

# RNA-seq 检测报告

2018-08-08

Pan

## 一、 样本信息

本次分析使用数据集为 NCBI SRA 编号 [SRP029245](#)。实验敲除了 HEK293T / 17 细胞中的 PRC2 必需亚基 SUZ12，用以检测 PRC2 与转录活性基因的关联。

表 1-1 样本信息

SRA 编号	类型	上传时间	其他信息	检测平台
SRR957677	RNA-seq	2013-08-27	siCtrl_BiolRep1	Illumina
SRR957678	RNA-seq	2013-08-27	siCtrl_BiolRep2	Illumina
SRR957679	RNA-seq	2013-08-27	siSUZ12_BiolRep1	Illumina
SRR957680	RNA-seq	2013-08-27	siSUZ12_BiolRep2	Illumina

## 二、 数据质量统计

### 2.1 测序数据情况汇总

使用 ReSeqTools (He W , et al.) 进行统计。

表 2-1 测序情况汇总

SRA ID	Sample Name	Raw reads	Bases	Q20(%)	Q30(%)	GC(%)
SRR957677	siCtrl_1	20803937	1040196850	97.74	93.83	47.52
SRR957678	siCtrl_2	8828013	441400650	97.41	93.44	46.97
SRR957679	siSUZ12_1	19909740	995487000	97.65	93.62	48.60
SRR957680	siSUZ12_2	24231941	1211597050	97.68	93.68	48.04

注：

- (1) Sample Name：对应样本的命名；
- (2) Raw reads：原始序列总数；
- (3) Bases：原始序列碱基数；
- (4) Q20(%)：计算 phred 数值大于 20 的碱基占总碱基数的比例；
- (5) Q30(%)：计算 phred 数值大于 30 的碱基占总碱基数的比例；
- (6) GC(%)：计算 G 和 C 的数量占总碱基数的比例。

2.2 测序质量分布图

使用 FastQC (Andrews S, et al.) 以及 MultiQC (Ewels P , et al.) 统计。



图 2-1 测序质量分布

横坐标为 phred 得分，纵坐标为 reads 数目。其中，siCtrl\_1 为红色、siCtrl\_2 为蓝色、siSUZ12\_1 为绿色、siSUZ12\_2 为紫色。可以看出，绝大多数 reads 的 phred 得分在 30 以上。

2.3 GC 含量图

使用 FastQC (Andrews S, et al.) 以及 MultiQC (Ewels P , et al.) 统计。横坐标为碱基在 reads 中的位置，纵坐标为碱基含量。

