RNA-seq analysis

ZTY 4.12

如何界定RNA-seq

- RNA-protein
- scRNA-seq
- RNA-RNA
- RNA-structure

‘typical’ RNA-seq

参考文献

REVIEWS

RNA sequencing: the teenage years

Rory Stark[1], *Marta Grzelak*[1] *and James Hadfield*[2]*

Abstract | Over the past decade, RNA sequencing (RNA-seq) has become an indispensable tool for transcriptome-wide analysis of differential gene expression and differential splicing of mRNAs. However, as next-generation sequencing technologies have developed, so too has RNA-seq. Now, RNA-seq methods are available for studying many different aspects of RNA biology, including single-cell gene expression, translation (the translatome) and RNA structure (the structurome). Exciting new applications are being explored, such as spatial transcriptomics (spatialomics). Together with new long-read and direct RNA-seq technologies and better computational tools for data analysis, innovations in RNA-seq are contributing to a fuller understanding of RNA biology, from questions such as when and where transcription occurs to the folding and intermolecular interactions that govern RNA function.

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

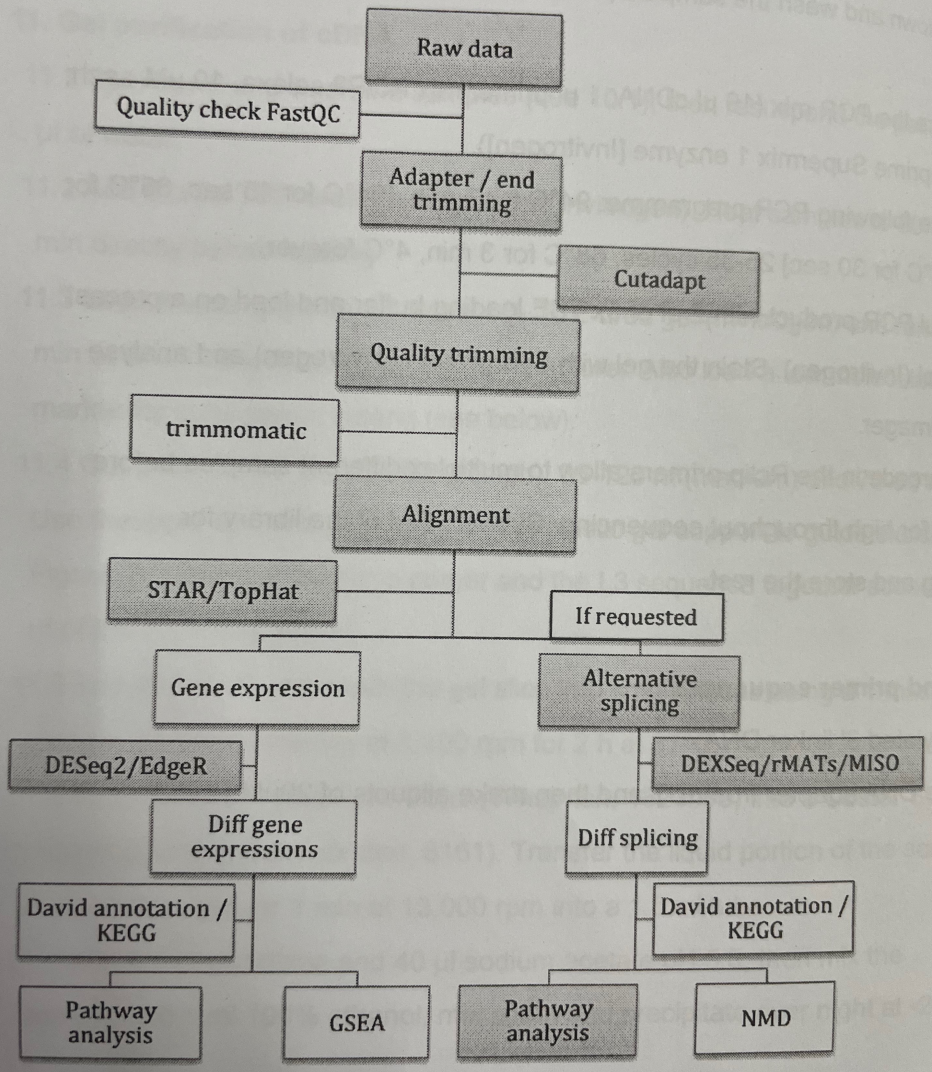
Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

2. Work flow


Raw data
Quality check FastQC
Adapter / end trimming
Cutadapt
Quality trimming
trimmomatic
Alignment
STAR/TopHat
If requested
Gene expression
Alternative splicing
DESeq2/EdgeR
DEXSeq/rMATs/MISO
Diff gene expressions
Diff splicing
David annotation / KEGG
David annotation / KEGG
Pathway analysis
GSEA
Pathway analysis
NMD

FastQC

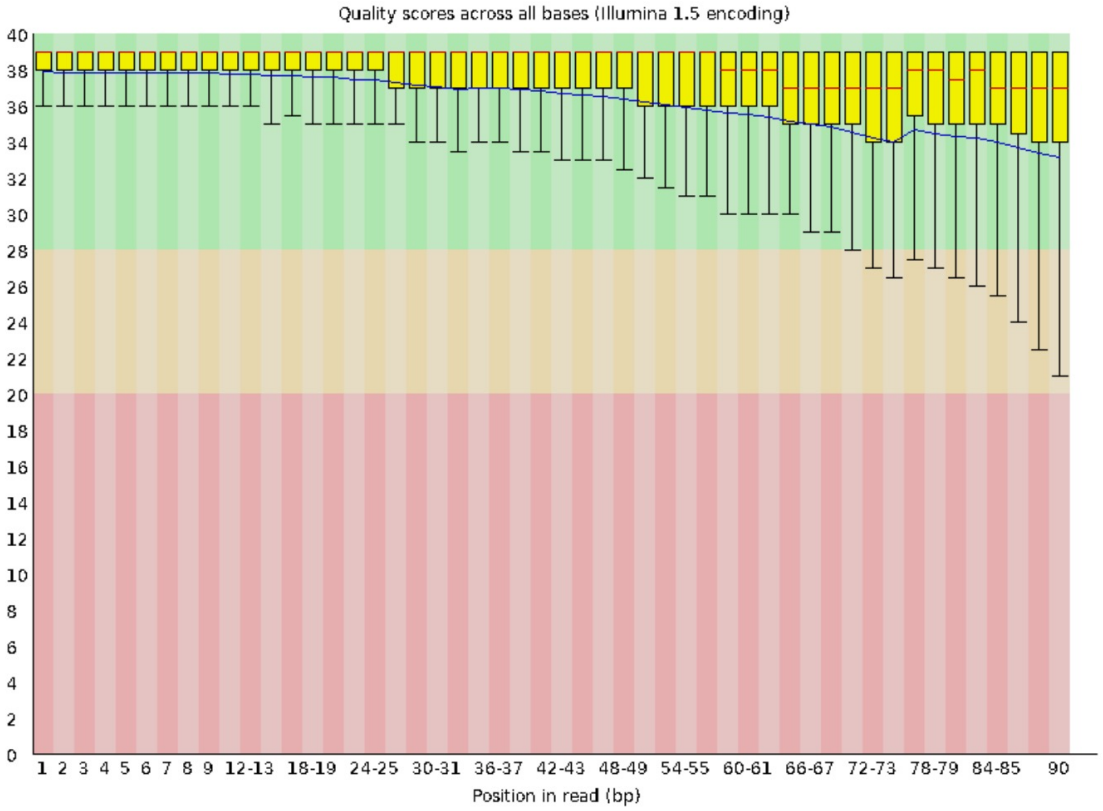
cd ~/download/wdc/

ls | while read id; do fastqc -t 8 -o /home/zhushu/zhouty19990625/download/wdc_data/fastqc1/ ${id}; done

