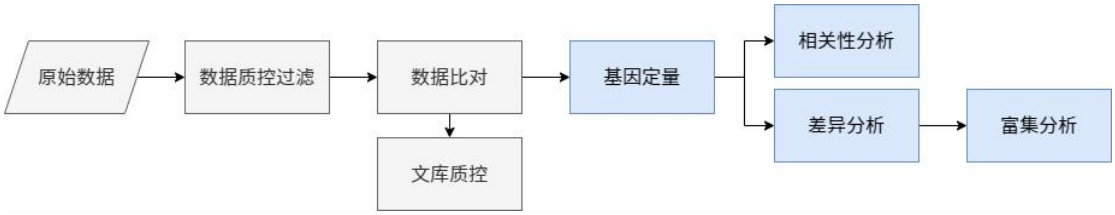
目录

金墁利 RNA-seq 数据分析

1.分析流程介绍

有参转录组分析是一种基于已知参考基因组的转录组测序数据分析方法，旨在研究基因表达水平、差异表达基因等分子特征。经过严格数据质控保证数据分析可靠性，详细的分析内容包括：数据质控、数据比对、基因定量、相关性分析、差异分析、富集分析等。分析流程图如下：



2. 原始数据质控

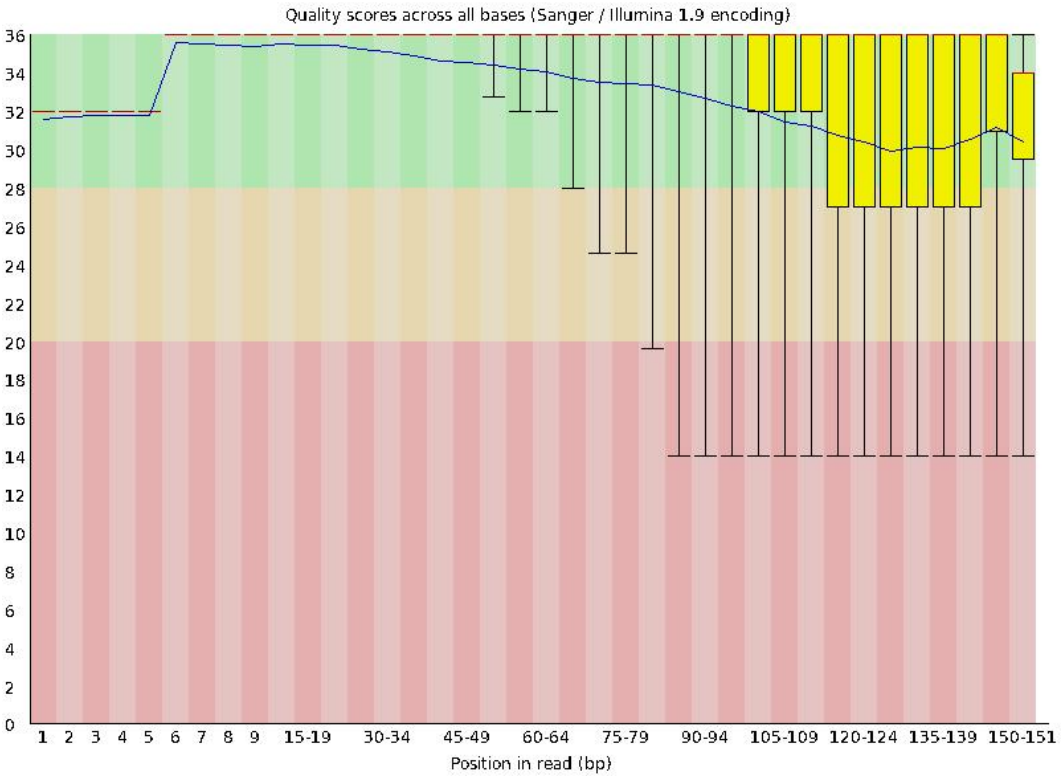
从测序平台获取 FASTQ 格式的原始数据（Raw Data），使用 FastQC、MultiQC 进行质量评估，fastp 进行数据过滤。

目录路径：RNASEQ\qc

2.1 FastQC

使用 FastQC 通过多个模块对测序数据进行全面评估，包括碱基质量、GC 含量、序列长度分布等，帮助用户判断测序数据的质量是否适合后续分析。分析结果目录中包含各样本测序质量报告。

目录路径：RNASEQ\qc\fastqc



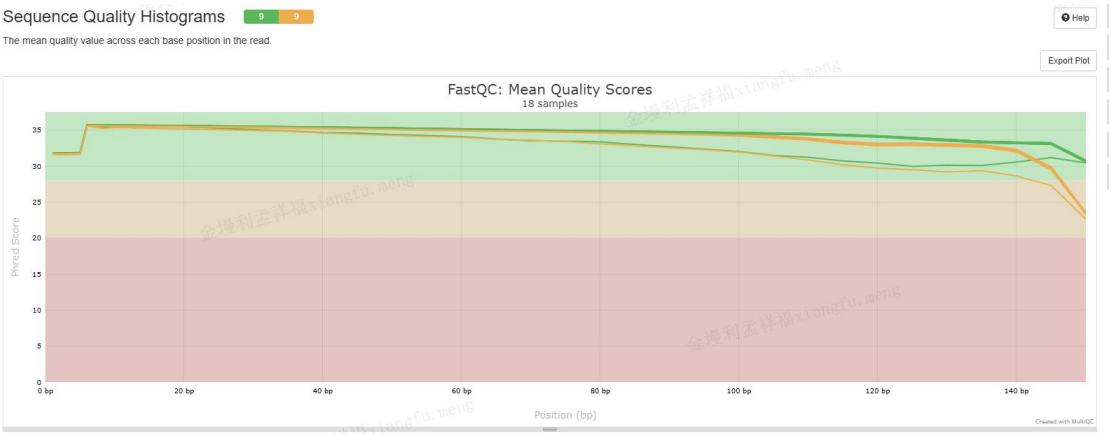
测序质量图

横坐标为测序循环数，纵坐标为测序 Q 值

2.2 MultiQC

MultiQC 将来自 FastQC 结果汇总成一个交互式的 HTML 报告，生成的 HTML 报告包含图表和统计数据，用户可以通过交互式界面查看详细信息，通过整合 FastQC 等工具的输出，快速评估测序数据的质量，识别潜在问题（如低质量序列、接头污染等）。

文件路径：RNASEQ\qc\multiqc_data\multiqc.html



序列质量图

横坐标为横坐标为测序循环数，纵坐标为测序 Q 值，每条线代表一个样本的测序质量趋势

2.3 fastp

用 fastp 软件对每一个样本的测序数据原始 FASTQ 数据做质控处理，得到过滤后 FASTQ 数据用于下游分析。

参数： -q 15 -u 40 -l 15 --cut_right --cut_window_size 4 --cut_mean_quality 20 --correction

参数解析：

-q 15：合格的质量值阈值为 15；

-u 40：Read 中允许的未合格碱基的百分比限制为 40%；

-l 15：Read 的最小长度阈值为 15；

--cut_window_size 4 --cut_mean_quality 20：从左到右的滑动窗口（4bp）修剪，如果遇到一个窗口的平均质量低于 20，则会丢弃该窗口及其右侧的所有碱基，并停止修剪；

--correction：启用基于配对末端（PE）数据重叠区域的碱基校正。

目录路径：RNASEQ\trimmed\fastp.stats.tsv

Sample	Clean_Reads	Total_Base	Q20	Q30	Q20_Rate	Q30_Rate	Average_Length	GC
RNStest5a	69249280	9861274208	9728598545	9519575388	0.986546	0.965349	142	0.526699
RNStest6a	71700580	10199472821	10058163088	9831513607	0.986145	0.963924	142	0.521106
RNStest7a	68432346	9639191100	9512466245	9318807638	0.986853	0.966762	140.5	0.545584
RNStest8a	65762254	9325199329	9194479077	8984196152	0.985982	0.963432	141	0.522801

FASTQ 质量表

Sample：样本名称；

Clean Reads：过滤后的 Read 数量；

Total Base：为过滤后的 Base 数量；

Q20：达到 Q20 的碱基数量；

Q30：达到 Q30 的碱基数量；

Q20 Rate：达到 Q20 的碱基数量比例；

Q30 Rate：达到 Q30 的碱基数量比例；

Average Length：过滤后的平均 Read 长度；

GC：平均的 Read GC 含量比例

3. 数据比对

使用 HISAT2 软件将过滤后的 FASTQ 数据比对到参考基因组 hg38，生成的 SAM 文件将包含过滤后的 FASTQ 数据与 hg38 参考基因组的详细比对记录，为后续的转录本组装或差异表达分析提供基础数据。

参数：--dta

参数解析：

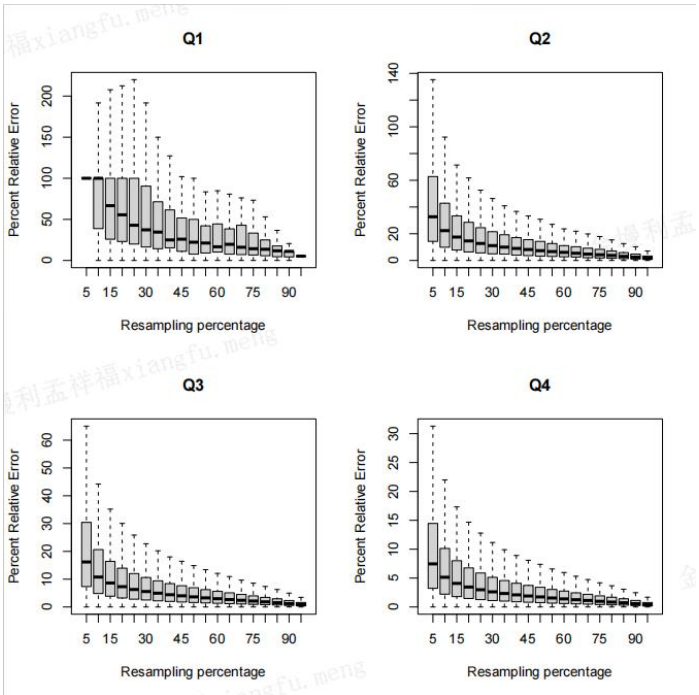
--dta：用于优化比对结果，调整 HISAT2 的比对策略，使其在发现新的剪接位点时使用更长的锚定长度，从而减少短锚定比对的数量。这有助于提高转录本组装的准确性和效率。

