1.当选择最后一项“使用自定义的基因集”时，会出现上传文件的功能组件。
2.若需要使用自己的基因集，请按照官方的 gmt 格式整理，格式说明见：http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#GMT:_Gene_Matrix_Transposed_file_format_.28.2A.gmt.29。



5. 关于 Gene Symbols 和 Gene Entrez ID

	A	B	C	D	E	F
1	HALLMARK_	http://ww	3726	2920	467	4792
2	HALLMARK_	http://ww	5230	5163	2632	5211
3	HALLMARK_	http://ww	2224	1595	3422	2222
4	HALLMARK_	http://ww	9181	23332	3832	9493
5	HALLMARK_	http://ww	4609	1499	3714	4851
6	HALLMARK_	http://ww	7046	4092	7040	64750
7	HALLMARK_	http://ww	3566	3572	6772	3554
8	HALLMARK_	http://ww	5437	5430	5436	5434
9	HALLMARK_	http://ww	6790	890	7153	9133

MSigDB 基因集

	A	B
1	ID	Expression
2	4312	4.572613
3	8318	4.514594
4	10874	4.418218
5	55143	4.144075
6	55388	3.876258
7	991	3.677857
8	6280	3.501963
9	2305	3.291812
10	9493	3.286223
11	1062	3.219761

输入数据

如上图，红框区的数据类型需要对应（MSigDB 的 3~最后列<=>输入数据的第一列）。只要两个红框内数据对应，您可以将内容替换成 miRNA，蛋白、代谢等，分析任何组学的数据。区别在于基因集是事先根据了解数据库或者文献而整理好的数据，输入数据是您的实验结果数据。

gene symbols：基因名称。是科研工作者按照基因功能起的名字，通常是对功能描述的缩写。

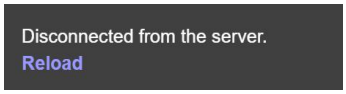
gene entrez id：基因 ID。是一串数字。是 Entrez 对基因的命名方式。