实例：https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

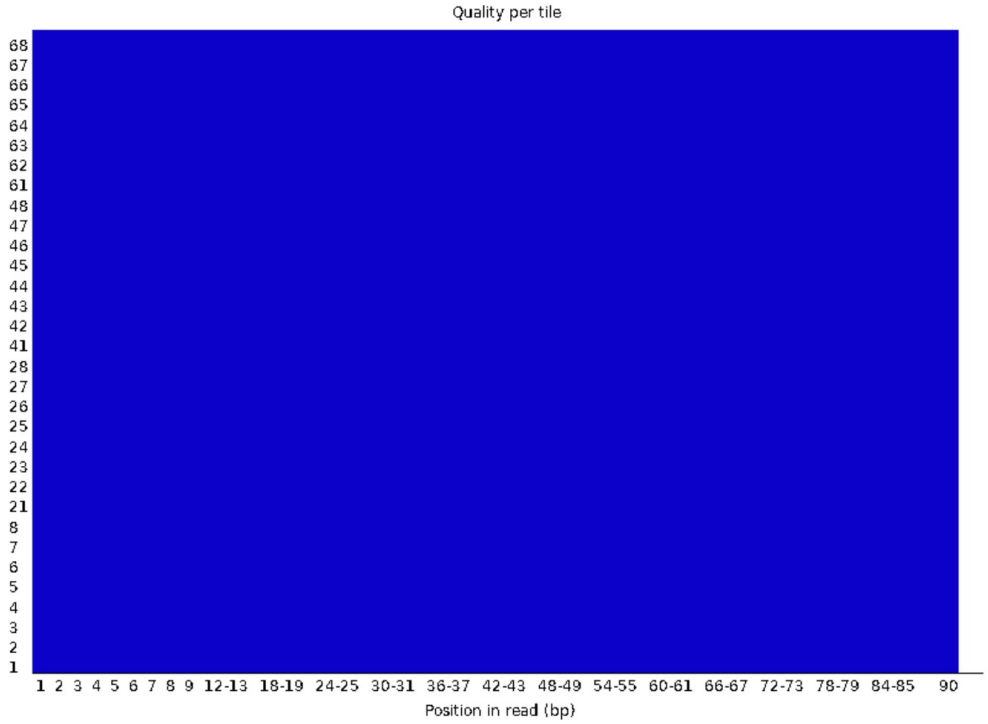
教学：https://zhuanlan.zhihu.com/p/88655260

FastQC



- **Per base sequence quality**
- 这是 read length = 90 的scRNAseq数据，横轴为read位置，纵轴是quality。
- quality = -10*log10(p)，p为测错的概率。
- 根据quality给出质量结果：正常区间（28 - 40），警告区间（20-28），错误区间（0-20）。
- 比如，当read的某一位置的p=0.01,quality=20，那么它就处于错误区间。

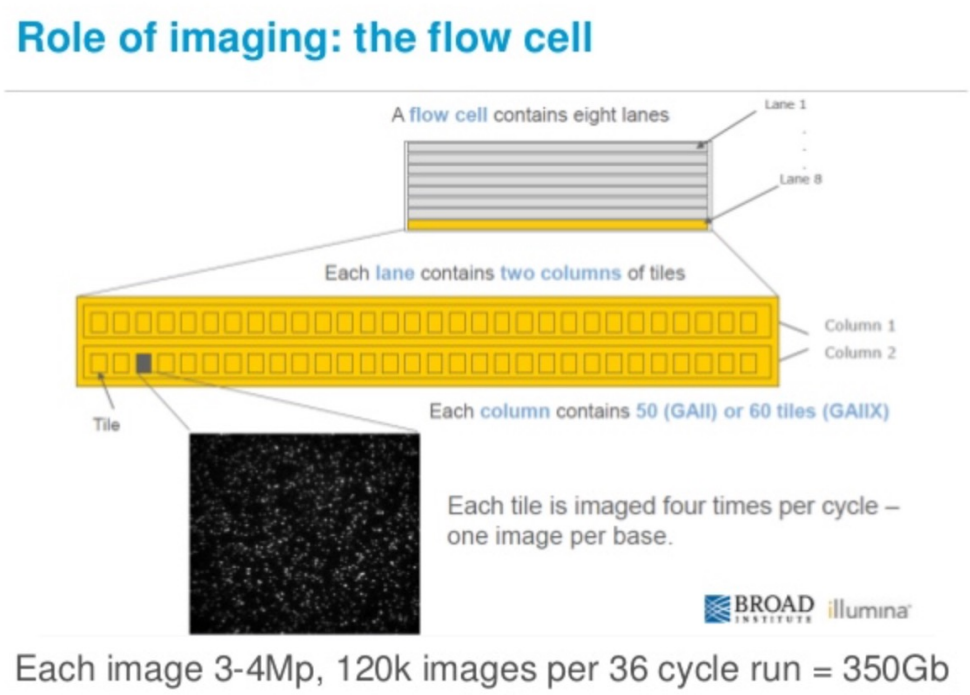
FastQC



- **Per tile sequence quality**

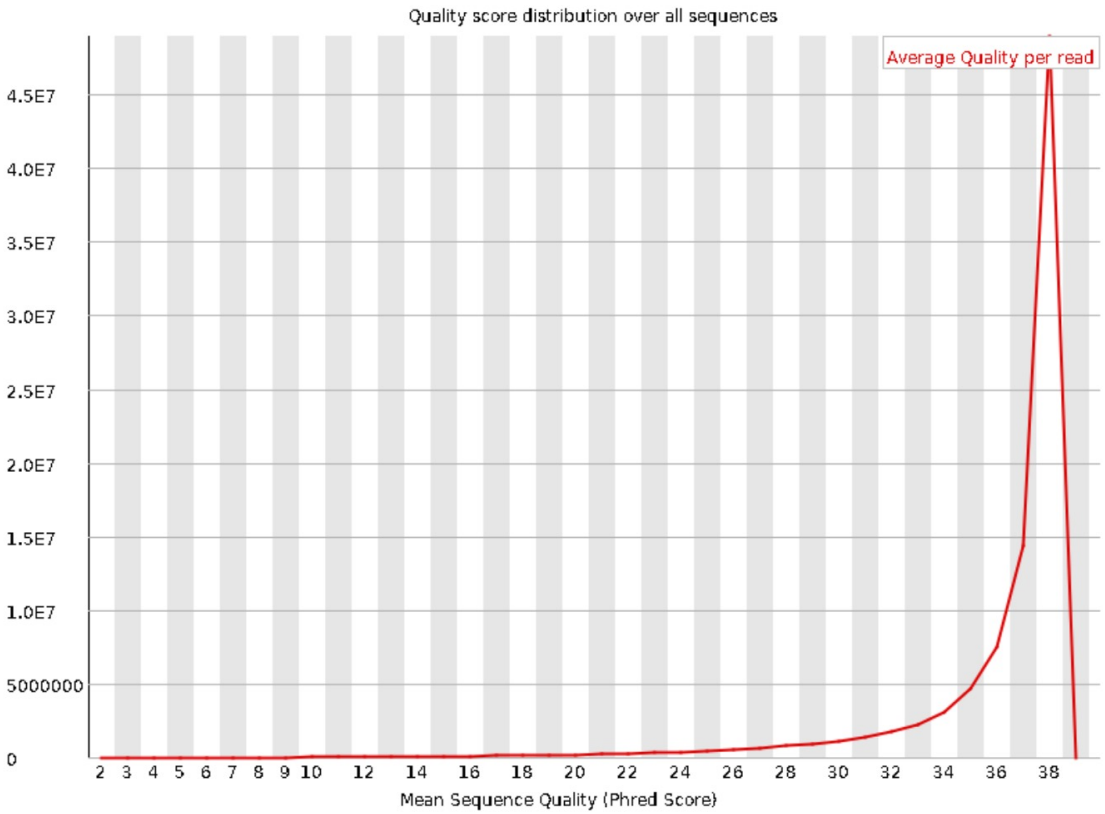
当Per tile sequence quality显示fail或者warning，表明测序的lane或某个run中出现出现了部分故障，从而影响一些特定的区域和循环，进而使测序数据的质量下降。另外，如果read的3'端的质量是好的，就意味着存在瞬时质量损失(Transient quality loss)的区域难以被剪切处理。