使用 RSeQC-RPKM_saturation 和 qualimap rnaseq 工具评估文库质量，保障数据质量与后续分析准确性。

3.1 RSeQC-RPKM_saturation

从总 RNA 读取中重抽取一系列子集，并计算每个子集的 RPKM 值。通过这种方式，我们可以检查当前的测序深度是否达到饱和状态（即 RPKM 值是否稳定）以估计基因表达。如果测序深度饱和，估计的 RPKM 值将是稳定的。

目录路径：RNASEQ\qc\RPKM_saturation



RNA-seq 测序 RPKM 饱和度分析图

横坐标表示 20 个重抽样比例，依次从 5%, 10%, …, 95%, 100%；

纵坐标表示“百分比相对误差”或“百分比误差”，用于衡量从读取子集估计的 RPKM 与真实表达水平之间的偏差；

所有转录本根据表达水平（RPKM）按升序排序后，分为四组：

Q1(0-25%)：表达水平排名在 25 百分位以下的转录本；

Q2(25-50%)：表达水平排名在 25 百分位到 50 百分位之间的转录本；

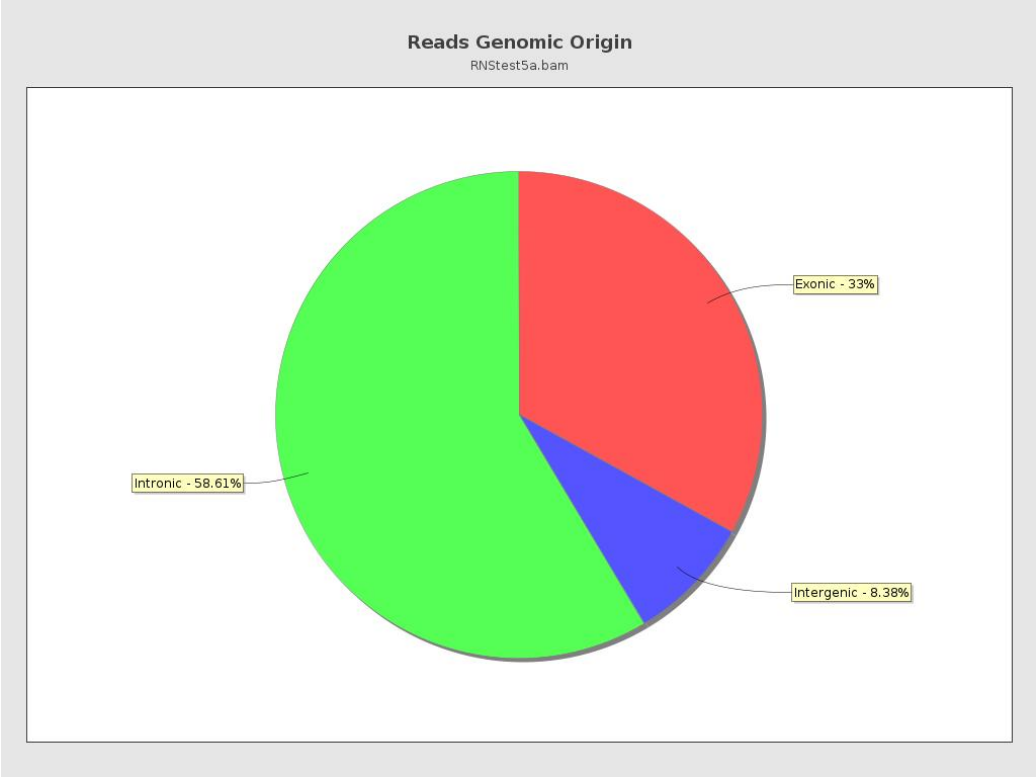
Q3(50-75%)：表达水平排名在 50 百分位到 75 百分位之间的转录本；

Q4(75-100%)：表达水平排名在 75 百分位以上的转录本

3.2 Qualimap-rnaseq

RNA-seq QC 报告提供了针对全转录组测序的质量控制指标和偏差估计，包括读取的基因组来源、接头分析、转录本覆盖率和 5'-3'偏差计算。

目录路径：RNASEQ\qc\qualimap_rnaseq



基因区域占比图

展示注释到各个功能区域的比例分布

4. 基因定量

使用 subread 软件包中的 featurecounts 软件计算基因丰度，获取原始的基因定量表格。

参数：-O --fracOverlap 0.2 -J -p -t exon -g gene_id

参数解析：

-O：在计数时允许重叠的 reads 被算入多个基因或特征中；

--fracOverlap 0.2：设置重叠计算的阈值为 20%，如果一个 reads 与某个基因区域的重叠部分占该区域长度的 20%或更多，该 Reads 才会被计入该特征的计数；

-J：将相同基因的多个 exon 合并计算；

-p：输入为双端 Reads 数据；

-t exon：指定了只针对外显子计数；

-g gene_id：指定在注释文件中 gene_id 列包含基因的 ID

目录路径：RNASEQ\feature counts

4.1 原始表格

Geneid	Length	RNStest5a	RNStest6a	RNStest7a	RNStest8a
ENSG00000223972	1735	66	23	10	2
ENSG00000227232	1351	119	193	116	57
ENSG00000278267	68	16	10	10	3
ENSG00000243485	1021	0	0	0	0
ENSG00000274890	138	0	0	0	0
ENSG00000237613	1219	0	0	0	0
ENSG00000268020	840	0	0	0	0
ENSG00000240361	940	0	0	0	0
ENSG00000186092	918	0	0	0	0
ENSG00000238009	3726	23	3	4	47
ENSG00000239945	1319	0	0	0	2
ENSG00000233750	3812	78	21	41	182
ENSG00000268903	755	192	69	91	401
ENSG00000269981	284	97	41	45	174
ENSG00000239906	323	15	4	13	8
ENSG00000241860	6195	31	4	12	15
ENSG00000222623	104	0	0	0	0
ENSG00000241599	457	0	0	0	0
ENSG00000279928	718	82	21	13	6
ENSG00000279457	1982	175	284	278	190

基因表达表

行为基因 Ensembl 编号，第一列为基因长度，之后的列为样本名称，表中数值为基因的 Reads

数量

4.2 RPKM

RPKM（Reads Per Kilobase per Million mapped reads）和 TPM（Transcripts Per Kilobase per Million mapped reads）是 RNA-seq 数据分析中常用的两种标准化方法，用于校正基因表达量中的测序深度和基因长度的影响，以便在不同样本或基因之间进行比较。

RPKM 计算公式如下：

$$RPKM = \frac{\text{reads mapped to gene}}{\text{total mapped reads (in millions)} \times \text{gene length (in kilobases)}}$$

其中：

reads mapped to gene：比对到某个基因的 reads 数。

total mapped reads：样本中所有比对到基因组的 reads 总数，以百万为单位。

gene length：基因的长度，以千碱基（kb）为单位。

Geneid	RNStest5a	RNStest6a	RNStest7a	RNStest8a
ENSG00000223972	1.481360354	0.515613027	0.24699162	0.047936628
ENSG00000227232	3.430108623	5.556450947	3.679462138	1.754512522
ENSG00000278267	9.162781512	5.719875975	6.301918545	1.83463345
ENSG00000243485	0	0	0	0
ENSG00000274890	0	0	0	0
ENSG00000237613	0	0	0	0
ENSG00000268020	0	0	0	0
ENSG00000240361	0	0	0	0
ENSG00000186092	0	0	0	0
ENSG00000238009	0.240381614	0.031316551	0.046004344	0.524556138
ENSG00000239945	0	0	0	0.063055383
ENSG00000233750	0.796815863	0.214270275	0.46090632	1.985434031
ENSG00000268903	9.903085713	3.554656698	5.165069133	22.08688076
ENSG00000269981	13.30055168	5.615145851	6.790095334	25.47814904
ENSG00000239906	1.808443719	0.481673766	1.724735602	1.029969656
ENSG00000241860	0.194866257	0.025113903	0.083008322	0.100690133
ENSG00000222623	0	0	0	0
ENSG00000241599	0	0	0	0
ENSG00000279928	4.447394647	1.137602074	0.775890807	0.347507172
ENSG00000279457	3.438354566	5.573271686	6.010669434	3.986455461

RPKM 表

行为基因 Ensembl 编号，列为样本名，值为 RPKM 标准化数值

4.3 TPM

TPM 是对 RPKM 的改进方法，其计算公式为：

$$TPM = \frac{\text{reads mapped to gene / gene length (in kilobases)}}{\text{sum of (reads mapped to all genes / gene lengths)}} \times 10^6$$