*** 了解 entrez ID ： http://www.360doc.com/content/17/0919/14/19913717_688385068.shtml

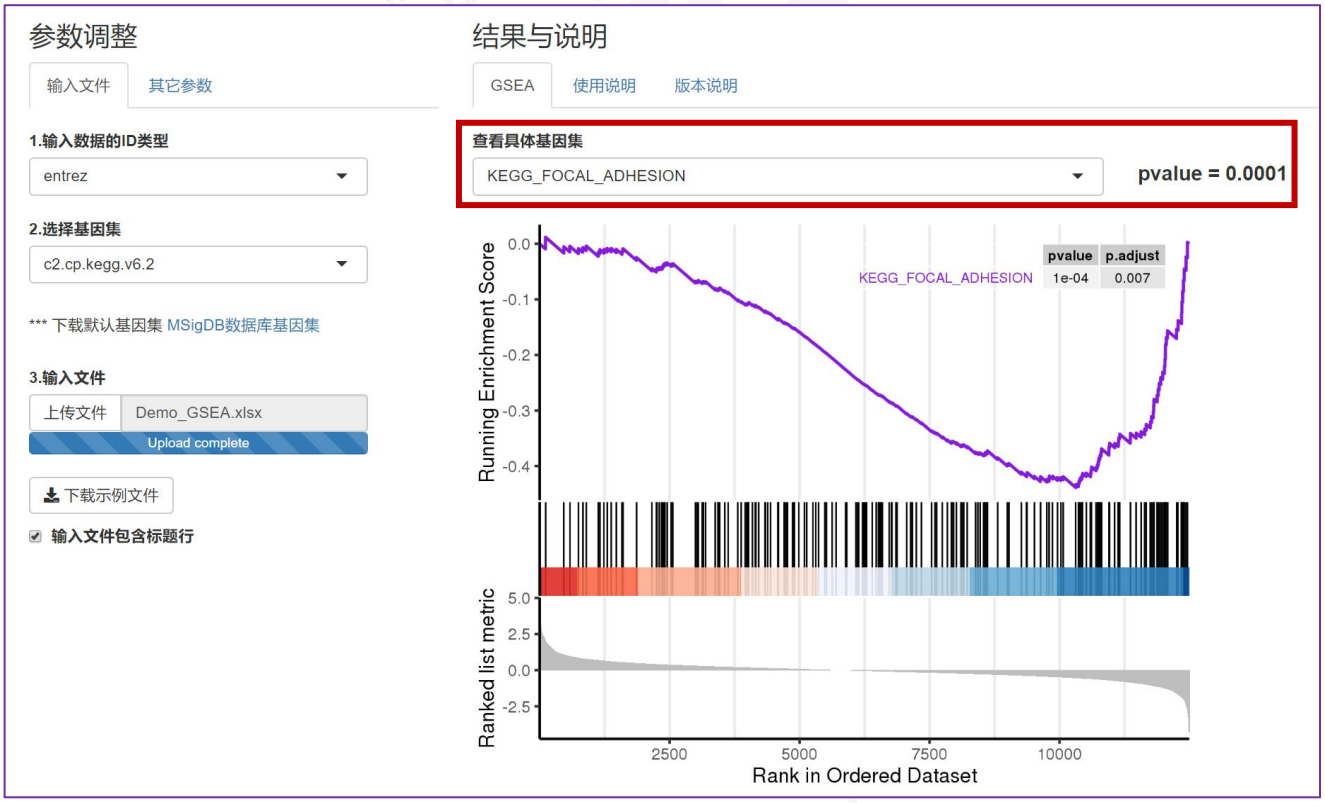
6. 上传数据（以示例数据为例）

按照顺序分别选择 ID 类型、基因集、上传输入文件，上传完成后，自动开始分析，等待十几秒即可出图：

- 在上传数据前**务必确认 ID 类型**，否则上传完成后立即报错退出