FastQC



- 在illumina 的测序设备中，根据flow cell的表面，人为的将其切分为swaths，这些swaths再进一步被切分为tiles。 通过查看per tile，识别因flow cell 或 run 的故障造成的测序的错误

FastQC

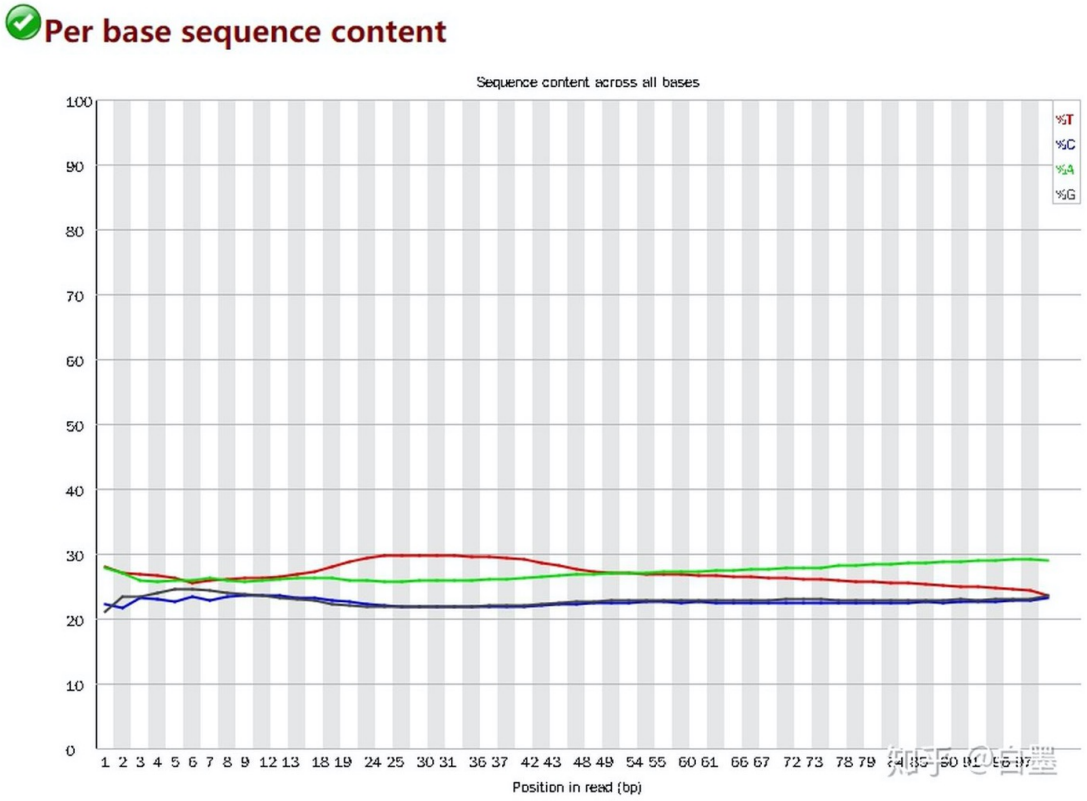


- **Per sequence quality scores**

这个图可以看出各个序列质量的分布情况，上图可以看出绝大部分序列质量都在30以上，质量可以说是很好了

- 横轴为quality，纵轴为reads计数。
- 当峰值处于quality为0-20时，报错。

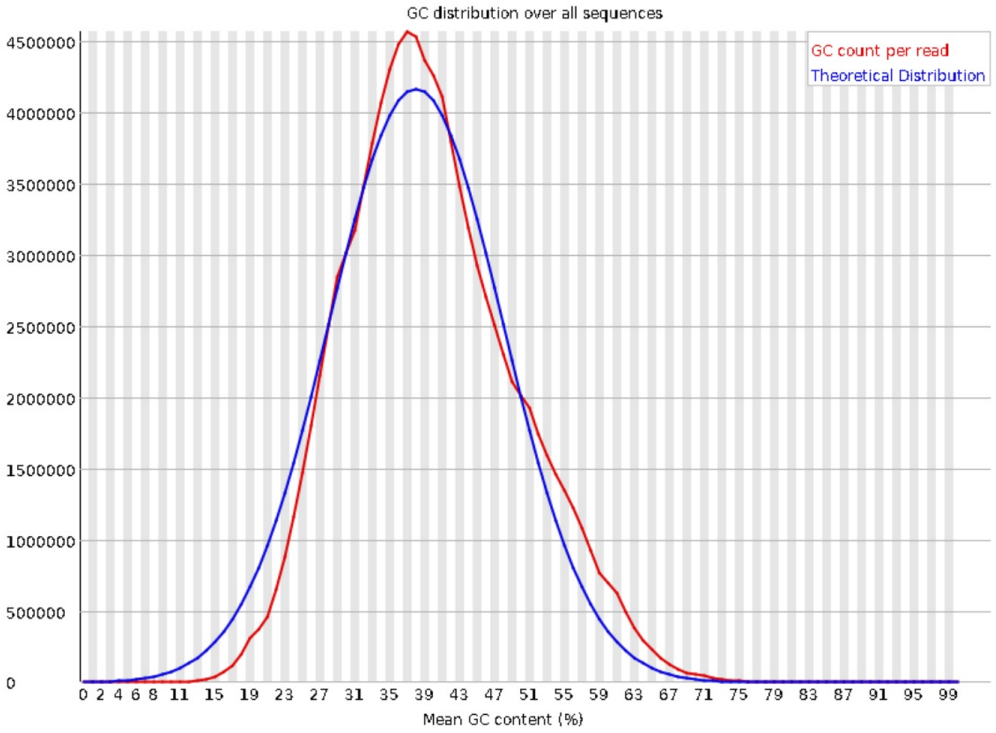
FastQC



- 横轴为位置，纵轴为百分比
- 正常测序数据为频率相近的四种碱基，无位置差异。表现在图上的话，四条线应该是平行且接近。
- 当任意位置A/T与G/C相差大于10%报警告，大于20%报错