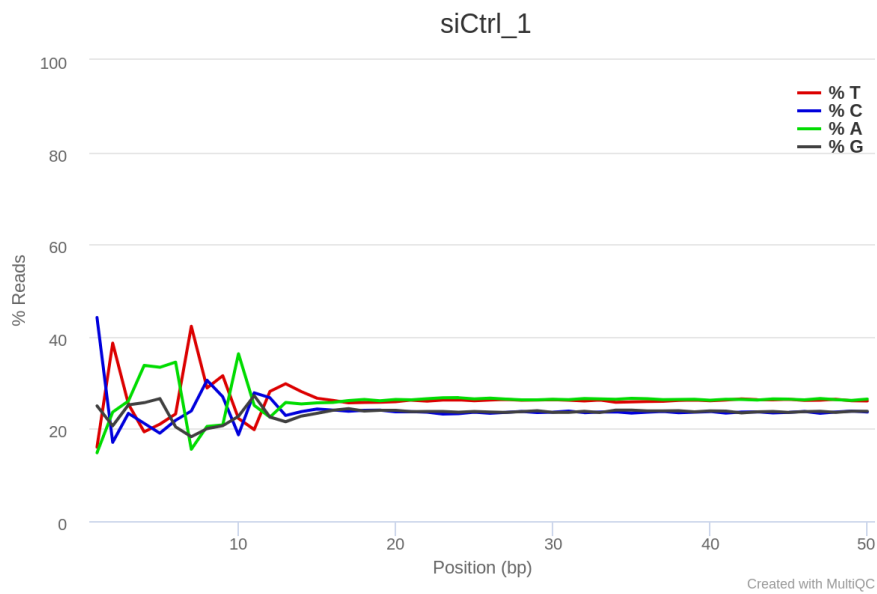


图 2-2 siCtrl\_1 GC 含量图

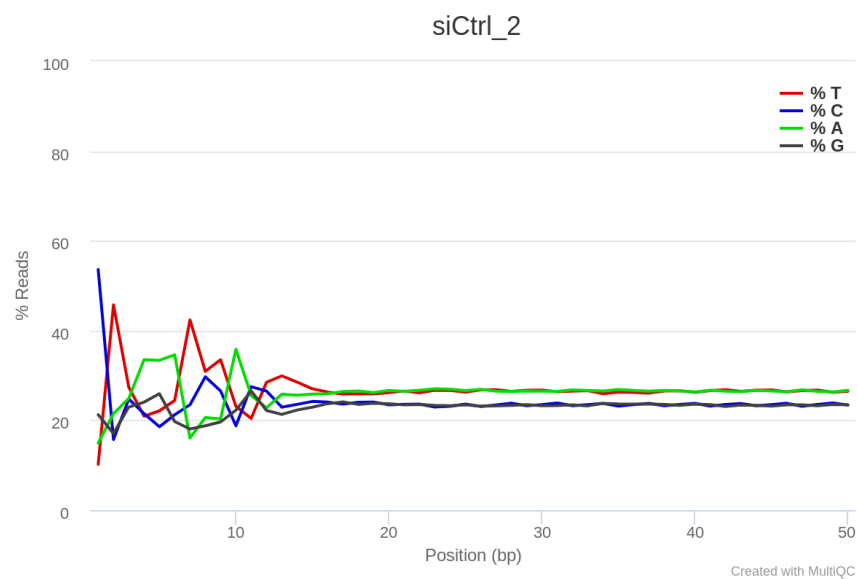


图 2-3 siCtrl\_2 GC 含量图

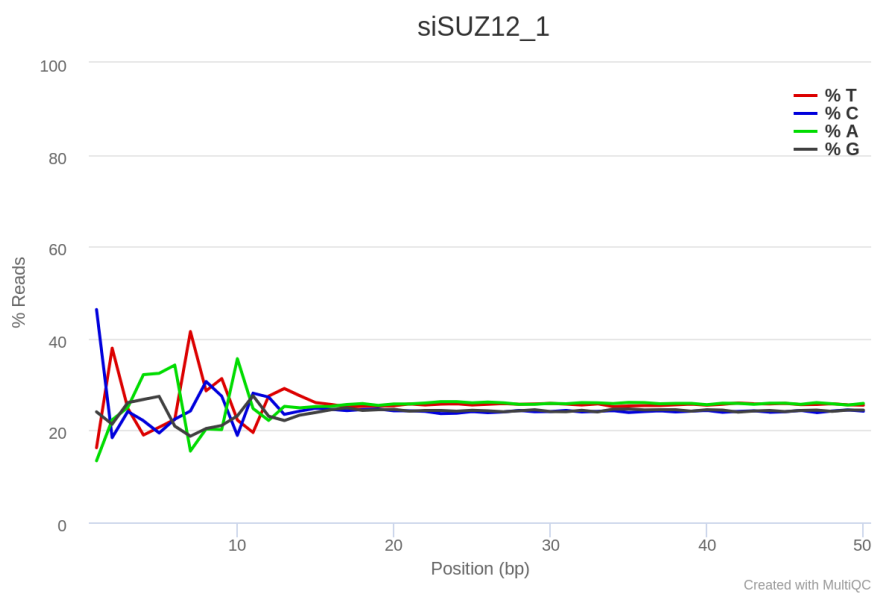


图 2-4 siSUZ12\_1 GC 含量图

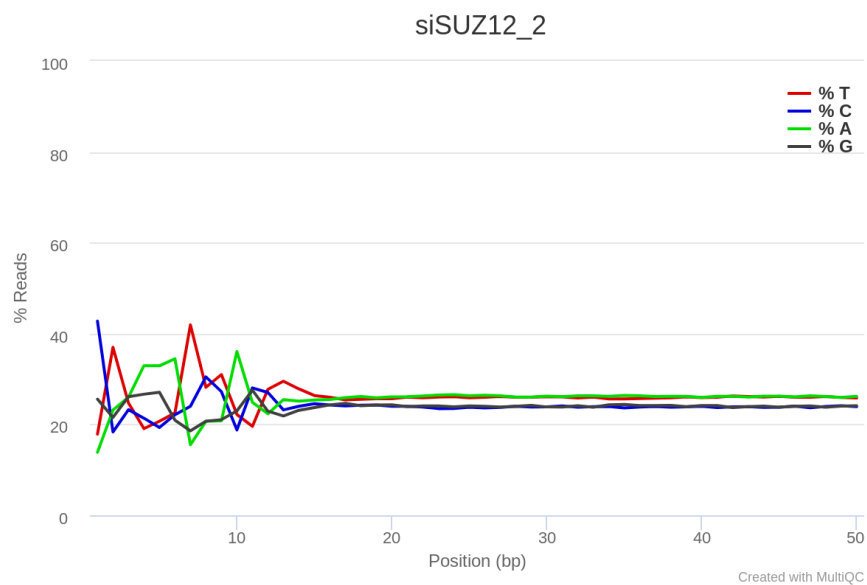


图 2-5 siSUZ12\_2 GC 含量图

三、差异分析

将测序数据通过 Hisat2（Kim D , et al.）与参考基因组（GRCh38）进行比对，比对结果使用 samtools（Li H）进行排序以及转换为 bam 格式文件。再使用 featureCounts（Liao Y , et al.）对 bam 文件与数据库（Homo\_sapiens.GRCh38.92.chr\_patch\_hapl\_scaff.gtf）进行比对注释以及统计 reads 数。

3.1 DESeq2 分析结果

使用 DESeq2（Love MI , et al.）对统计结果进行处理得到基因表达结果。下列为 DESeq2 分析结果。提取其中 padj<0.1 的基因，本次检测到差异较显著的基因 88 个（只显示前 10 行）。

表 3-1 DESeq2 分析结果

GeneName	BaseMean	Log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000178691	1020.461	-2.83048	0.225455	-12.5545	3.75E-36	3.70E-32
ENSG00000135535	2433	-1.22818	0.183795	-6.68233	2.35E-11	1.16E-07