- 默认显示 **p 值最小**的一张图。
- “查看具体基因集”处的基因集默认**按 p 值从小到大排列。**
- 右边自动显示所选基因集的 p 值。



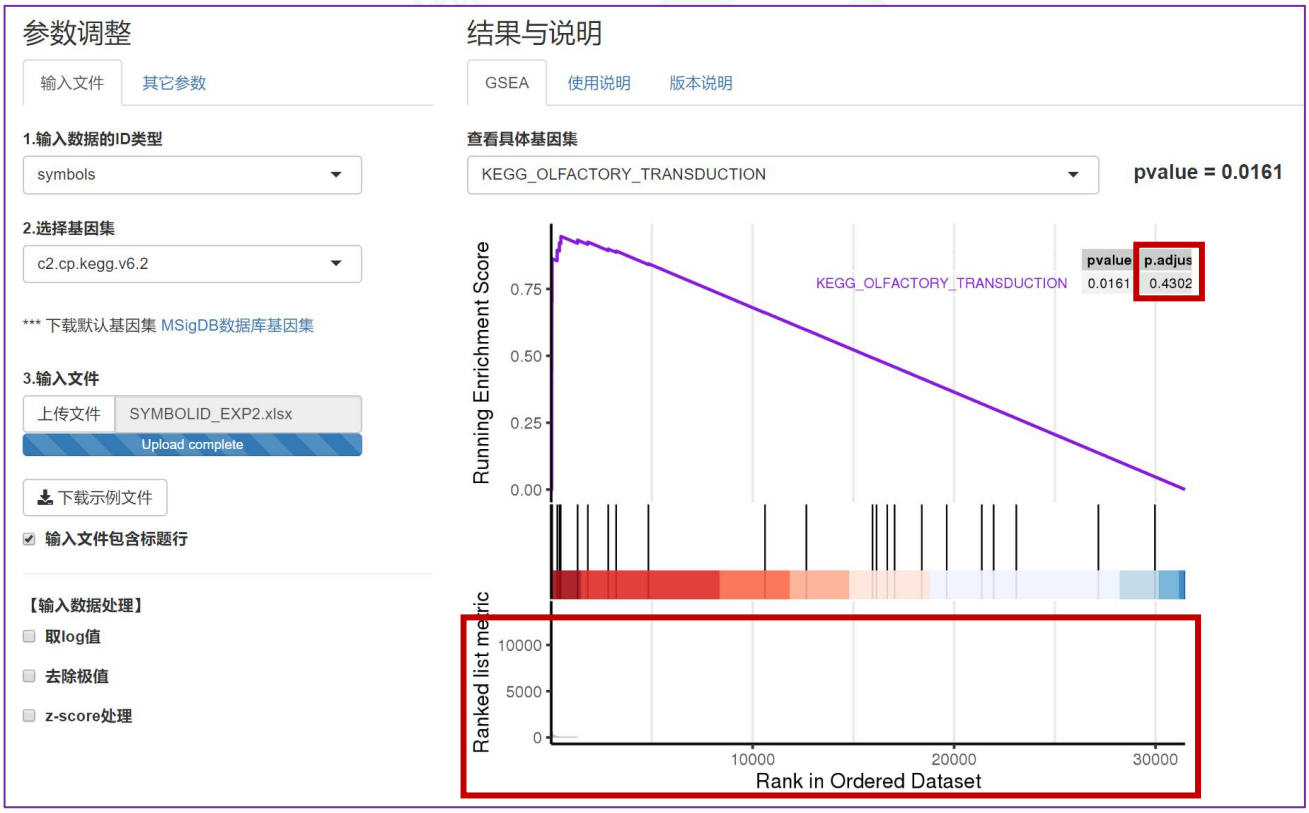
7. 输入数据处理

默认情况不对您输入的数据进行任何处理，十分建议您在上传数据前先自行整理（每种组学或者数据情况的整理方式可能不尽相同），如归一化或者去除极值等；若您不便进行处理，以下数据处理方案供您参考，但可能未必完全适合您的数据，请知悉。

