**数据集**

Patel AP, Tirosh I, Trombetta JJ, Shalek AK et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 2014 Jun (<http://science.sciencemag.org/content/344/6190/1396.full)>

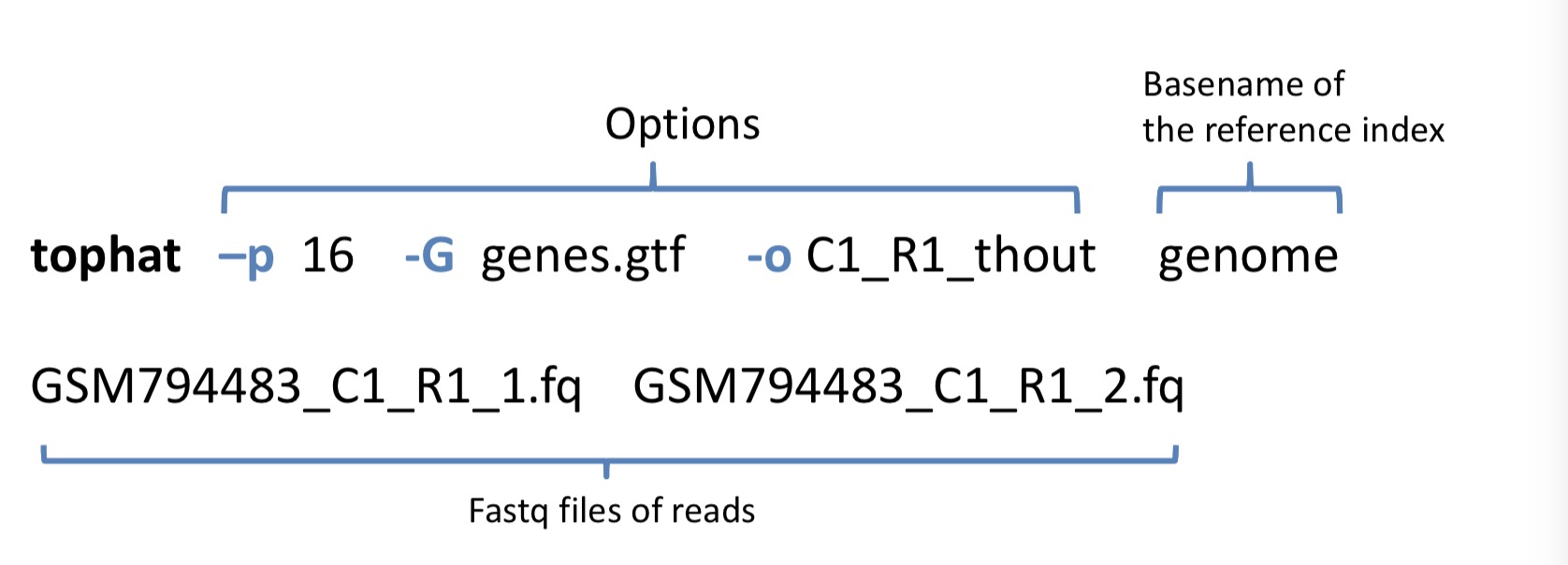
**GEO下载地址**

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57872>

**预处理代码（李翔宇）**

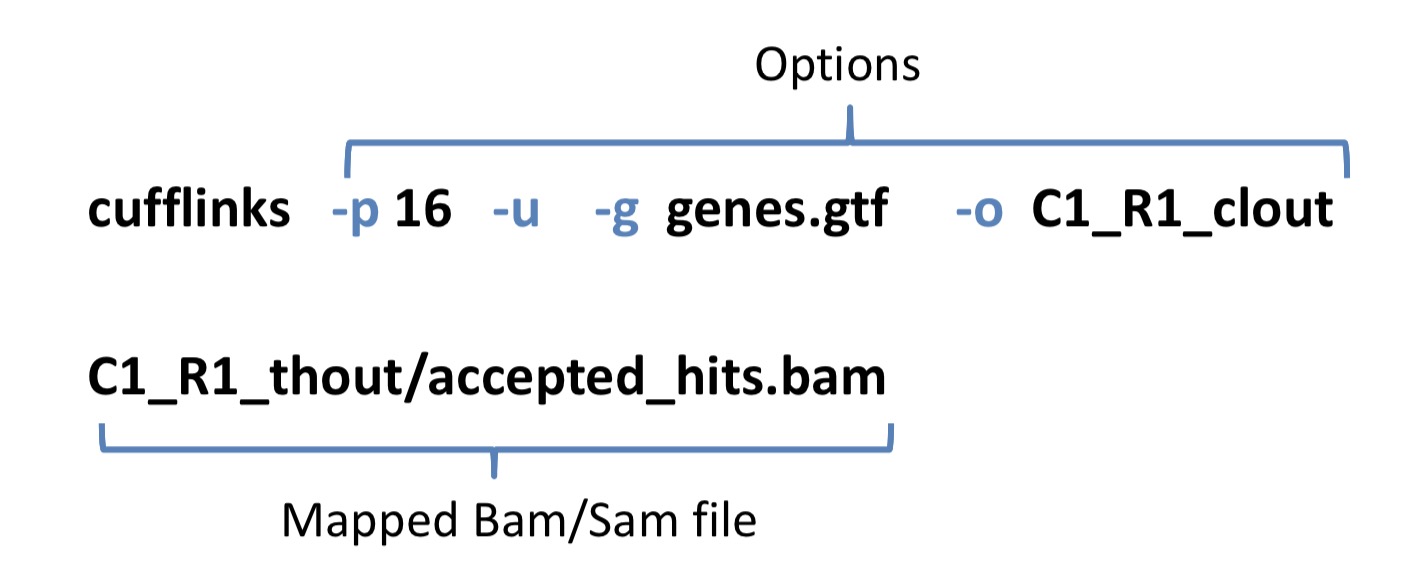
# mapping

tophat -p 8 -G mouse.gtf -o C1\_R1\_thout /bowtie2\_index/mm9 C1\_R1\_1.fq C1\_R1\_2.fq



# 把mapping结果组装成转录本，并估计转录本丰度

cufflinks -p 8 -G mouse.gtf -o C1\_R1\_clout C1\_R1\_thout/accepted\_hits.bam



# 转换sam格式为bam格式

samtools view -o accepted\_hits.sam accepted\_hits.bam

# 计算基因表达量

python -m HTSeq.scripts.count -s no accepted\_hits.sam mouse.gtf > C1\_R1.count

计算出matrix后的分析步骤可以参考以下站点以及各R包的tutorial

<http://hms-dbmi.github.io/scw/heterogeneity.html>

预处理代码（苗准）

<https://paste.ubuntu.com/p/RZhVxTjVKR/>

# QC:

fastqc --extract ${i##${data\_dir}/}\_1.fastq ${i##${data\_dir}/}\_2.fastq;

# mapping:

# Tophat

tophat -p 30 -o ./tophat\_out\_GTF -G /home/miaozhun/data/genome/mm9\_transcriptome/mm9\_transcriptome.GTF --transcriptome-index /home/miaozhun/data/genome/bowtie\_index\_mm9\_transcriptome/mm9\_transcriptome /home/miaozhun/data/genome/bowtie2\_index\_mm9/musMusculusMasked\_mm9 ${i##${data\_dir}/}\_1.fastq ${i##${data\_dir}/}\_2.fastq;