其中：

reads mapped to gene：比对到某个基因的 reads 数。

gene length：基因的长度，以千碱基（kb）为单位。

Geneid	RNStest5a	RNStest6a	RNStest7a	RNStest8a
ENSG00000223972	3.11682551	1.203411735	0.418649273	0.098912063
ENSG00000227232	7.217048863	12.96844323	6.236665634	3.620247411
ENSG00000278267	19.27876029	13.34986803	10.68171307	3.785568307
ENSG00000243485	0	0	0	0
ENSG00000274890	0	0	0	0
ENSG00000237613	0	0	0	0
ENSG00000268020	0	0	0	0
ENSG00000240361	0	0	0	0
ENSG00000186092	0	0	0	0
ENSG00000238009	0.505769946	0.073091065	0.077977079	1.08236503
ENSG00000239945	0	0	0	0.130107983
ENSG00000233750	1.676523881	0.500094742	0.781233369	4.096729045
ENSG00000268903	20.83638198	8.29636832	8.75476033	45.57389695
ENSG00000269981	27.98474975	13.10543383	11.50916972	52.57141338
ENSG00000239906	3.805018477	1.124199413	2.923416209	2.12523133
ENSG00000241860	0.410004305	0.058614433	0.140698593	0.207763232
ENSG00000222623	0	0	0	0
ENSG00000241599	0	0	0	0
ENSG00000279928	9.357448412	2.655099102	1.315130133	0.717043579
ENSG00000279457	7.234398568	13.00770189	10.18804762	8.225621009

TPM 表

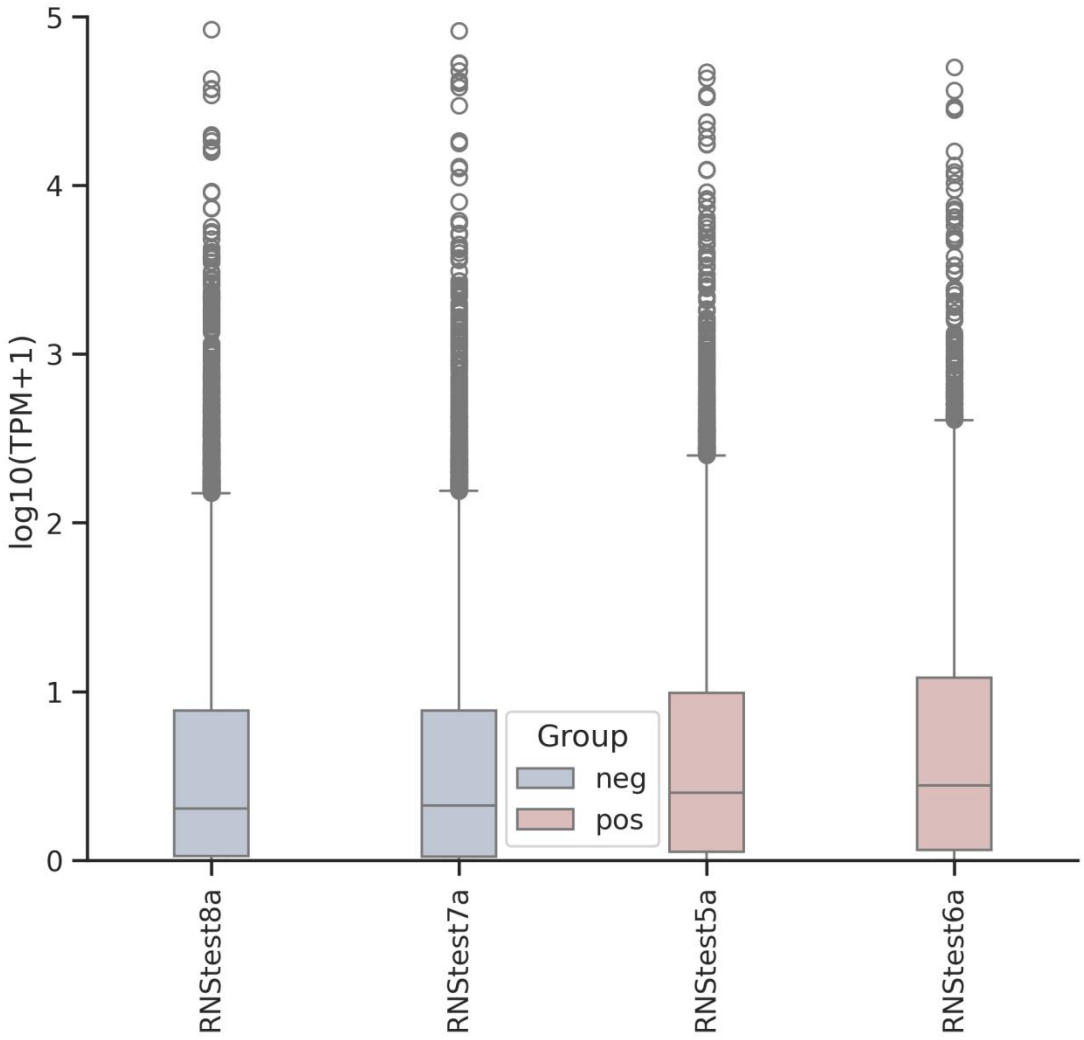
行为基因 Ensembl 编号，列为样本名，值为 TPM 标准化数值

4.4 基因表达分布

为了全面展示各个样本中基因表达量的整体分布情况，我们使用 R 语言对所有样本的基因 TPM 值进行了可视化分析，分别绘制了“小提琴图”和“箱线图”。小提琴图结合了箱线图和核密度图的优势，能够直观地展示数据的分布形态和概率密度，而箱线图则清晰地呈现了

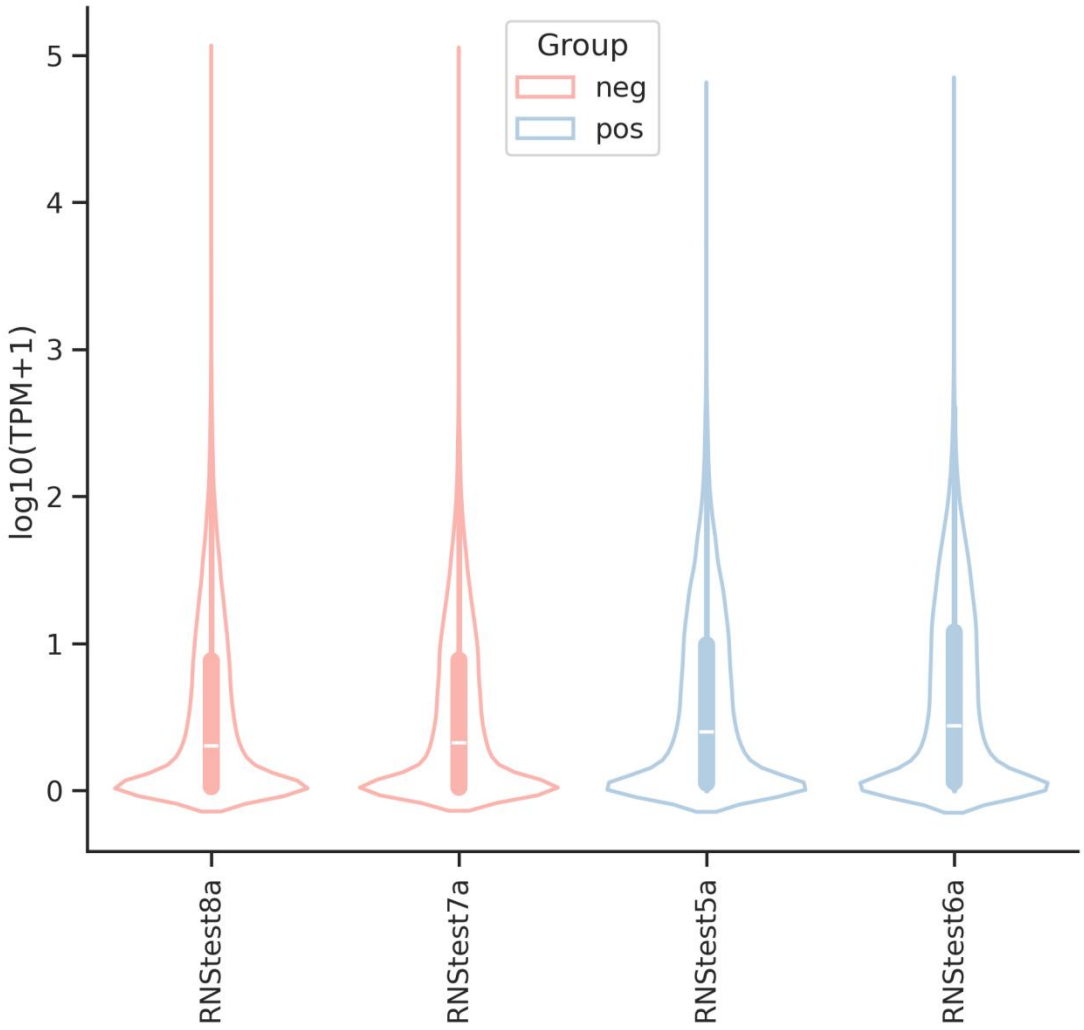
数据的中位数、四分位数以及可能的离群值。由于 TPM 值的跨度较大，为了避免图中数据点的分布过于分散，我们对 TPM 值进行了 log10(TPM+1)的变换处理。