ENSG00000164172	542.4684	-1.30612	0.205653	-6.35111	2.14E-10	7.02E-07
ENSG00000172239	493.5468	-1.31923	0.214485	-6.15067	7.72E-10	1.90E-06
ENSG00000196504	3716.472	-1.10346	0.202081	-5.46047	4.75E-08	9.36E-05
ENSG00000163848	635.4076	-1.14847	0.219016	-5.24379	1.57E-07	0.000221
ENSG00000173905	1111.492	-1.12328	0.214041	-5.24796	1.54E-07	0.000221
ENSG00000187772	1434.25	-1.26702	0.24713	-5.12691	2.95E-07	0.000322
ENSG00000141425	2648.092	-0.91524	0.178091	-5.13917	2.76E-07	0.000322
ENSG00000077549	1207.774	-0.91644	0.187561	-4.88607	1.03E-06	0.001013

注：

- (1) GeneName：基因名称（ENSEMBL ID）；
- (2) BaseMean：所有样本经过校正的平均 reads 数；
- (3) log2FoldChange：取 log2 后的表达量差异；
- (4) lfcSE：log2FoldChange 标准误差值；
- (5) stat：log2FoldChange 除以 lfcSE，用于计算 pvalue；
- (6) pvalue：统计学差异显著性检验指标；
- (7) padj：校正后的 pvalue，padj 越小，表示基因表达差异越显著。BaseMean 值较低，padj 值将设置为 NA。