若输入数据没有经过处理，GSEA 的分析结果可能不太理想，如下图所示：

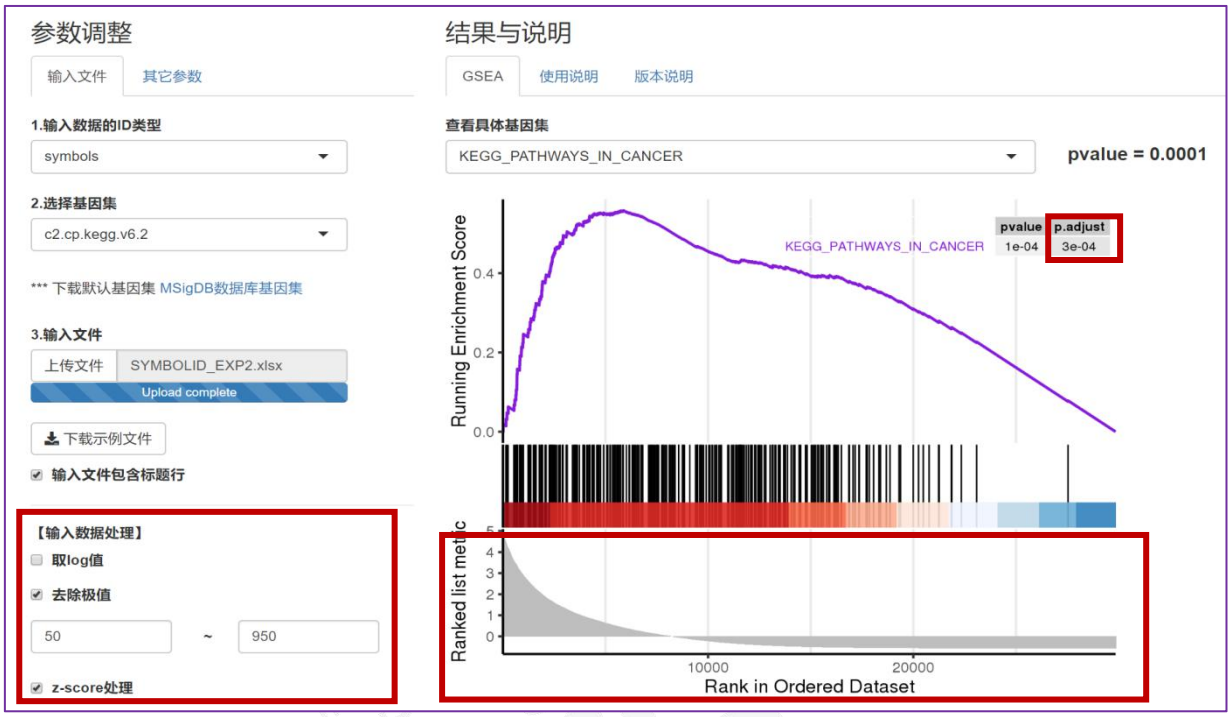
1.p 值较大（图例为该次分析所得到的的最小 p 值结果）；

2. “Ranked list metric” 图几乎没有内容，因为极值太大，导致小的数字在图上几乎不可见。



同样的数据，经过数据处理后，可得到如下结果：

1.p 值变得显著；2. “Ranked list metric” 图能相对清楚的展示您的数据分布



数据处理：

- **取 log 值**

  缩小数据间的差距，有助于分析分布跨数量级的数据，如：0.01~1000，跨了 6 个数量级。

- **去除极值**

  1. 极值可能是噪音，可以去掉这些数据进行分析，尝试能否使您的结果更显著。

  2. 选择去除极值后会显示一个区间范围，如 2~998，意为保留数值在总体中排序为 0.2%~99.8%的数据，即删除最小和最大的 0.2%的数据。

  3. 举个简单例子：如图为将数据从小到大四等分，若您选择 25~75，即意味着保留值在 0~3.59076 之间的数据（分别为 25%和 75%对应的数值）。在实际应用中，考虑到极值的个数只占总数的极小一部分，我们将数据千等份，即您可选的数字范围为 1~1000。

0%	25%	50%	75%	100%
0.000000e+00	0.000000e+00	2.573474e-01	3.590760e+00	1.404877e+04

  4. 在上个例子中您可能已经观察到，0%和 25%所对应 的数字均为零，此时，填 0~75 和 25~75 的效果是一样的。因为在 0%~25%之间的数字均为零，在这个区间内，删除掉任何比例的 0 值数据都将会是一种随机删除的过程，所以我们不进行删除。