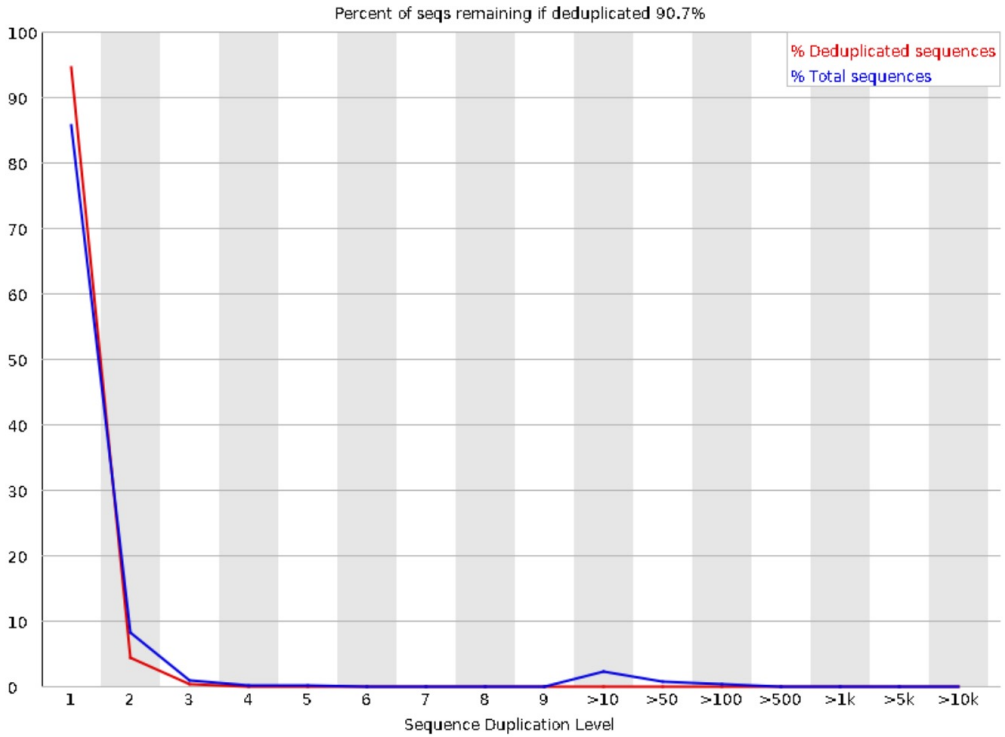
FastQC

7、Per sequence GC content



- 横轴为GC含量，纵轴为read计数。红色为实际测得，蓝色为理论分布。
- 如果曲线形状不符，代表文库污染
- 偏离大于15%，报警告；大于30%，报错

FastQC



- 理解：蓝线：在所有reads中，重复一次的read，占所有reads的比例。
- sequences duplication是指在测序前建库PCR过程中导致的一些序列扩增次数过多导致的。若重复较高则需要进行处理这些dup

FastQC

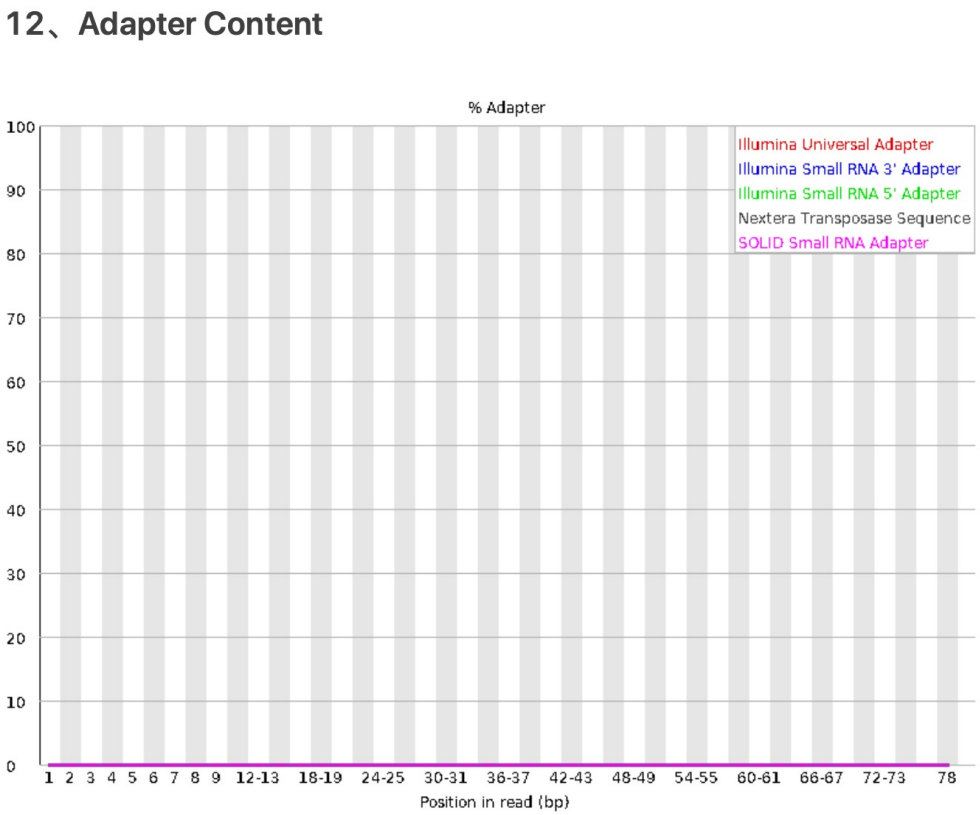
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit
CAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTT	1684	0.42601849790532476	No Hit
TGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACT	1668	0.4219708162150128	No Hit
CAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTA	1668	0.4219708162150128	No Hit
TATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAA	1630	0.4123575722005221	No Hit
GTCATGGAAGCGATAAAACTCTGCAGGTTGGATACGCCAA	1620	0.40982777114407726	No Hit

- 如果有某个序列大量出现，就叫做over-represented。fastqc的标准是占全部reads的0.1%以上

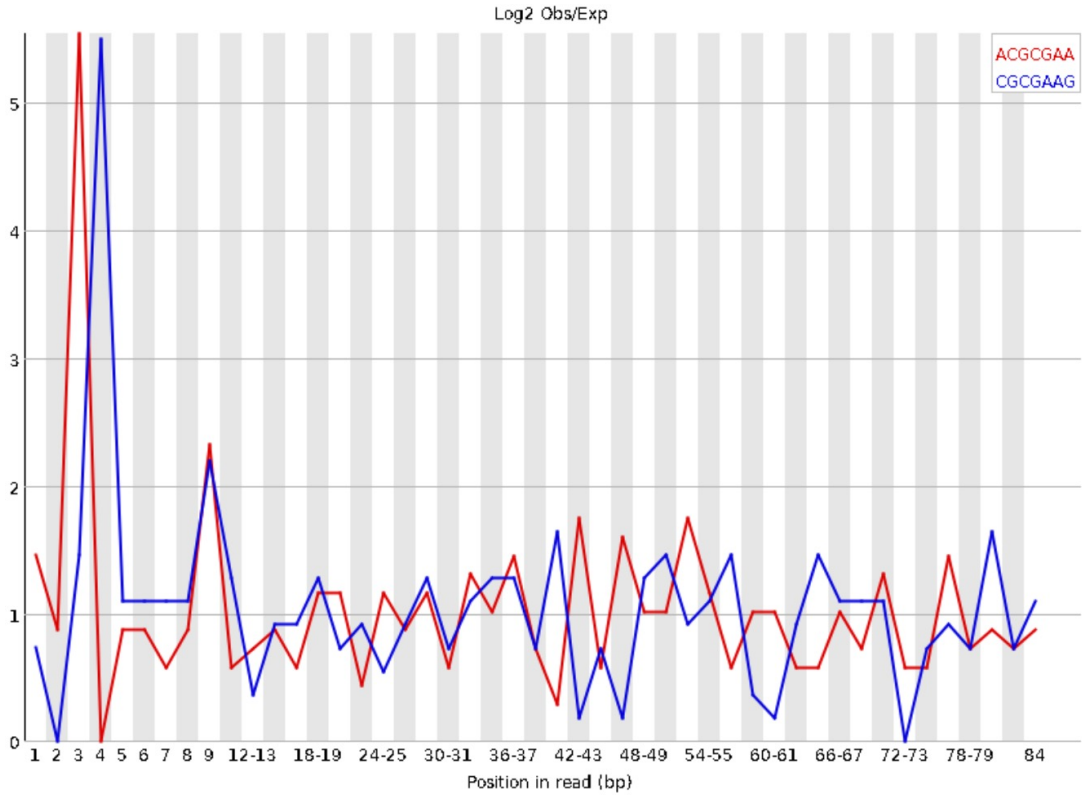
- 当发现超过总reads数0.1%的reads时报"黄色!"，当发现超过总reads数1%的reads时报"红色×"。

FastQC



- Adapter Content:
- 横坐标：read长度
- 纵坐标：有接头序列的比例
- Trimmomatic(java)

FastQC



- K-mer content
- 如果某k个bp的短序列在reads中大量出现，其频率高于统计期望的话，fastqc将其记为over-represented k-mer。默认的k = 5，可以用-k --kmers选项来调节，范围是2-10。出现频率总体上3倍于期望或是在某位置上5倍于期望的k-mer被认为是over-represented。fastqc除了列出所有over-represented k-mers，还会把前6个的per base distribution画出来。
- 当有出现频率总体上3倍于期望或是在某位置上5倍于期望的k-mer时，报"黄色!"；当有出现频率在某位置上10倍于期望的k-mer时报"红色×"。本图所显示的结果来自于表格中前六个序列。