目录路径：RNASEQ\gene_quant\{boxplot|violinplot}



基因表达分布箱线图

横坐标为样本，纵坐标为 log10 标准化数值，不同颜色代表不同分组



基因表达分布小提琴图

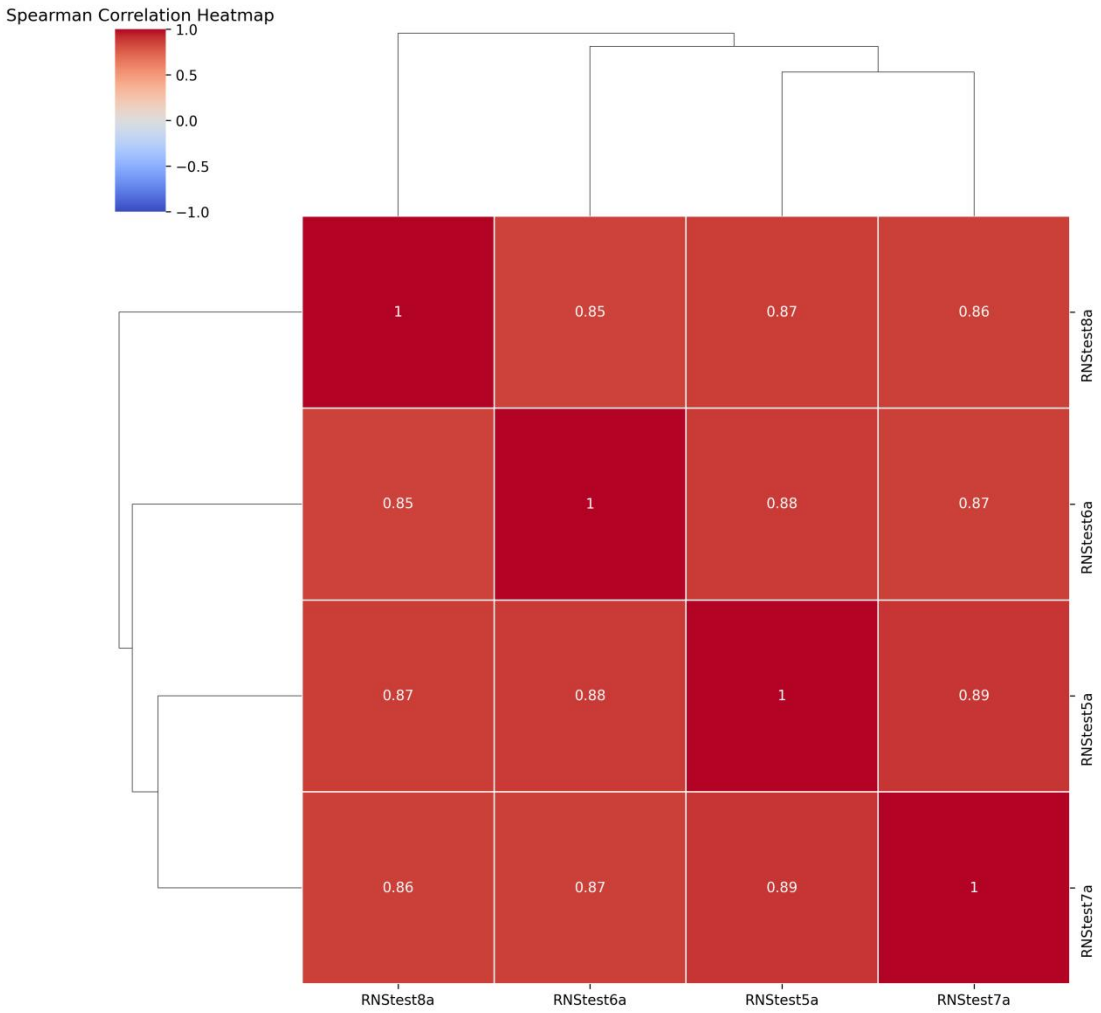
横坐标为样本，纵坐标为 log10 标准化数值，不同颜色代表不同分组

4.5 样本间相关性

在本研究中，基于样本 TPM 或 FPKM 数据，计算两两样本间相关性系数，以此评估样本组内生物学实验重复性。数据仅保留所有样本中 TPM 之和大于 0 的基因，过滤掉完全不表达基因的信息。

ENCODE project 2016 年的 RNA-Seq 指导方案推荐样本间相关系数要在 0.9 或 0.8 以上，来自同一个人的样本需要 0.9 以上，不同的人的样本需要在 0.8 以上。

目录路径：RNASEQ\gene_quant\intersample_correlation



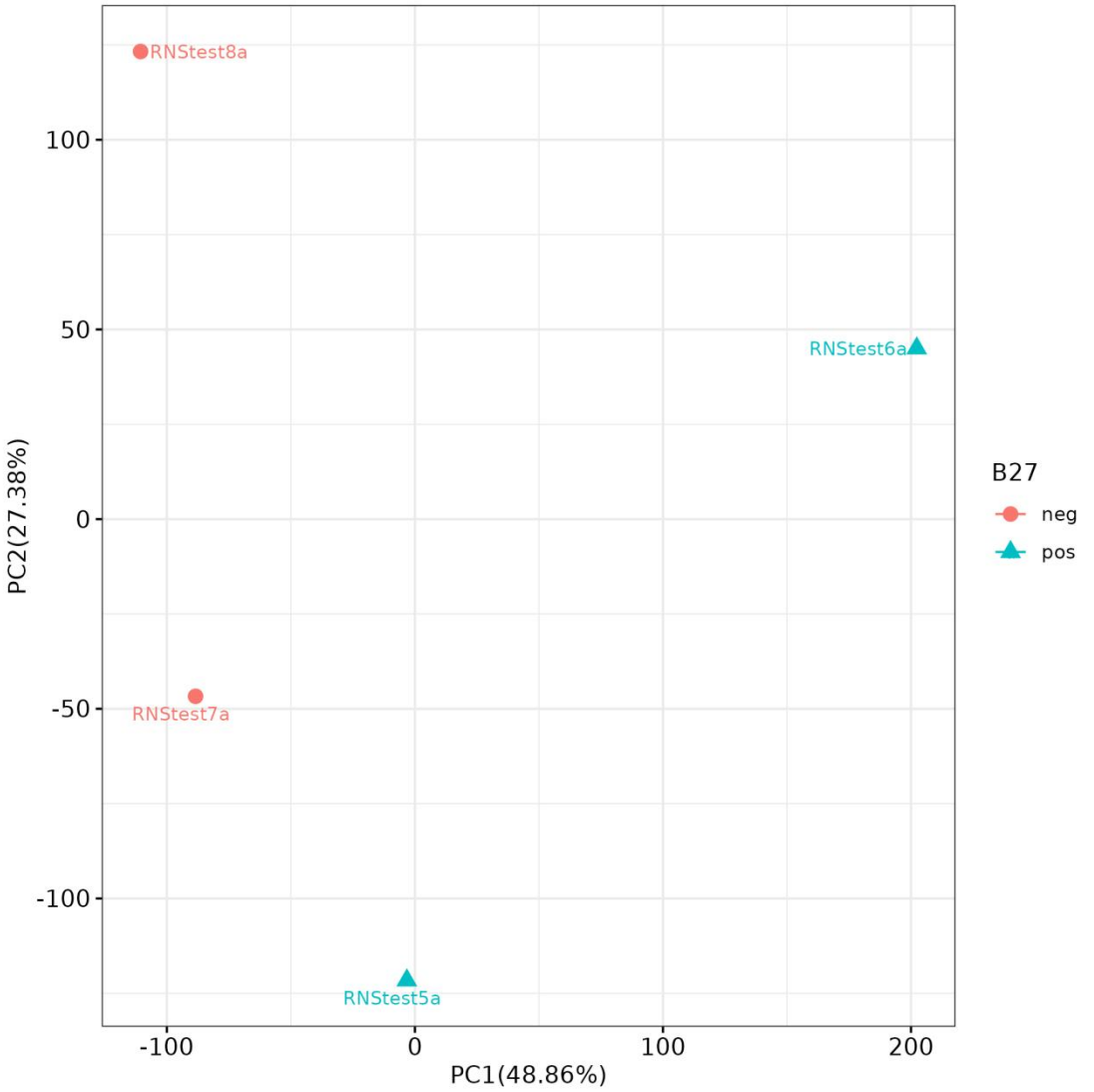
样本间相关性热图

横纵坐标为样本名，数值为样本间相关系数，根据相关性系数颜色深浅不同，越深红相关性越高，越深蓝色相关性越低

4.6 主成分分析

主成分分析（PCA）常被用于评估组间差异与组内样本重复情况。该方法运用线性代数算法，对大量基因变量降维并提取主成分。我们对所有样本的基因表达值（TPM）开展 PCA 分析，结果如图所示。理想状态下，PCA 图中组间样本分散，组内样本聚集。

目录路径：RNASEQ\gene_quant\pca



组间主成分 PCA 图

图中横坐标为第一主成分,纵坐标为第二主成分,不同的颜色表示不同的分组。样本点之间的距离近似于样本之间各基因 TPM 差异的总和

5. 差异表达基因

在转录组中,确定某个基因在不同的样品中的表达量是否有差异是分析的核心内容之一。获得基因表达量后,即可对表达数据进行统计学分析,进而筛选不同样本之间显著差异的基因。运用 DESeq2 这一强大的 R 包进行差异基因计算，它基于负二项分布模型，将原始计数数据导入，结合样本分组构建矩阵，经标准化与统计检验，识别出样本组间的差异基因，为探究分子机制提供线索，输出差异基因表格。

5.1 原始差异基因表

log2FoldChange（log2FC）表示处理组相对于对照组的表达变化倍数的对数值。log2FC>0 表示基因在处理组中的表达上调；log2FC<0 表示基因在处理组中的表达下调；log2FC 的绝对值表示差异表达的幅度，绝对值越大，表示差异越显著。