- **z-score 处理**

  一种常见的归一化方法，公式为：$x_{new} = \frac{x - \mu}{\sigma}$

- 不论您以何种顺序选择，数据处理的顺序始终是先取 log、后去极值最后做 z-score（同它们在工具中的排列顺序）

8. 分析结果下载

如右图所示，提供了三种下载方式：

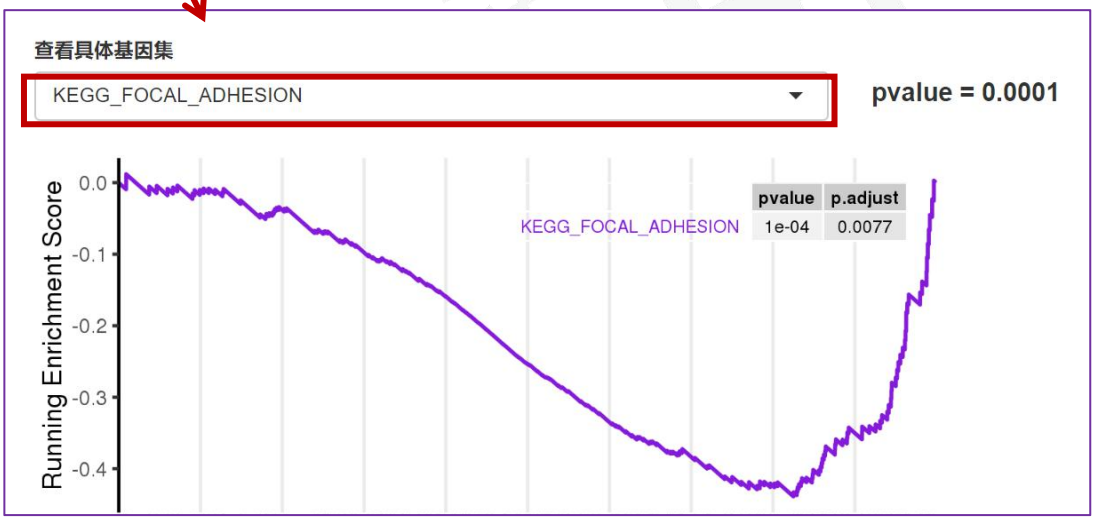
➢ 数据下载

1. “输出数据名”决定了该文件名，该名称根据输入数据名自动填充，如无必要不必修改。

2. 该内容为 GSEA 分析的表格结果，根据 p 值从小到大排列，您可以根据该表决定查看哪个基因集的图片结果 。

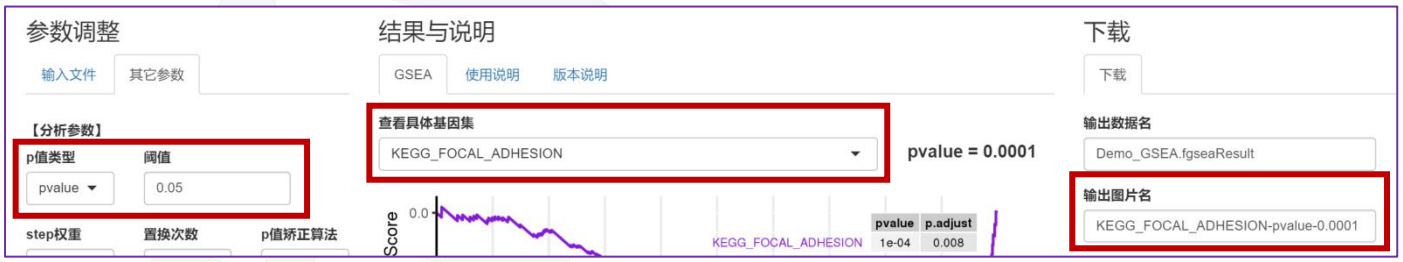
3. 如下图所示，做完一次分析后您可以先查看表格结果，挑选一些图片进行查看，然后决定是否需要调整参数重新分析。

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Description	setSize	enrichment	NES	pvalue	p.adjust	qvalues	rank	leading_ed	core_enrichment
2	KEGG_FOC	KEGG_FOC	192	-0.43886	-1.79554	0.00014	0.007728	0.005848	2183	tags=29%,	5228/7424/1499/4
3	KEGG_ECM	KEGG_ECM	81	-0.51076	-1.85878	0.000154	0.007728	0.005848	1985	tags=40%,	1288/3910/3371/1
4	KEGG_DNA	KEGG_DNA	33	0.722768	2.354554	0.000253	0.007728	0.005848	1905	tags=64%,	4174/4171/4175/4
5	KEGG_PRI	KEGG_PRI	34	0.622965	2.040386	0.000254	0.007728	0.005848	1452	tags=41%,	29851/6890/3932/
6	KEGG_PRO	KEGG_PRO	41	0.705994	2.415856	0.000257	0.007728	0.005848	2516	tags=66%,	5688/5709/5698/5
7	KEGG_OOC	KEGG_OOC	91	0.464904	1.858435	0.000286	0.007728	0.005848	896	tags=21%,	991/9133/983/408
8	KEGG_CELL	KEGG_CELL	114	0.669521	2.778754	0.0003	0.007728	0.005848	1234	tags=40%,	8318/991/9133/89
9	KEGG_CHE	KEGG_CHE	166	0.383167	1.678926	0.000335	0.007728	0.005848	1414	tags=22%,	3627/10563/6373/
10	KEGG_CYT	KEGG_CYT	233	0.341013	1.558165	0.000376	0.007728	0.005848	2228	tags=30%,	3627/10563/6373/
11	KEGG_TOL	KEGG_TOL	97	0.416527	1.683969	0.000575	0.010632	0.008046	2820	tags=35%,	3627/6373/4283/3
12	KEGG_TYR	KEGG_TYR	39	-0.57345	-1.81413	0.000822	0.013826	0.010463	1722	tags=38%,	7173/2184/1621/2
13	KEGG_P53	KEGG_P53	62	0.495909	1.839148	0.001081	0.015935	0.012059	1077	tags=26%,	9133/6241/983/11