FastQC

cd ~/download/wdc/

ls | while read id; do fastqc -t 8 -o /home/zhushu/zhouty19990625/download/wdc_data/fastqc1/ ${id}; done

实例：https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

教学：https://zhuanlan.zhihu.com/p/88655260

MultiQC

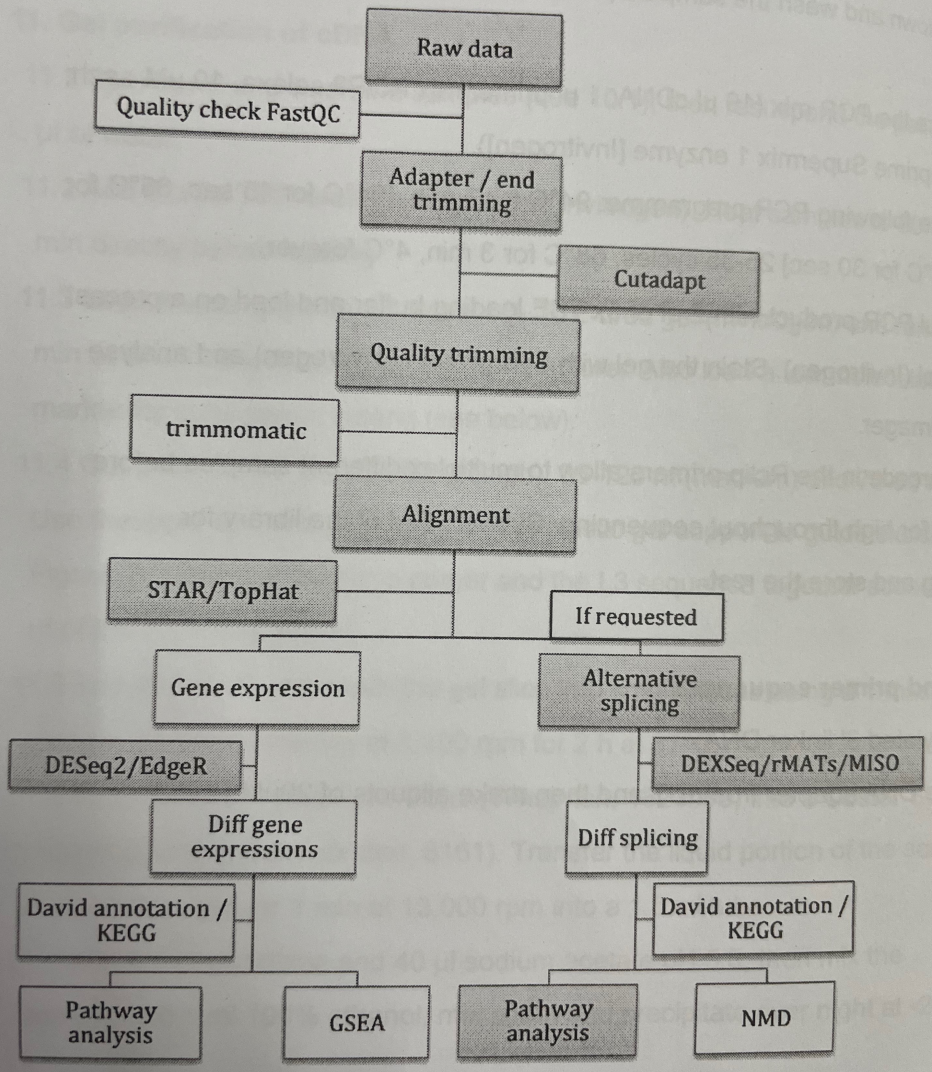
cd /home/zhushu/zhouty19990625/download/wdc_data/fastqc1/

multiqc *fastqc.zip --pdf -o /home/zhushu/zhouty19990625/download/wdc_data/output

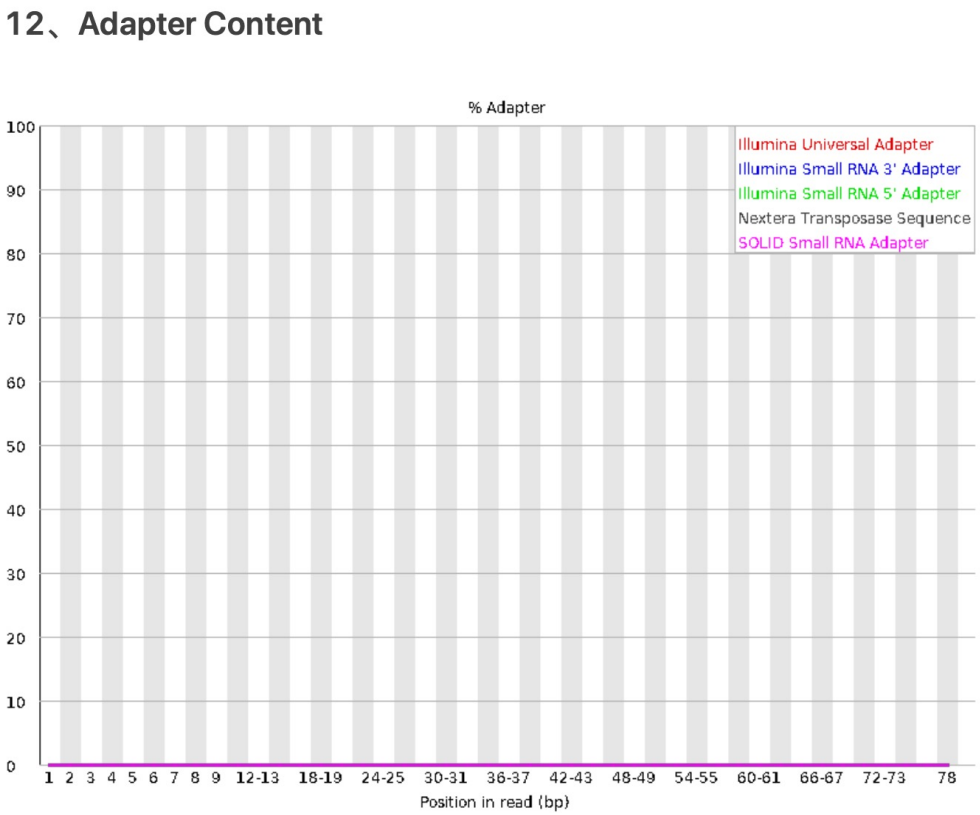
网站： http://multiqc.info

教学：https://www.jianshu.com/p/d06b0e3d6a78?from=singlemessage

2. Work flow


Raw data
Quality check FastQC
Adapter / end trimming
Cutadapt
Quality trimming
trimmomatic
Alignment
STAR/TopHat
If requested
Gene expression
Alternative splicing
DESeq2/EdgeR
DEXSeq/rMATs/MISO
Diff gene expressions
Diff splicing
David annotation / KEGG
David annotation / KEGG
Pathway analysis
GSEA
Pathway analysis
NMD

FastQC



- Adapter Content:
- 横坐标：read长度
- 纵坐标：有接头序列的比例
- Trimmomatics(java)