文件路径：RNASEQ\gene diff\DESeq2\ALL\{scheme}.{case} vs {control}.tsv

gene_id	symbol	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000158747	NBL1	280.417584	2.561180133	0.195332461	13.11190226	2.81E-39	6.26E-35
ENSG00000180537	RNF182	994.9571122	-4.17760113	0.332740712	-12.5551247	3.73E-36	4.15E-32
ENSG00000225073	DDX39B	162.6646959	-8.017562204	0.724110558	-11.07229016	1.71E-28	1.27E-24
ENSG00000232280	FLOT1	56.67507056	5.625281119	0.619330096	9.082847989	1.06E-19	4.71E-16
ENSG00000170160	CCDC144A	48.01225067	3.988546784	0.491705602	8.111656181	4.99E-16	1.59E-12
ENSG00000059122	FLYWCH1	1366.453154	1.329432021	0.175060043	7.594148849	3.10E-14	7.13E-11
ENSG00000066926	FECH	3013.266388	-1.501087503	0.197778483	-7.589741228	3.21E-14	7.13E-11
ENSG00000165801	ARHGEF40	3356.035009	-1.943963426	0.265489423	-7.322187835	2.44E-13	4.52E-10
ENSG00000214022	REPIN1	3469.215641	-1.158437567	0.159451417	-7.265144402	3.73E-13	6.38E-10
ENSG00000168389	MFSD2A	992.6438164	-1.31954389	0.190816338	-6.915256333	4.67E-12	6.55E-09
ENSG00000180304	OAZ2	3720.232443	-1.405703908	0.20331063	-6.914069906	4.71E-12	6.55E-09
ENSG00000196329	GIMAP5	811.2358777	1.565588904	0.238116783	6.574878442	4.87E-11	6.37E-08
ENSG00000221869	CEBPD	1752.99734	-1.791846613	0.273790534	-6.544589346	5.97E-11	7.38E-08
ENSG00000171798	KNDC1	70.51688673	3.160406223	0.485555338	6.508848682	7.57E-11	8.87E-08
ENSG00000132199	ENOSF1	393.5886906	1.186362505	0.183326246	6.47131837	9.72E-11	1.08E-07
ENSG00000115306	SPTBN1	6831.905311	0.903373361	0.142218709	6.35200084	2.13E-10	2.15E-07
ENSG00000275584	LILRA6	36.02231829	-5.036950998	0.795411589	-6.332508941	2.41E-10	2.33E-07
ENSG00000127507	ADGRE2	2956.627619	-0.810204228	0.128745417	-6.293072387	3.11E-10	2.89E-07
ENSG00000228628	ATF6B	24.48940686	-4.06111797	0.646516523	-6.281537792	3.35E-10	2.98E-07
ENSG00000144152	FBLN7	184.8401274	2.068583108	0.331851431	6.233461467	4.56E-10	3.91E-07

差异基因 DESeq2 原始表

gene_id：Ensembl 基因编号；

symbol：基因名称；

baseMean：DESeq2 结果特有，所有样本校正后表达量均值；

log2FoldChange：处理组与对照组基因表达水平的比值,再经过差异分析软件收缩模型处理,最后以 2 为底取对数；

lfcSE：log2 倍数变化估计的标准误差估计；

stat：该基因或转录本的检验统计量值；

pvalue：显著性检验的 p 值；

padj：BH 方法校正后的 p 值；

5.2 显著差异基因表

显著的差异基因挑选规则：

symbol 不为 NA，即需要注释出基因名；

|log2FoldChange|>1 & padj<0.05 是常用的经验值

文件路径：RNASEQ\gene_diff\DESeq2\DEG\{scheme}.{case}_vs_{control}.tsv

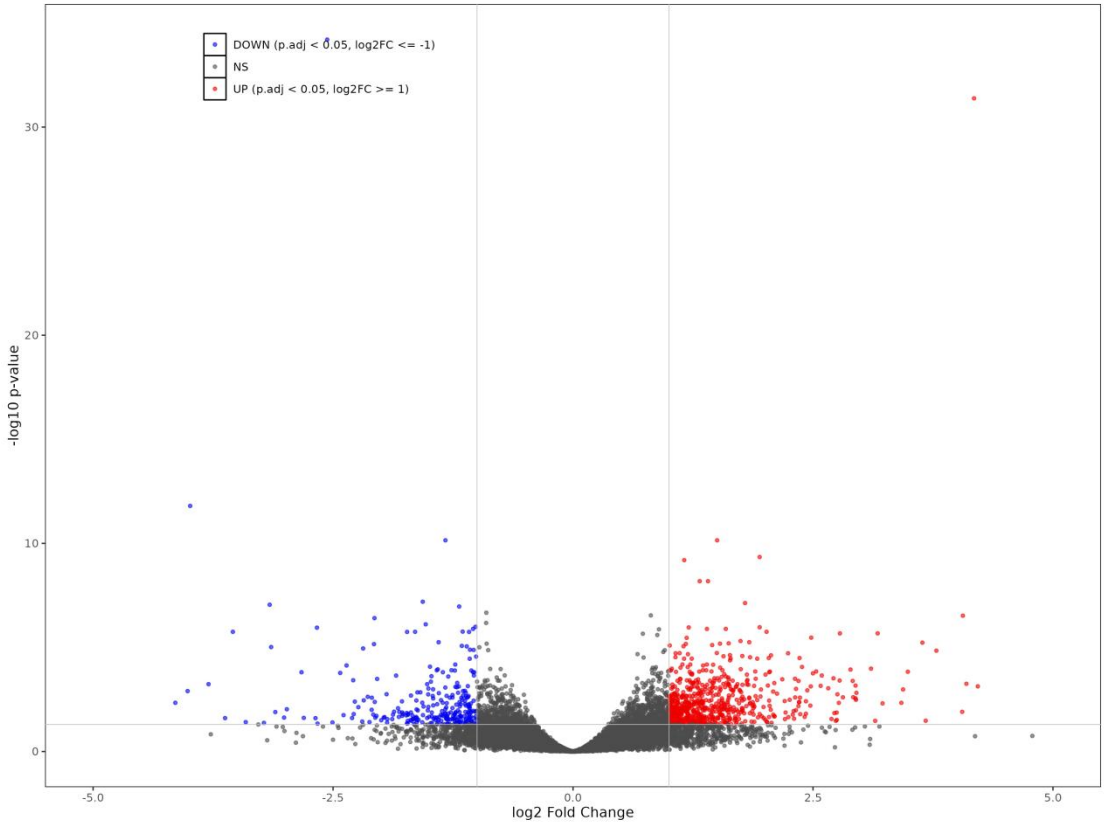
gene_id	symbol	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000236884	HLA-DRB1	65.33304156	-9.650523483	3.384045632	-2.851771085	0.00434764	0.042383298
ENSG00000227993	HLA-DRA	275.7587263	8.708618752	2.426655367	3.588733229	0.000332289	0.008003697
ENSG00000229074	HLA-DRB1	403.4641054	-8.631340495	1.515716174	-5.694562507	1.24E-08	5.01E-06
ENSG00000225073	DDX39B	162.6646959	-8.017562204	0.724110558	-11.07229016	1.71E-28	1.27E-24
ENSG00000168468	ATF6B	15.50513177	-7.571962126	1.928827236	-3.925681877	8.65E-05	0.003284636
ENSG00000206433	LST1	13.77409376	-7.392944212	1.32836539	-5.565444771	2.61E-08	8.43E-06
ENSG00000233209	HLA-DQB1	129.956497	-6.361558916	1.865324949	-3.41042933	0.000648607	0.012520379
ENSG00000230034	PSMB8	13.60737619	-5.971362015	1.116991808	-5.345931789	9.00E-08	2.27E-05
ENSG00000278046	LILRA3	18.34358107	-5.808572269	1.271746066	-4.567399442	4.94E-06	0.000436123
ENSG00000232367	TAP1	22.42184613	5.790078663	1.164006613	4.974266122	6.55E-07	0.000109599
ENSG00000232280	FLOT1	56.67507056	5.625281119	0.619330096	9.082847989	1.06E-19	4.71E-16
ENSG00000236353	PEX2	36.43841832	-5.384803169	1.584856582	-3.397659592	0.000679649	0.012895375
ENSG00000278415	FCAR	166.9737599	-5.357083748	1.566319506	-3.420173042	0.000625813	0.012236009
ENSG00000274619	CFD	133.287423	-5.191040435	1.285061735	-4.039526112	5.36E-05	0.002328156
ENSG00000275584	LILRA6	36.02231829	-5.036950998	0.795411589	-6.332508941	2.41E-10	2.33E-07
ENSG00000233490	GPSM3	18.43191472	-4.216679978	0.956584372	-4.408058612	1.04E-05	0.000739277
ENSG00000180537	RNF182	994.9571122	-4.17760113	0.332740712	-12.5551247	3.73E-36	4.15E-32
ENSG00000225201	HLA-E	458.2648904	4.143456374	1.089256938	3.803929293	0.000142419	0.004567253
ENSG00000232126	HLA-B	25935.11904	-4.097822355	0.911641097	-4.494995198	6.96E-06	0.000556973
ENSG00000228628	ATF6B	24.48940686	-4.06111797	0.646516523	-6.281537792	3.35E-10	2.98E-07

显著的差异基因表

5.2 差异基因火山图

差异基因火山图是有参转录组 RNAseq 分析中的关键可视化结果，直观展示差异基因分布。通过设定阈值，可清晰筛选出显著上调或下调基因，助力锁定关键基因开展后续研究。