➢ 图片下载

1. “输出图片名”决定了该文件名，该名称根据**所选基因集、p 值类型**和**该基因集的对应类型 p 值**自动填充，如无必要不必修改。



2. “输出图片类型”决定了图片后缀和类型，其中，pdf 和 svg 是矢量图，png 和 tiff 是高清非矢量图，其分辨率已经标注（足够用来发文章）。

3. 图片高度和宽度用于决定图片大小，同时还有以下作用：

A. 调整图片比例；

B. 调整字号，若原图为 7x7，改为 14x14 后，图片视觉比例不变，但是图中的字号看上去小一半。

4. 该按键只能下载当前展示的图片，用于下载您感兴趣但是 p 值不够显著的基因集图片。

➢ 打包下载

1.打包下载的内容包括：

A. GSEA 分析的表格结果；

B. 根据您选择的 p 值类型和阈值筛选出 top10 的基因集，并根据您选择的图片类型做出相应图片；

C. 本次分析所使用的参数，方便您后续回溯。见右图。

2.因为要绘制十张图片，打包下载可能耗时较久，建议您测试考虑成熟后再最终打包下载。

3.那些您关注的但是并不在 p 值最显著的 10 张图中的基因集建议您使用“图片下载”功能来下载。

	A	B
1	【输入信息】	
2	输入数据ID类型	entrez
3	选择的基因集	c2.cp.kegg.v6.2
4	输入文件	Demo_GSEA.xlsx
5	输入文件标题	是
6	输入数据处理	none
7		
8	【分析参数】	
9	p值类型	pvalue
10	阈值	0.05
11	step权重	1
12	置换次数	10000
13	矫正p值的算法	BH
14	基因集大小	10~5000
15		
16	【绘图参数】	
17	图片标题	
18	选择绘制的图片	ES图,热图,Rank图
19	字号	18
20	ES图颜色	#851ADA
21	ES图类型	线
22	是否在图中绘制p值	是

9.分析参数调整

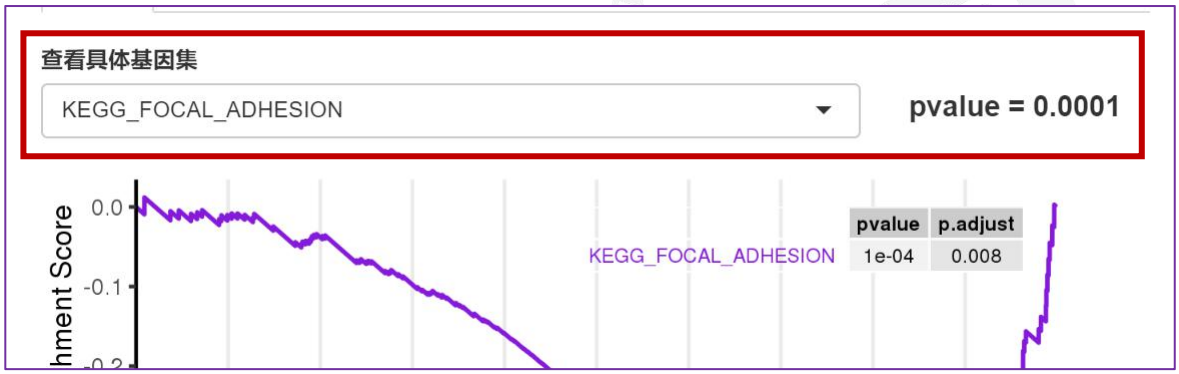
一般情况下，推荐您使用默认的参数值，此部分的参数调整不是必须的。若您有明确的需求，可以在这个界面进行调整。