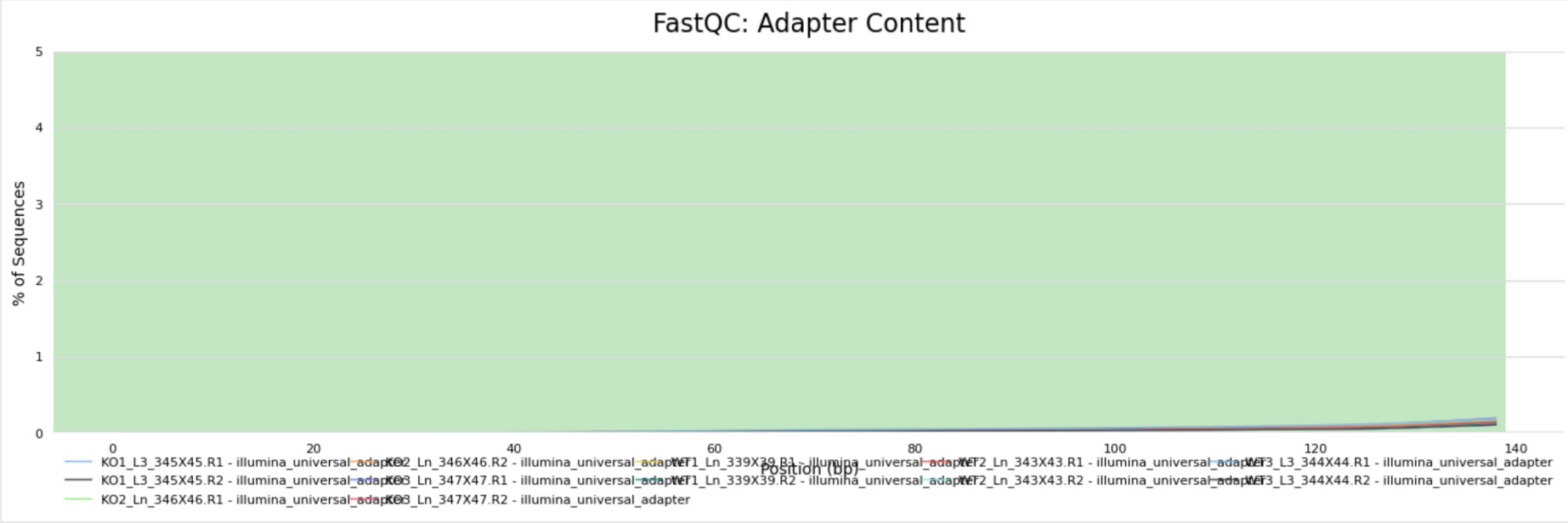
FastQC

Adapter Content

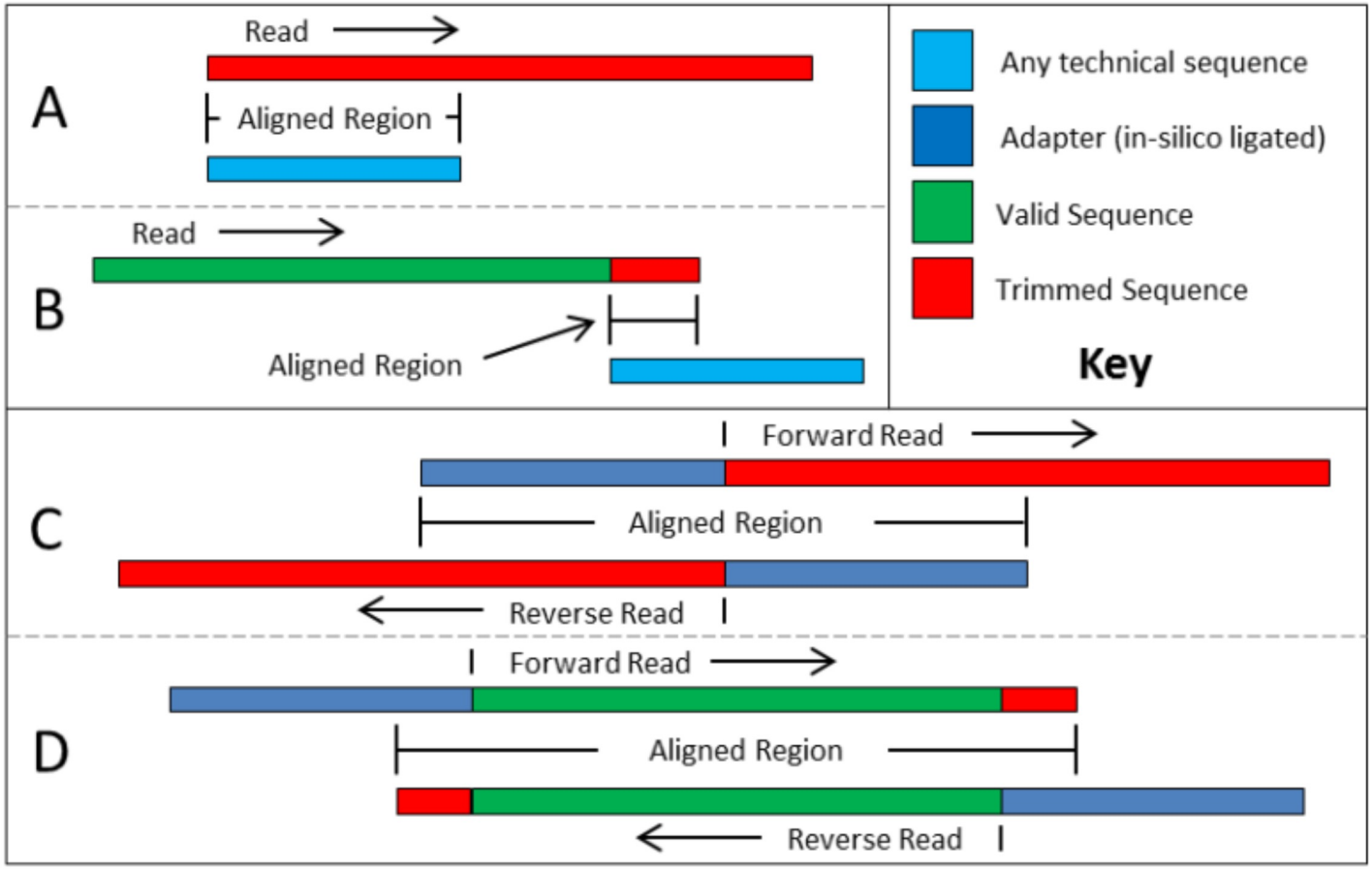
☐ Help

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

☐ Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs).



Why remove adapter?（实际情况要更复杂）



https://www.bilibili.com/video/BV1Cx411p7dm?from=search&seid=15560186713602196341

Trimmomatic

- cat ~/test_wdc.txt | while read id; do java -jar /home/zhushu/zhouty19990625/bioinfo_tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE "$id".R1.fastq "$id".R2.fastq -baseout /home/zhushu/zhouty19990625/download/wdc_data/wdc_after_trimming_fastq/"$id".fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:true SLIDINGWINDOW:4:15 MINLEN:50; done

- 网站： http://www.usadellab.org/cms/index.php?page=trimmomatic
- 教学： https://www.jianshu.com/p/a8935adebaae