文件路径：RNASEQ\gene_diff\DESeq2\ALL\{scheme}.{case}_vs_{control}.volcano.png



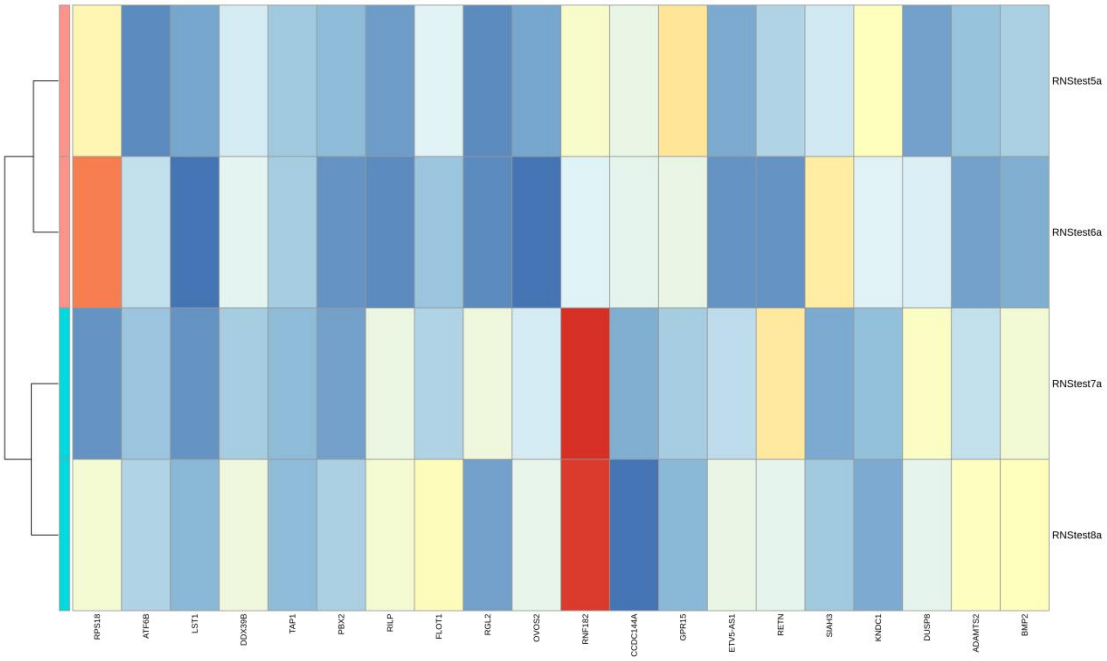
差异基因火山图

纵轴为统计学显著性（P 值），横轴为基因表达变化倍数（log2 Fold Change），红色点代表上调基因，蓝色点代表下调基因，黑色点代表不显著基因。

5.3 组间差异基因热图

差异基因聚类热图呈现了表达量矩阵中最为显著的基因。考虑到图片可读性，我们对展示基因数量加以控制，最终呈现的基因不超过 20 个，以此清晰直观地展现关键基因表达情况。

目录路径：RNASEQ\gene diff\DESeq2\pheatmap



组间 TOP20 差异基因热图

横坐标为基因名，纵坐标为样本名，左侧为聚类结果，颜色代表表达量。

6. 富集分析

在有参转录组 RNAseq 研究里，我们借助 R 语言的 clusterProfiler 包开展富集分析。该包整合丰富数据库资源，将筛选出的差异基因作为输入。它能从基因本体（GO）的生物过程、细胞组分、分子功能，以及 KEGG 通路等层面，利用超几何分布等算法，识别基因显著富集的功能类别与通路，助力解析差异基因参与的生物学过程，挖掘潜在调控机制。

6.1 GO

在 RNAseq 分析中，GO 富集分析（Gene Ontology Enrichment Analysis）用于解释差异表达基因（DEGs）的潜在生物学功能，其中 Biological Process（生物过程，BP）、Cellular Component（细胞组分，CC）、Molecular Function（分子功能，MF）是 GO（Gene Ontology）的三个核心分类，分别代表基因功能的三个层次。

目录路径：RNASEQ\enrich\GO

6.1.1 GO 通路注释

ONTOLOGY	ID	Descripti	GeneRatic	BgRatio	RichFactor	FoldEnrichment	zScore	pvalue	p.adjust	qvalue	geneID	Count
BP	GO:0002768	immune re	40/744	369/18986	0.108401084	2.766267448	6.919430339	6.31E-09	3.01E-05	2.65E-05	3123/3119/	40
BP	GO:0050867	positive	40/744	391/18986	0.10230179	2.610620686	6.498888752	3.21E-08	7.65E-05	6.73E-05	3123/3122/	40
BP	GO:0002429	immune re	36/744	336/18986	0.107142857	2.734158986	6.477011589	4.90E-08	7.79E-05	6.85E-05	3123/3119/	36
BP	GO:0002696	positive	36/744	371/18986	0.09703504	2.476219459	5.79911743	5.71E-07	0.000589423	0.00051823	3123/3122/	36
BP	GO:0045088	regulatic	42/744	470/18986	0.089361702	2.280404942	5.676468437	6.18E-07	0.000589423	0.00051823	10211/3133	42
BP	GO:0050863	regulatic	37/744	393/18986	0.094147583	2.402534953	5.674018881	8.34E-07	0.000595313	0.000523408	3123/3122/	37
BP	GO:0002697	regulatic	38/744	410/18986	0.092682927	2.365158668	5.643577598	8.73E-07	0.000595313	0.000523408	3123/3122/	38
BP	GO:0002507	tolerance	9/744	30/18986	0.3	7.655645161	7.36771775	1.42E-06	0.000823608	0.000724129	3133/3106/	9
BP	GO:0045089	positive	36/744	387/18986	0.093023256	2.373843461	5.514466199	1.55E-06	0.000823608	0.000724129	10211/3133	36
BP	GO:0051453	regulatic	15/744	90/18986	0.166666667	4.253136201	6.247325063	2.04E-06	0.000951761	0.000836803	6556/6521/	15
BP	GO:0002758	innate im	30/744	297/18986	0.101010101	2.577658303	5.534185129	2.19E-06	0.000951761	0.000836803	10211/2858	30
BP	GO:0002833	positive	37/744	416/18986	0.088942308	2.269702492	5.288082575	3.21E-06	0.001230571	0.001081937	3123/10211	37
BP	GO:0030641	regulatic	15/744	94/18986	0.159574468	4.072151682	6.030081133	3.57E-06	0.001230571	0.001081937	6556/6521/	15
BP	GO:0006885	regulatic	16/744	106/18986	0.150943396	3.851896936	5.946224571	3.61E-06	0.001230571	0.001081937	6556/6521/	16
BP	GO:0050870	positive	27/744	260/18986	0.103846154	2.650031017	5.410181257	4.17E-06	0.001327831	0.00116745	3123/3122/	27
BP	GO:0034113	heterotyp	12/744	63/18986	0.19047619	4.860727087	6.198690112	5.10E-06	0.00152029	0.001336662	10211/1028	12
BP	GO:0032609	type II i	17/744	124/18986	0.137096774	3.49854752	5.637156979	6.91E-06	0.001830209	0.001609148	3123/11006	17
BP	GO:0032649	regulatic	17/744	124/18986	0.137096774	3.49854752	5.637156979	6.91E-06	0.001830209	0.001609148	3123/11006	17
BP	GO:1903039	positive	28/744	284/18986	0.098591549	2.515939724	5.198195633	7.55E-06	0.001895374	0.001666442	3123/3122/	28
BP	GO:0002218	activatic	30/744	317/18986	0.094637224	2.415030019	5.130861592	8.15E-06	0.001945298	0.001710336	10211/2858	30

GO 通路注释表

ONTOLOGY: 表示 GO 的三个本体类别，分别是 BP（Biological Process，生物过程）、MF（Molecular Function，分子功能）和 CC（Cellular Component，细胞组分）；

ID: 富集到的 GO term 的唯一标识符；

Description: 对 GO term 的描述，解释该 term 的功能或作用；

GeneRatio: 富集到该 GO term 中的基因数目与输入基因列表总数的比值；

BgRatio: 背景基因集中富集到该 GO term 的基因数目与背景基因集总数的比值；

RichFactor: 富集因子，计算公式为 GeneRatio/BgRatio，表示输入基因集中富集到该 term

的基因比例相对于背景基因集的富集程度；

FoldEnrichment：富集倍数，计算公式为 GeneRatio/BgRatio，与 RichFactor 类似，表示输入基因集中富集到该 term 的基因比例相对于背景基因集的倍数；

zScore：基于超几何分布计算的 z 分数，表示富集到该 term 的基因数目与期望值的偏离程度；

pvalue：富集分析的 p 值，表示观察到的富集程度是否显著。p 值越小，富集越显著；

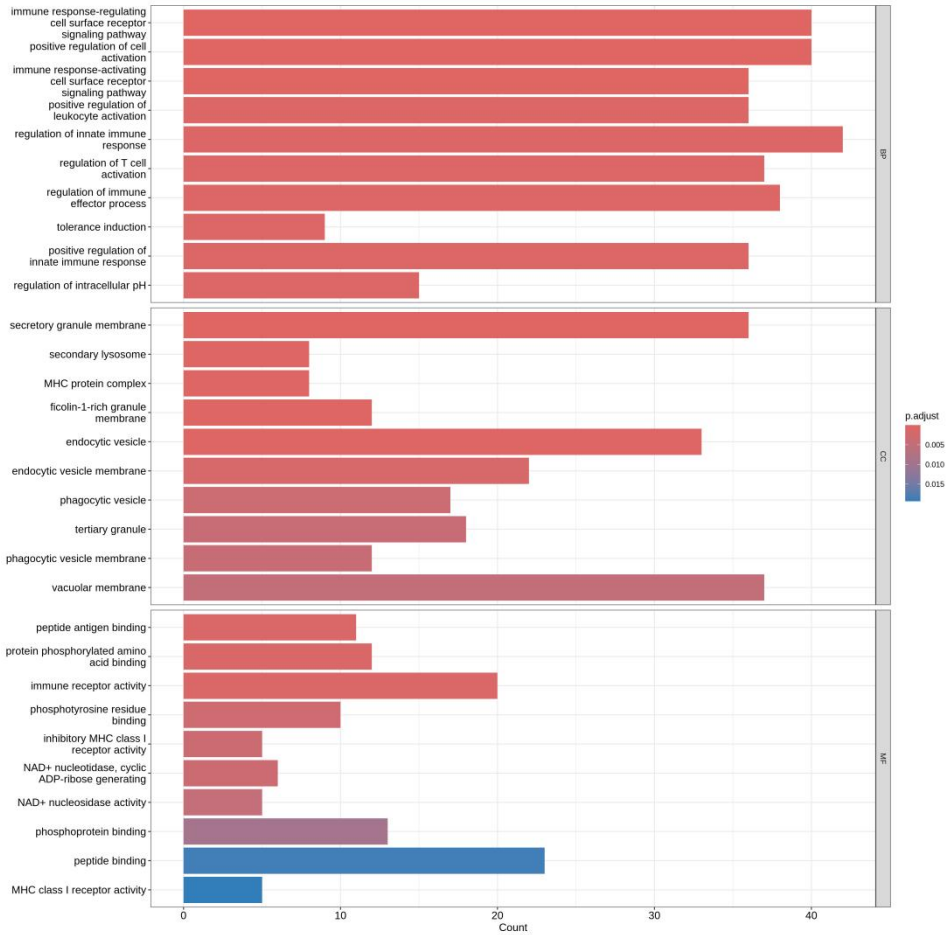
p.adjust：校正后的 p 值，通常使用 Benjamini-Hochberg(BH)等方法进行多重检验校正，以控制假阳性率；