### 3.1.1 Heatmap

对 DESeq2 结果绘制热图。热图 heatmap 可以实现基因表达模式可视化的需求。从这里可以看到这 4 个样本的表达差异。

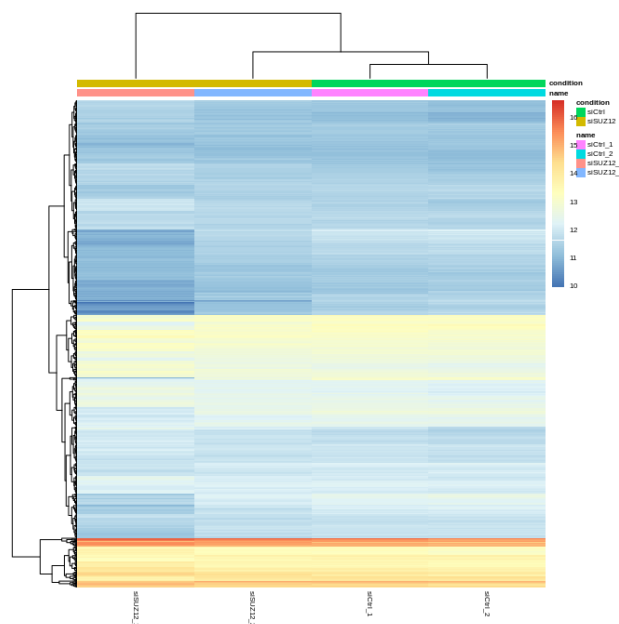


图 3-1 heatmap

图中每个小方格表示每个基因，其颜色表示该基因表达量大小，表达量越大红色越深。每列表示每个样品中所有基因的表达量情况。上方树形图表示对来自不同实验分组的不同样品的聚类分析结果，左侧树状图表示对来自不同样本的不同基因的聚类分析结果。

### 3.1.2 MA 图

在 DESeq2 结果中，MA 图表示经过标准化的 counts 数与  $\log_2\text{FoldChange}$  的关系。其中红色点为  $\text{padj}$  值（校正后 p 值）小于 0.1 的点。

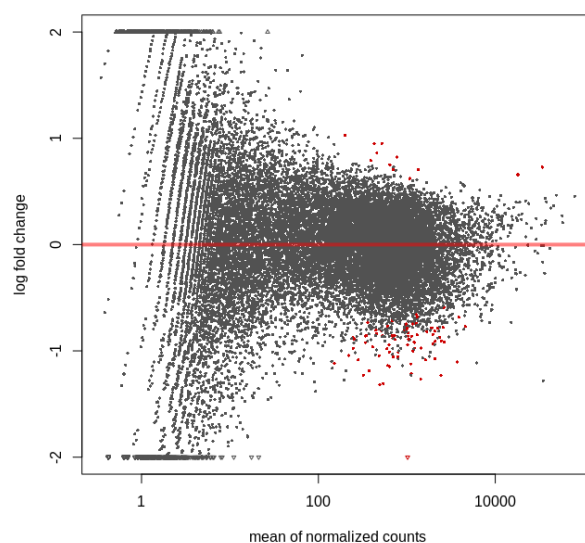


图 3-2 MA 图

3.2 edgeR 分析结果

使用 edgeR (Robinson MD , et al.) 对统计结果进行处理得到基因表达结果。下列为 edgeR 分析结果。提取其中 padj<0.1 的基因，本次检测到差异较显著的基因 83 个（只显示前 10 行）。

表 3-2 edgeR 分析结果

GeneName	logFC	LogCPM	F	PValue	FDR
ENSG000000178691	2.815727	6.382903	191.7282	5.50E-12	1.26E-07
ENSG000000172239	1.306663	5.333683	55.5025	9.03E-08	0.000517
ENSG000000164172	1.294146	5.474934	56.46352	7.78E-08	0.000517
ENSG000000135535	1.216325	7.651307	55.9175	8.46E-08	0.000517
ENSG000000163848	1.134842	5.707212	38.87853	1.68E-06	0.007711
ENSG000000196504	1.090359	8.264805	40.11477	2.16E-06	0.008256
ENSG000000173905	1.110058	6.520278	36.65775	3.48E-06	0.011402
ENSG000000213626	-1.63885	2.041093	33.05166	5.68E-06	0.016274
ENSG000000141425	0.903635	7.775512	32.49211	6.43E-06	0.016368
ENSG000000075618	-0.96067	5.455631	31.84296	7.43E-06	0.017028