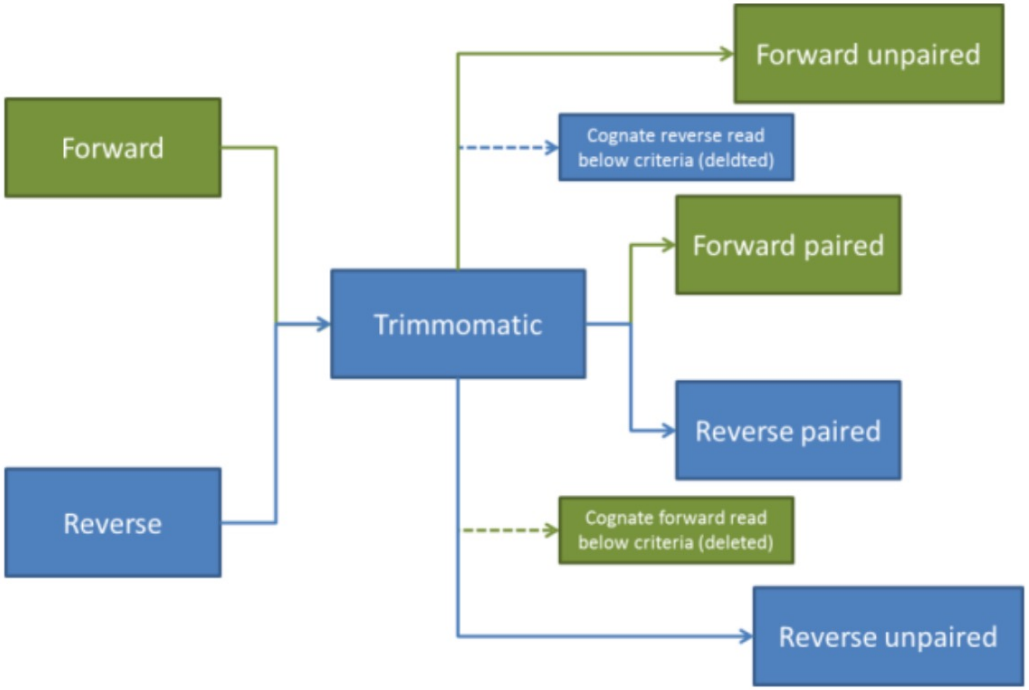
Trimmomatic

- cat ~/test_wdc.txt | while read id; do java -jar /home/zhushu/zhouty19990625/bioinfo_tools/Trimmomatic-0.39/trimmomatic-0.39.jar **PE "$id".R1.fastq "$id".R2.fastq -baseout /home/zhushu/zhouty19990625/download/wdc_data/wdc_after_trimming_fastq/"$id".fastq** ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:true SLIDINGWINDOW:4:15 MINLEN:50; done

- java -jar <path to trimmomatic.jar> PE [-threads <threads] [-phred33 | -phred64] [-trimlog <logFile>] >] [-basein <inputBase> | <input 1> <input 2>] [-baseout <outputBase> | <paired output 1> <unpaired output 1> <paired output 2> <unpaired output 2> <step 1> <step 2> ...

Trimmomatic

- cat ~/test_wdc.txt | while read id; do java -jar /home/zhushu/zhouty19990625/bioinfo_tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE "$id".R1.fastq "$id".R2.fastq -baseout /home/zhushu/zhouty19990625/download/wdc_data/wdc_after_trimming_fastq/**"$id".**fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:true SLIDINGWINDOW:4:15 MINLEN:50; done



- mySampleFiltered_1P.fq.gz - for paired forward reads
- mySampleFiltered_1U.fq.gz - for unpaired forward reads
- mySampleFiltered_2P.fq.gz - for paired reverse reads
- mySampleFiltered_2U.fq.gz - for unpaired reverse reads

Trimmomatic

- cat ~/test_wdc.txt | while read id; do java -jar /home/zhushu/zhouty19990625/bioinfo_tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE "$id".R1.fastq "$id".R2.fastq -baseout /home/zhushu/zhouty19990625/download/wdc_data/wdc_after_trimming_fastq/"$id".fastq **ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:true** SLIDINGWINDOW:4:15 MINLEN:50; done
- ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>:<minAdapterLength>:<keepBothReads>

Trimmomatic

- ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>:<minAdapterLength>:<keepBothReads>

- <fastaWithAdaptersEtc>:指定包含接头和引物序列（所有被视为污染的序列）的 fasta 文件路径，Trimmomatic 自带了一个包含 Illumina 平台接头和引物序列的 fasta 文件，可以直接用这个。

- <seed mismatches>:指定第一步 seed 搜索时允许的错配碱基个数，例如 2。

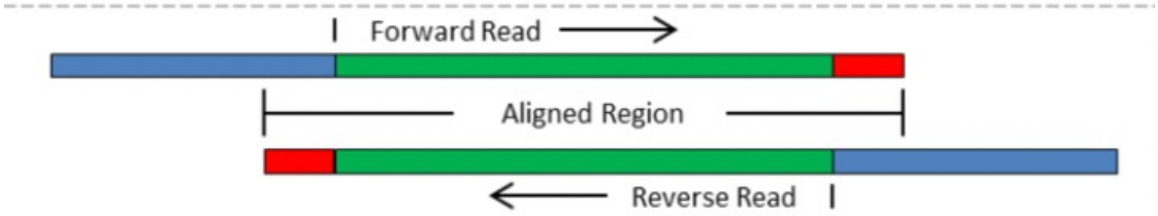
Trimmomatic

- ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>:<minAdapterLength>:<keepBothReads>

- <palindrome clip threshold>:指定针对 PE 的 palindrome clip 模式下，需要 R1 和 R2 之间至少多少比对分值（上图中 D 模式），才会进行接头切除，例如 30。
- <simple clip threshold>:指定切除接头序列的最低比对分值（上图 A/B 模式），通常 7-15 之间。

Trimmomatic

- ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>:<minAdapterLength>:<keepBothReads>
- <minAdapterLength>:只对 PE 测序的 palindrome clip 模式有效，指定 palindrome 模式下可以切除的接头序列最短长度，由于历史的原因，默认值是 8，但实际上 palindrome 模式可以切除短至 1bp 的接头污染，所以可以设置为 1 。
- <keepBothReads>:这种情况下要不要保留正反两条reads，默认不保留



Trimmomatic

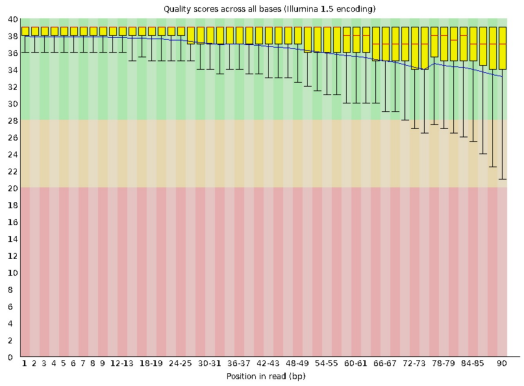
- cat ~/test_wdc.txt | while read id; do java -jar /home/zhushu/zhouty19990625/bioinfo_tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE "$id".R1.fastq "$id".R2.fastq -baseout /home/zhushu/zhouty19990625/download/wdc_data/wdc_after_trimming_fastq/"$id".fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:true **SLIDINGWINDOW:4:15** MINLEN:50; done
- SLIDINGWINDOW:<windowSize>:<requiredQuality>
- 滑窗剪切，统计滑窗口中所有碱基的平均质量值，如果低于设定的阈值，则切掉窗口。

Trimmomatic

- cat ~/test_wdc.txt | while read id; do java -jar /home/zhushu/zhouty19990625/bioinfo_tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE "$id".R1.fastq "$id".R2.fastq -baseout /home/zhushu/zhouty19990625/download/wdc_data/wdc_after_trimming_fastq/"$id".fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:8:true SLIDINGWINDOW:4:15 **MINLEN:50**; done
- 设定一个最短 read 长度，当 reads 经过前面的过滤之后，如果保留下来的长度低于这个阈值，整条 reads 被丢弃。被丢弃的 reads 数会被统计在 Trimmomatic 日志的 dropped reads 中。

Trimmomatic

- MAXINFO:<targetLength>:<strictness>
- <targetLength>:使得 reads 可以 map 到参考序列上唯一位置的最短长度
- <strictness>:一个介于 0 - 1 之间的小数，决定如何平衡 最大化 reads 长度 或者 最小化 reads 出现错误的概率，当参数设置小于 0.2 时倾向于最大化 reads 长度，当参数设置大于 0.8 时倾向于最小化 reads 中出现测序错误的概率。
- LEADING:<quality>
- TAILING:<quality>
- CROP:<length>
- HEADCROP:<length>