qvalue：q 值，类似于校正后的 p 值，用于控制错误发现率（FDR）；

geneID：富集到该 GO term 的基因名称，多个基因通常用斜杠/分隔；

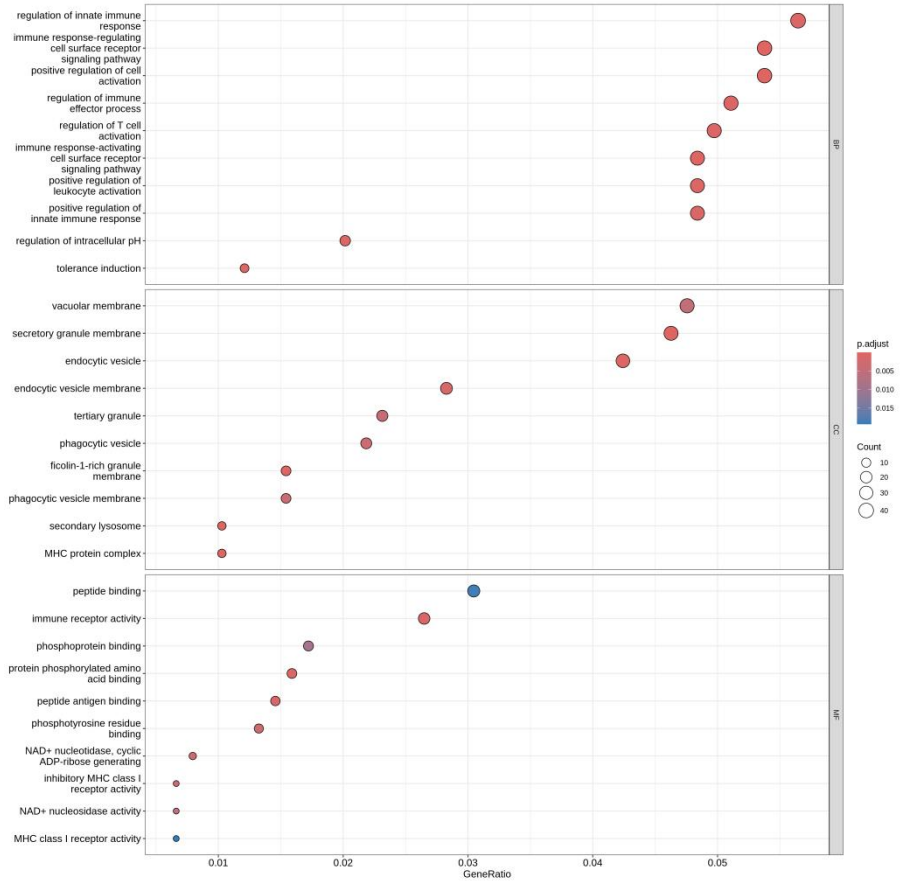
Count：富集到该 GO term 的基因数目

6.1.2 GO 显著通路分布



GO 显著富集通路分布柱形图

左侧纵坐标为通路名称，右侧纵坐标为各通路所在的本体类别，横坐标为各通路富集到的差异基因数量，颜色越深红表示该通路越显著。



GO 显著富集通路分布气泡图

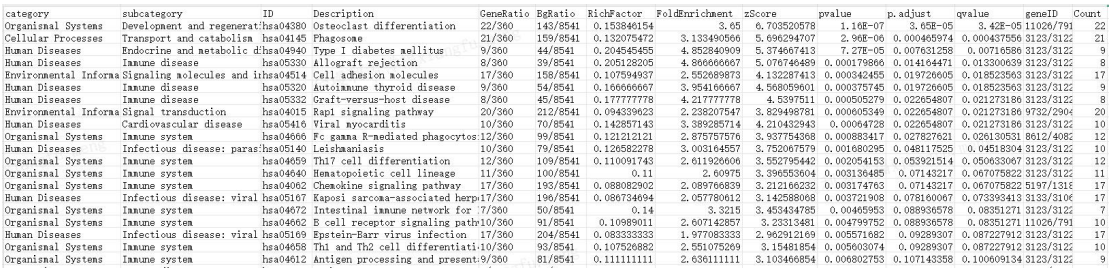
左侧纵坐标为通路名称，右侧纵坐标为各通路所在的本体类别，横坐标为富集到该 GO term 中的基因数目与输入基因列表总数的比值，气泡越大表示 GeneRaio 值越大，颜色越深红表示越显著。

6.2 KEGG

在 RNAseq 分析中，KEGG 富集分析（Kyoto Encyclopedia of Genes and Genomes Enrichment Analysis）用于揭示差异表达基因（DEGs）参与的生物学通路和代谢网络。

目录路径：RNASEQ\enrich\KEGG

6.2.1 KEGG 通路注释



category	subcategory	ID	Description	GeneRatio	BgRatio	RichFactor	FoldEnrichment	zScore	pvalue	p.adjust	qvalue	geneID	Count
Organismal Systems	Development and regenerat	hsa04380	Osteoclast differentiation	22/360	143/8541	0.153846154	3.65	6.703520578	1.16E-07	3.65E-05	3.42E-05	11026/791	22
Cellular Processes	Transport and catabolism	hsa04145	Phagosome	21/360	159/8541	0.132075472	3.133490566	5.696294707	2.96E-06	0.000465974	0.000437556	3123/3122	21
Human Diseases	Endocrine and metabolic d	hsa04940	Type I diabetes mellitus	9/360	44/8541	0.204545455	4.852840909	5.374667413	7.27E-05	0.007631258	0.00716586	3123/3122	9
Human Diseases	Immune disease	hsa05330	Allograft rejection	8/360	39/8541	0.205128205	4.866666667	5.076746489	0.000179866	0.014164471	0.013300639	3123/3122	8
Environmental Informa	Signaling molecules and i	hsa04514	Cell adhesion molecules	17/360	158/8541	0.107594937	2.552689873	4.132287413	0.000342455	0.019726605	0.018523563	3123/3122	17
Human Diseases	Immune disease	hsa05320	Autoimmune thyroid disease	9/360	54/8541	0.166666667	3.954166667	4.568059601	0.000375745	0.019726605	0.018523563	3123/3122	9
Human Diseases	Immune disease	hsa05332	Graft-versus-host disease	8/360	45/8541	0.177777778	4.217777778	4.5397511	0.000505279	0.022654807	0.021273186	3123/3122	8
Environmental Informa	Signal transduction	hsa04015	Rap1 signaling pathway	20/360	212/8541	0.094339623	2.238207547	3.829498781	0.000605349	0.022654807	0.021273186	9732/2904	20
Human Diseases	Cardiovascular disease	hsa05416	Viral myocarditis	10/360	70/8541	0.142857143	3.389285714	4.210432943	0.00064728	0.022654807	0.021273186	3123/3122	10
Organismal Systems	Immune system	hsa04666	Fc gamma R-mediated phagocytos	12/360	99/8541	0.121212121	2.875757576	3.937754368	0.000883417	0.027827621	0.026130531	8612/4082	12
Human Diseases	Infectious disease: paras	hsa05140	Leishmaniasis	10/360	79/8541	0.126582278	3.003164557	3.752067579	0.001680295	0.048117525	0.04518304	3123/3122	10
Organismal Systems	Immune system	hsa04659	Th17 cell differentiation	12/360	109/8541	0.110091743	2.611926606	3.552795442	0.002054153	0.053921514	0.050633067	3123/3122	12
Organismal Systems	Immune system	hsa04640	Hematopoietic cell lineage	11/360	100/8541	0.11	2.60975	3.396553604	0.003136485	0.07143217	0.067075822	3123/3122	11
Organismal Systems	Immune system	hsa04062	Chemokine signaling pathway	17/360	193/8541	0.088082902	2.089766839	3.212166232	0.003174763	0.07143217	0.067075822	5197/1318	17
Human Diseases	Infectious disease: viral	hsa05167	Kaposi sarcoma-associated herp	17/360	196/8541	0.086734694	2.057780612	3.142588068	0.003721908	0.078160067	0.073393413	3133/3106	17
Organismal Systems	Immune system	hsa04672	Intestinal immune network for	7/360	50/8541	0.14	3.3215	3.453434785	0.00465953	0.088936578	0.08351271	3123/3122	7
Organismal Systems	Immune system	hsa04662	B cell receptor signaling path	10/360	91/8541	0.10989011	2.607142857	3.23313481	0.004799752	0.088936578	0.08351271	11026/791	10
Human Diseases	Infectious disease: viral	hsa05169	Epstein-Barr virus infection	17/360	204/8541	0.083333333	1.977083333	2.962912169	0.005571682	0.09289307	0.087227912	3123/3122	17
Organismal Systems	Immune system	hsa04658	Th1 and Th2 cell differentiati	10/360	93/8541	0.107526882	2.551075269	3.15481854	0.005603074	0.09289307	0.087227912	3123/3122	10
Organismal Systems	Immune system	hsa04612	Antigen processing and present	9/360	81/8541	0.111111111	2.636111111	3.103466854	0.006802753	0.107143358	0.100609134	3123/3122	9