- 注：
- (1) GeneName: 基因名称 (ENSEMBL ID);
  - (2) logFC: 取 log2 后的表达量差异;
  - (3) LogCPM: 总体对数平均值;
  - (4) F: 拟合方程的显著性, F 越大, 表示方程越显著, 拟合程度也就越好;
  - (5) PValue: 统计学差异显著性检验指标;
  - (6) FDR: 校正后的 pvalue, FDR 越小, 表示基因表达差异越显著。



3.2.1 MD 图

在 edgeR 结果中，MD 图表示 logCPM 与 logFoldChange 的关系。其中红色点为上调基因，绿色点为下调基因。

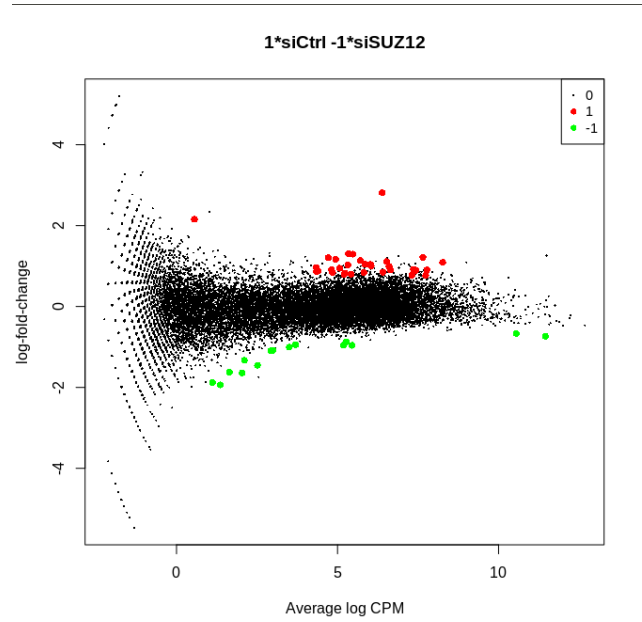


图 3-3 MD 图

3.3 结果整合与总结

总结 DESeq2 与 edgeR 结果，共得到差异较显著基因 118 个。下表为统计得到的差异基因（只列出其中 30 个）。

表 3-3 差异基因结果

ENSG00000178691	ENSG00000163848	ENSG00000152818	ENSG00000156650	ENSG00000184009	ENSG00000181904
ENSG00000135535	ENSG00000173905	ENSG00000100731	ENSG00000128512	ENSG00000170903	ENSG00000114520
ENSG00000164172	ENSG00000187772	ENSG00000054598	ENSG00000117335	ENSG00000011405	ENSG00000071794
ENSG00000172239	ENSG00000141425	ENSG00000075618	ENSG00000163376	ENSG00000064666	ENSG00000114978
ENSG00000196504	ENSG00000077549	ENSG00000196914	ENSG00000140526	ENSG00000144357	ENSG00000117620
ENSG00000178691	ENSG00000163848	ENSG00000152818	ENSG00000156650	ENSG00000184009	ENSG00000181904

