Salmon

[Salmon][https://salmon.readthedocs.io/en/latest/salmon.html]: a wicked-fast 转录本量化软件

salmon需要一套靶向转录本(参考转录本或de-nove组装)进行量；总之，所有需要的就是fasta格式参考转录本和fastq/fasta reads文件，同时也可以处理提前比对好的文件(sam/bam)

salmon的比对模型需要通过两阶段实现：indexing和quantification。索引步骤独立于reads文件，仅对参考转录本进行索引;定量过程对reads进行比对定量

推荐针对decoy-aware transcriptom采用selective alignment,来缓解reads潜在对错误比对

随着salmon v0。13.1,推荐都采用selective alignment：##--validateMappings

注意,若reads或比对文件针对靶向转录本不是随机顺序,需要randomize/shuffle之后在使用salmon定量

mapping-based mode

推荐使用selective alignment，因此使用脚本generateDecoyTranscriptome.sh脚本构建decoy-aware trasncriptom 文件

salmon index -t transcripts.fa -i transcripts_index -k 31

含两种定量模式，一种是根据基于比对文件(sam/bam)的方式，另一种是根据read定量方式

If you provide salmon with alignments '-a [--alignments]' then the alignment-based algorithm will be used, otherwise the algorithm for quantifying from raw reads will be used

根据reads定量

-l/--libType librarytype,针对alignments文件自动检测测序文件的reads类型: -l A

https://salmon.readthedocs.io/en/latest/library_type.html#fraglibtype

-l/--libType 分三部分:

1. 相对方向:I=inward;O=outward;M=matching
2. read文库是否为stranded/unstranded：S=stranded;U=unstranded
3. 若为unstranded，无需第三部分值;该部分指定strand from which the readoriginated:F=read 1 源于正链；R=read 1 源于负链

-i/--index salmon index

-r/--unmateReads 包含为成对reads的文件列表,例如单端reads

-1 --mates1 包含#1 mates文件；-2/--mates2 包含#2 mates文件

-o/--output 输出定量目录

-g/--geneMap 包含转录本比对到genes的文件，此时输出会包含quant.genes.sf文件，包含基因水平的丰度评估, 该文件可谓gtf文件或tab分隔的格式文件

--discardOrphansQuasi 舍弃仅单端比对上的双短reads

--validateMappings [Quasi-mapping mode only]: 使用alignment-based verification来验证比对

--writeOrphanLins 输出指向orphan reads的转录本

--writeUnmappedNames 输出没有比对上的read到unmapped_names.txt;其中单端read后缀为u, unmapped；双端read：u表示一对read都没比对上；m1指read1没比对；m2指read2没比对；m12表read1/read2比对到了不同转录本

```
salmon quant –i transcripts_index –l IU –1 read1.fq.gz –2 read2.fq.gz –o transcripts_quant ––writeUnmappedNames
```

根据必读文件定量

期待reads是直接比对到了转录本(RSEM, eXpress等), 而不是比对到了基因组(Cufflinks);若比对到了基因组，需要将SAM/BAM转回FASTA/Q文件，然后使用lightweight-alignment-based mode；或将其比对到转录本；或着使用sam-xlate将BAM文件到基因组坐标转换为转录本坐标

That is, Salmon expects that the reads have been aligned directly to the transcriptome (like RSEM, eXpress, etc.) rather than to the genome (as does, e.g. Cufflinks).若

提供aln.bam文件和需要定量的转录组序列文件transcripts.fa

-a 空格分开多个bam/sam文件

```
salmon quant –t transcripts.fa –l A –a aln.bam –o salmon_quant ––writeUnmappedNames
```

输出文件quant.sf

quant.sf文件共5列：Name, Length, EffectiveLength, TPM, NumReads

Name, target transcripts 名称

Length, target transcript 长度，即多少个核苷酸

EffectiveLength, target transcript 计算的有效长度:It takes into account all factors being modeled that will effect the probability of sampling fragments from this transcript, including the fragment length distribution and sequence-specific and gc-fragment bias (if they are being modeled)

TPM, 估计转录本的表达量

NumReads, 估计比对到每个转录本的reads数

TPM的计算公式：

$$TPM_i = (N_i/L_i) * 1000000/sum(N_i/L_i + ... + N_m/L_m)$$

N_i：mapping到基因i上的read数；
L_i：基因i的外显子长度的总和。
在一个样本中一个基因的TPM：先对每个基因的read数用基因的长度进行校正，之后再用校正后的这个基因read数(N_i/L_i)与校正后的这个样本的所有read数（$sum(N_i/L_i + ... + N_m/L_m)$）求商。由此可知，TPM概括了基因的长度、表达量和基因数目。TPM可以用于同一物种不同组织间的比较，因为sum值总是唯一的。

[补充][http://blog.sciencenet.cn/blog-1113671-1038659.html]：

FPKM (推荐软件，cufflinks/Stringtie) 和RPKM (推荐软件，Range/Deseq) 的计算方法基本一致，公式如下(外显子的表达)：

RPKM= total exon reads/ (mapped reads (Millions) * exon length(KB))

你可以用这个公式计算基因，外显子，转录本的表达，这里以基因的表达为例进行说明：

total exon reads：某个样本mapping到特定基因的外显子上的所有的reads

mapped reads (Millions) :某个样本的所有reads总和

exon length(KB)：某个基因的长度（外显子的长度的总和，以KB为单位）

在一个样本中一个基因的RPKM等于落在这个基因上的总的read数(total exon reads)与这个样本的总read数(mapped reads (Millions))和基因长度(exon length(KB)) 的乘积的比值。

而RPM的计算公式：

RPM=total exon reads / mapped reads (Millions)

RPM per gene is calculated as the number of reads per gene divided by the number ofsingle-mapping reads per sample library times one million

##

而TPM (推荐软件，RSEM/Stringtie) 的计算公式：

$TPM_i=(N_i/L_i)*1000000/sum(N_i/L_i+.....+ N_m/L_m)$

N_i：mapping到基因i上的read数； L_i：基因i的外显子长度的总和

在一个样本中一个基因的TPM：先对每个基因的read数用基因的长度进行校正，之后再用校正后的这个基因read数(N_i/L_i)与校正后的这个样本的所有read数（$sum(N_i/L_i+.....+ N_m/L_m)$）求商。