KEGG 通路注释表

category: 表示 KEGG 数据库中的一级分类，即主要的生物学功能类别。KEGG 数据库将通路分为 7 大类，包括代谢（Metabolism）、遗传信息处理（Genetic Information Processing）、环境信息处理（Environmental Information Processing）、细胞过程（Cellular Processes）、生物体系统（Organismal Systems）、人类疾病（Human Diseases）和药物开发（Drug Development）；

subcategory: 表示 KEGG 数据库中的二级分类，即在一级分类下的更具体的功能类别

ID: 表示 KEGG 数据库中分配给特定通路的唯一标识符；

Description: 对 KEGG 通路的描述，解释该通路的功能或作用；

GeneRatio: 富集到该 KEGG 通路中的基因数目与输入基因列表总数的比值；

BgRatio: 背景基因集中富集到该 KEGG 通路的基因数目与背景基因集总数的比值；

RichFactor: 富集因子，计算公式为 GeneRatio/BgRatio，表示输入基因集中富集到该通路的基因比例相对于背景基因集的富集程度；

FoldEnrichment: 富集倍数，计算公式为 GeneRatio/BgRatio，与 RichFactor 类似，表示输入基因集中富集到该通路的基因比例相对于背景基因集的倍数；

zScore: 基于超几何分布计算的 z 分数，表示富集到该通路的基因数目与期望值的偏离程度；

pvalue: 富集分析的 p 值，表示观察到的富集程度是否显著。p 值越小，富集越显著；

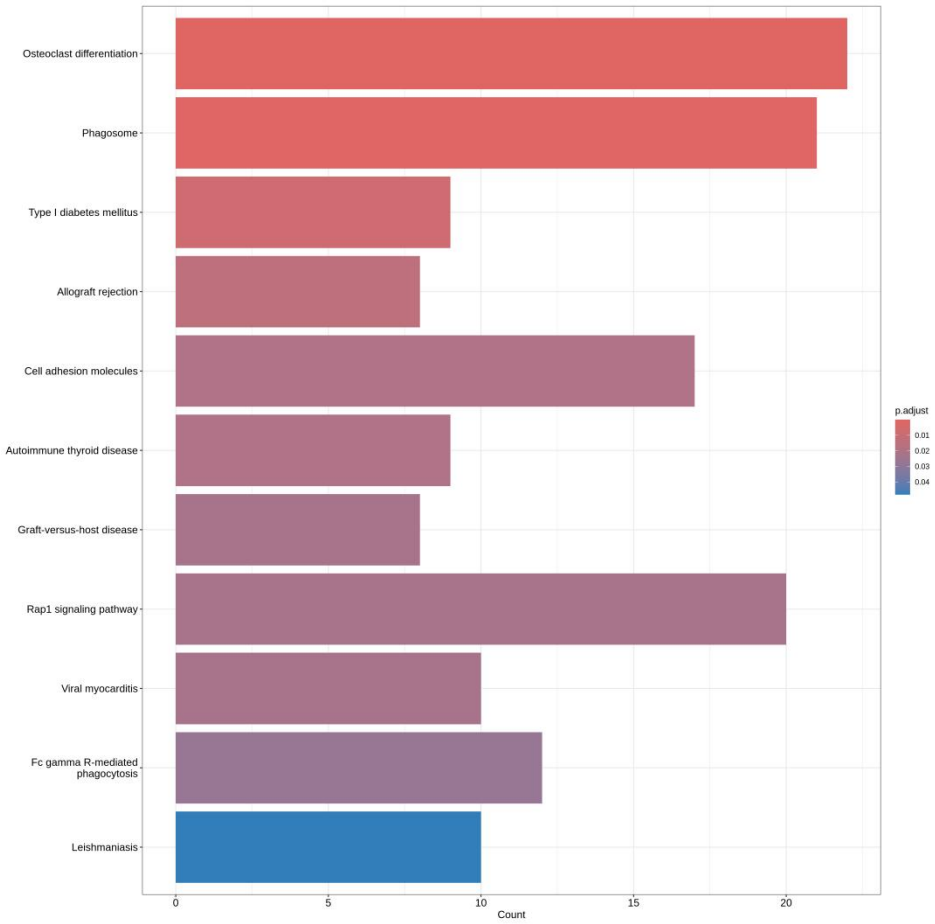
p.adjust: 校正后的 p 值，通常使用 Benjamini-Hochberg(BH)等方法进行多重检验校正，以控制假阳性率；

qvalue: q 值，类似于校正后的 p 值，用于控制错误发现率（FDR）；

geneID: 富集到该 KEGG 通路的基因名称，多个基因通常用斜杠/分隔；

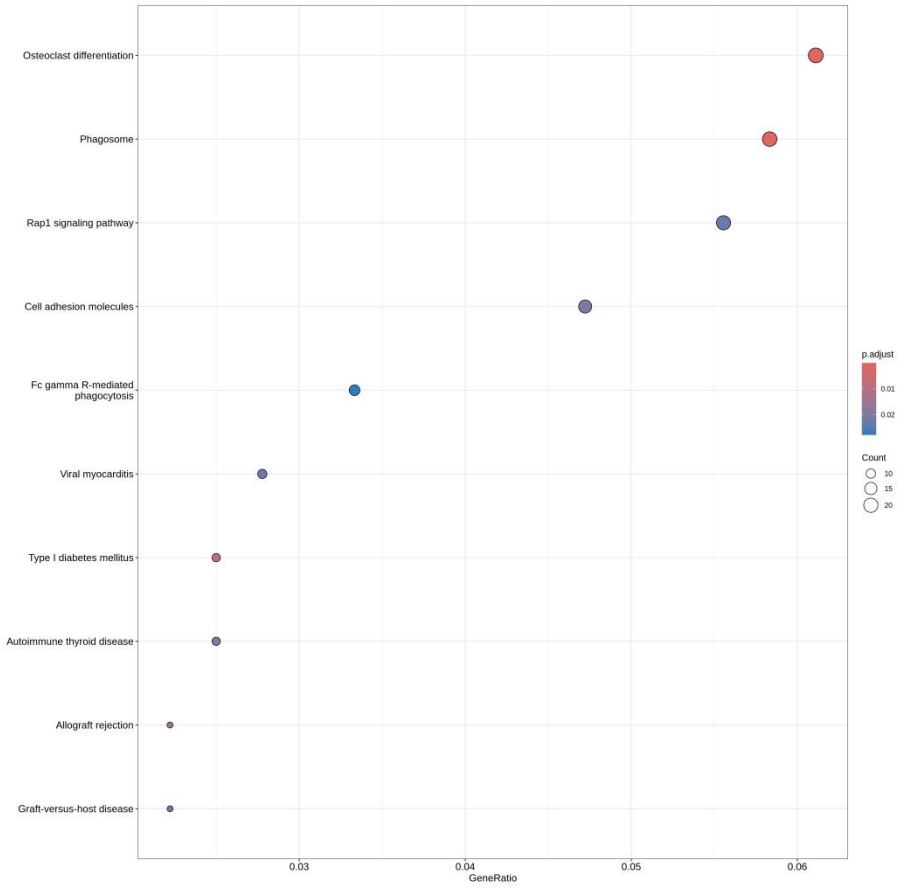
Count: 富集到该 KEGG 通路的基因数目；

6.2.2 KEGG 显著通路分布



KEGG 显著富集通路分布柱形图

纵坐标为通路名称，横坐标为各通路富集到的差异基因数量，颜色越深红表示越显著

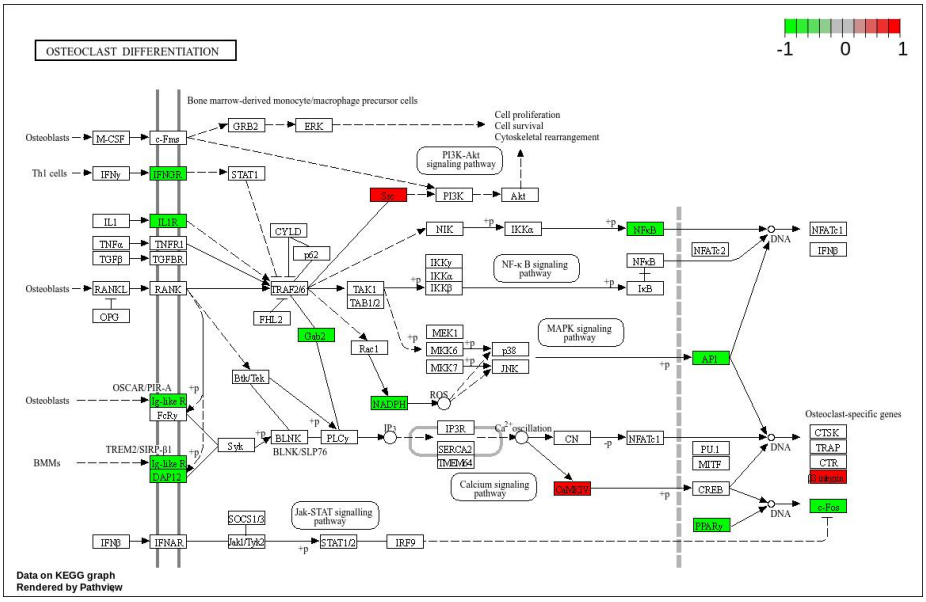


显著富集通路分布气泡图

纵坐标为通路名称，横坐标为富集到该 KEGG 通路中的基因数目与输入基因列表总数的比值，气泡越大表示 GeneRaio 值越大，颜色越深红表示越显著。

6.2.3 KEGG 通路图

仅输出最显著的通路。通路图绘制使用 R pathview 包，该工具能够自动下载通路图数据，解析数据文件，并将用户数据映射到通路，最终生成带有映射数据的通路图。



KEGG 最显著通路图

图中红色表示上调基因,绿色表示下调基因