四、GO 富集分析

对差异基因进行富集分析。其中最显著的是 early endosome 通路（GO:0005769）。

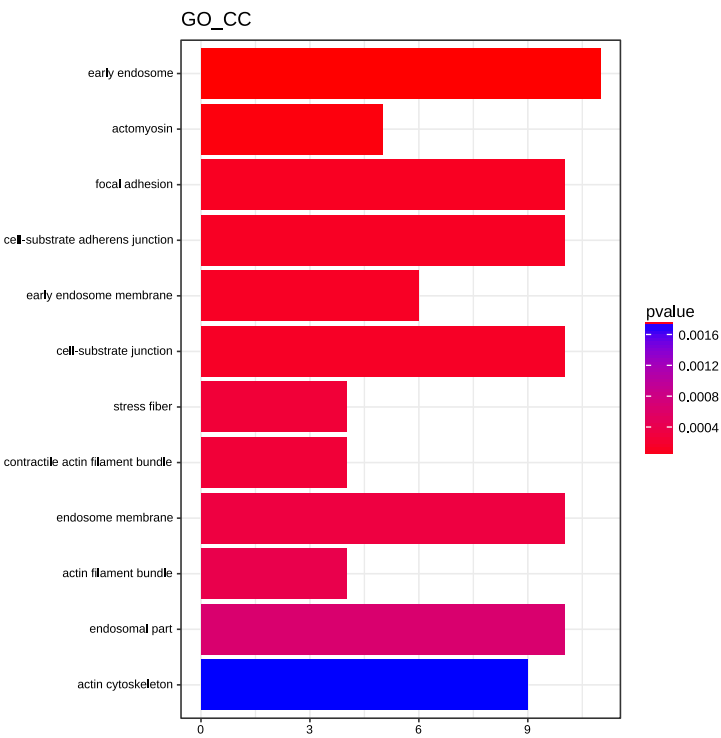


图 4-1 GO 细胞组件富集条形图

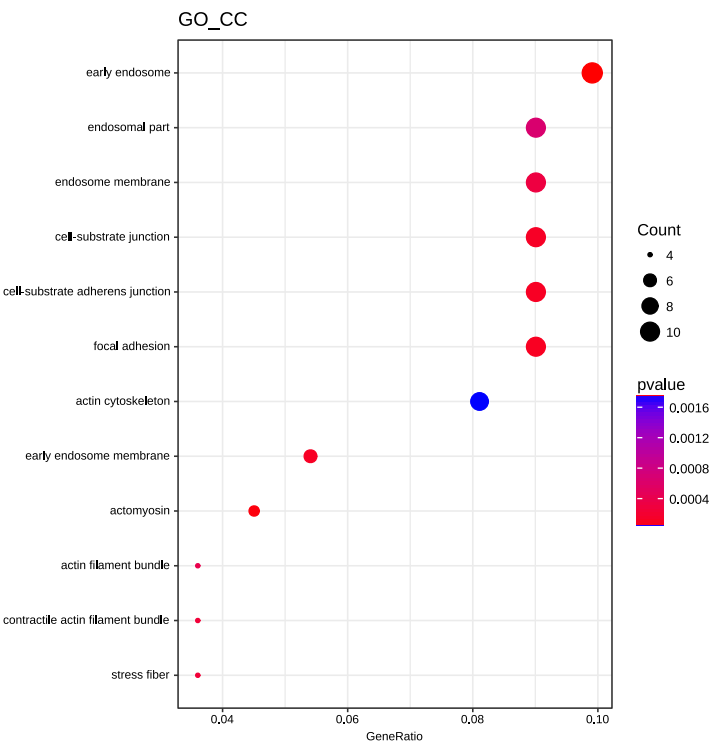


图 4-2 GO 细胞组件富集散点图

## 五、参考文献

1. Promiscuous RNA binding by Polycomb repressive complex 2. Davidovich C , et al. Nat Struct Mol Biol. 2013 Nov;20(11):1250-7. doi: 10.1038/nsmb.2679. Epub 2013 Sep 29.
2. ReSeqTools: an integrated toolkit for large-scale next-generation sequencing based resequencing analysis. He W , et al. Genet Mol Res. 2013 Dec 4;12(4):6275-83. doi: 10.4238/2013.December.4.15.
3. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Leggett RM , et al. Front Genet. 2013 Dec 17;4:288. doi: 10.3389/fgene.2013.00288.
4. MultiQC: summarize analysis results for multiple tools and samples in a single report. Ewels P , et al. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16.
5. HISAT: a fast spliced aligner with low memory requirements. Kim D , et al. Nat Methods. 2015 Apr;12(4):357-60. doi: 10.1038/nmeth.3317. Epub 2015 Mar 9.
6. The Sequence Alignment/Map format and SAMtools. Li H , et al. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8.
7. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Liao Y , et al. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13.
8. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Love MI , et al. Genome Biol. 2014;15(12):550.
9. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Robinson MD , et al. Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11.
10. clusterProfiler: an R package for comparing biological themes among gene clusters. Yu G , et al. OMICS. 2012 May;16(5):284-7. doi: 10.1089/omi.2011.0118. Epub 2012 Mar 28.