- **p 值类型** 和 **阈值**：决定进行阈值筛选的 p 值类型。如下图所示，分析结果中给出三种 p 值类型。

  1.默认输出 pvalue（p 值类型）<0.05（阈值）的数据表格；

  2.默认展示 pvalue<0.05 的基因集的图片（即决定了“查看具体基因集”中可选择的基因集）

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	Description	setSize	enrichment	NES	pvalue	p.adjust	qvalues	rank	leading_ed	core_enrichment	
2	KEGG_FOC	KEGG_FOC	192	-0.43886	-1.79554	0.00014	0.007728	0.005848	2183	tags=29%,	5228/7424/1499/46	
3	KEGG_ECM	KEGG_ECM	81	-0.51076	-1.85878	0.000154	0.007728	0.005848	1985	tags=40%,	1288/3910/3371/12	
4	KEGG_DNA	KEGG_DNA	33	0.722768	2.354554	0.000253	0.007728	0.005848	1905	tags=64%,	4174/4171/4175/41	
5	KEGG_PRIN	KEGG_PRIN	34	0.622965	2.040386	0.000254	0.007728	0.005848	1452	tags=41%,	29851/6890/3932/3	
6	KEGG_PRO	KEGG_PRO	41	0.705994	2.415856	0.000257	0.007728	0.005848	2516	tags=66%,	5688/5709/5698/56	
7	KEGG_OOC	KEGG_OOC	91	0.464904	1.858435	0.000286	0.007728	0.005848	896	tags=21%,	991/9133/983/4085	
8	KEGG_CELL	KEGG_CELL	114	0.669521	2.778754	0.0003	0.007728	0.005848	1234	tags=40%,	8318/991/9133/890	
9	KEGG_CHE	KEGG_CHE	166	0.383167	1.678926	0.000335	0.007728	0.005848	1414	tags=22%,	3627/10563/6373/4	
10	KEGG_CYT	KEGG_CYT	233	0.341013	1.558165	0.000376	0.007728	0.005848	2228	tags=30%,	3627/10563/6373/4	
11	KEGG_TOL	KEGG_TOL	97	0.416527	1.683969	0.000575	0.010632	0.008046	2820	tags=35%,	3627/6373/4283/35	
12	KEGG_TYR	KEGG_TYR	39	-0.57345	-1.81413	0.000822	0.013826	0.010463	1722	tags=38%,	7173/2184/1621/29	
13	KEGG_P53_	KEGG_P53_	62	0.495909	1.839148	0.001081	0.015935	0.012059	1077	tags=26%,	9133/6241/983/111	



- **step 权重和置换次数**：与计算 ES 值的算法有关，在此不再赘述，感兴趣的可以去 GSEA 官网查看相关资料（http://software.broadinstitute.org/gsea/index.jsp）。修改这两个值会改变分析结果，包括 ES、NES 和 p 值等。
- **p 值矫正算法**：计算 p.adjust 值的算法。
- **基因集大小**：包含的基因个数在这个范围内的基因集才会进行分析。若基因集内基因个数太小，分析结果不具有统计学意义，基因集内基因个数太多，容易得到 p 值显著的结果，但是范围太广不利于您筛选目标基因。

10.绘图参数调整

- **图片标题**：可以为空，但是不能出现中文
- **选择绘制的图片**：GSEA 的分析结果图片从上到下由三张图片组成——ES 图、热图和 Rank 图，您可以根据您的情况调整图片。

  1.调整图片顺序；2.删除某张图（如 Rank 图绘制效果不理想）。
- **字号**：可以调整字体大小（除了下图红框内的文本）。
- **ES 颜色**：决定了 ES 图的线/点和图内的基因集名称的颜色。可以修改，既可以是#开头的颜色命名方式，也可以是 R 语言内置的颜色名，参考“使用说明”中的“参考的颜色值”。
- **ES 图形类型**：线或点。
- **在图中绘制 p 值**：决定了如图内容是否显示。