# MapSplice (Tophat以外的另一种mapping工具，可供选择)

python /home/miaozhun/share/tool/MapSplice-v2.1.8/mapsplice.py -p 30 -c /home/miaozhun/data/genome/mouse\_mm9/chromFaMasked/ -x /home/miaozhun/data/genome/bowtie\_index\_mm9/musMusculusMasked\_mm9 -o ./mapsplice\_out\_Masked/ -1 ${i##${data\_dir}/}\_1.fastq -2 ${i##${data\_dir}/}\_2.fastq;

# Quantification:

# HTSeq

python -m HTSeq.scripts.count -s no -f bam -r name accepted\_hits\_namesorted\_SRX324${i}.bam /home/miaozhun/data/genome/mm9\_transcriptome/mm9\_EnsemblGenes.GTF > /data/miaozhun/analysis/readCounts/GSE48968/biological/readCounts\_EnsemblGenes\_SRX324${i}.txt

差异表达分析scde包 R代码（苗准）

<https://paste.ubuntu.com/p/xrWjHTkH3m/>

load(paste0(geneNum, "\_Genes\_", sampleNum, "\_Cells\_SplatDefault.Rdata"))

groups <- factor(rep(c("A", "B"), each=sampleNum))

names(groups) <- colnames(counts)

counts<-apply(counts,2,function(x) {storage.mode(x) <- 'integer'; x})

cd <- clean.counts(counts, min.lib.size=100, min.reads = 0, min.detected = 0)

timeUsed\_1\_1 <- system.time(o.ifm <- scde.error.models(counts = cd, groups = groups, n.cores = 8, min.size.entries = 100, threshold.segmentation = TRUE, save.crossfit.plots = FALSE, save.model.plots = FALSE, verbose = 1))

timeUsed\_1\_2 <- system.time(o.prior <- scde.expression.prior(models = o.ifm, counts = cd, length.out = 400, show.plot = FALSE))

timeUsed\_1\_3 <- system.time(result <- scde.expression.difference(o.ifm, cd, o.prior, groups = groups, n.randomizations = 100, n.cores = 8, verbose = 1))

timeUsed\_1 <- timeUsed\_1\_1 + timeUsed\_1\_2 + timeUsed\_1\_3