FastQC again

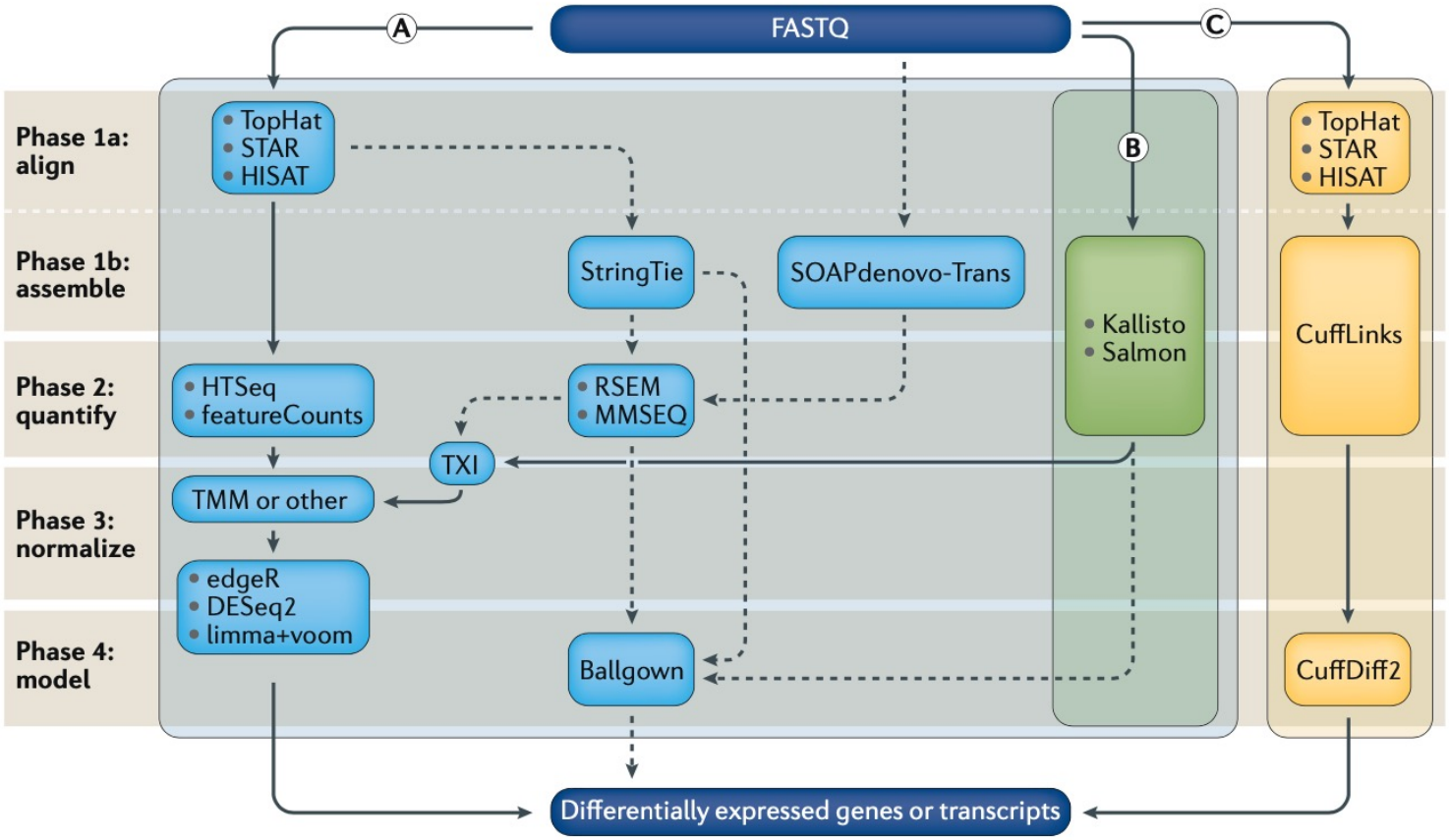
cd /home/zhushu/zhouty19990625/download/wdc_data/wdc_after_trimming_fastq/

ls | while read id; do fastqc -t 8 -o /home/zhushu/zhouty19990625/download/wdc_data/fastqc2/ $id; done

cd /home/zhushu/zhouty19990625/download/wdc_data/fastqc2/

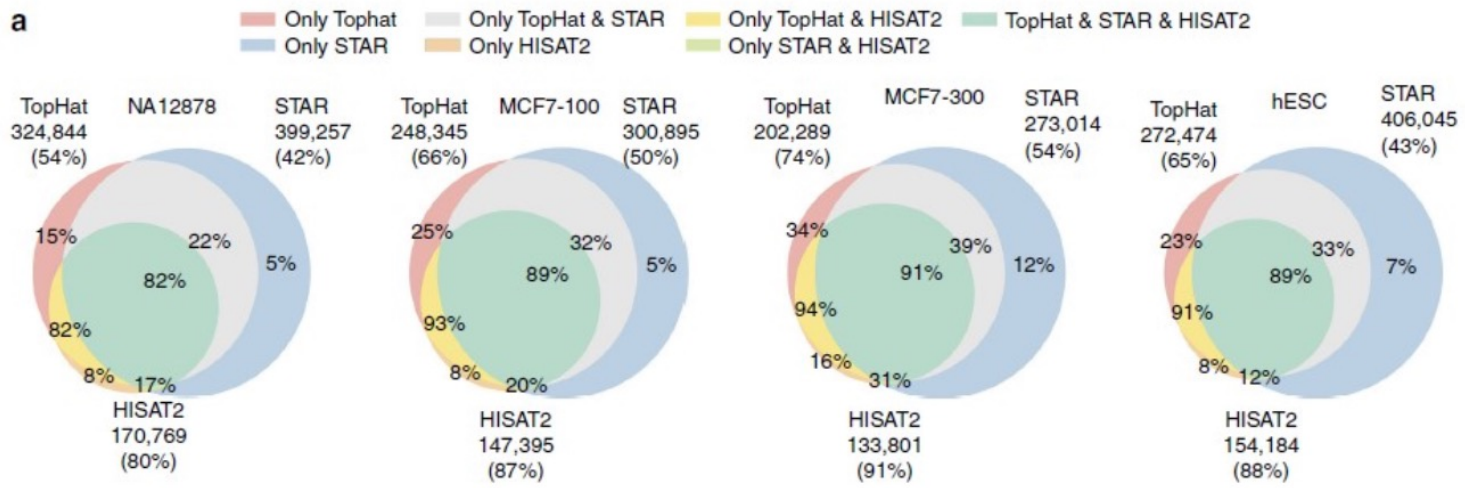
multiqc *fastqc.zip --pdf -o /home/zhushu/zhouty19990625/download/wdc_data/output2

RNA-seq正式步骤



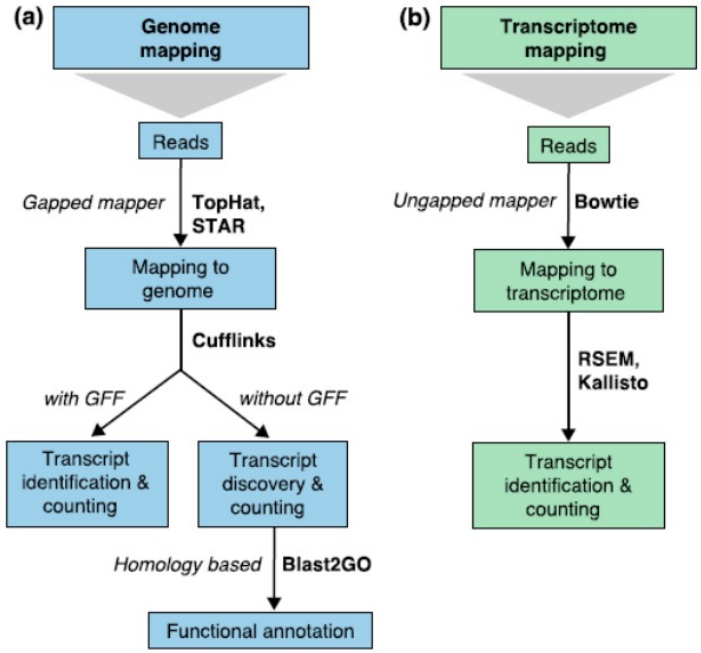
Gene level Transcript level

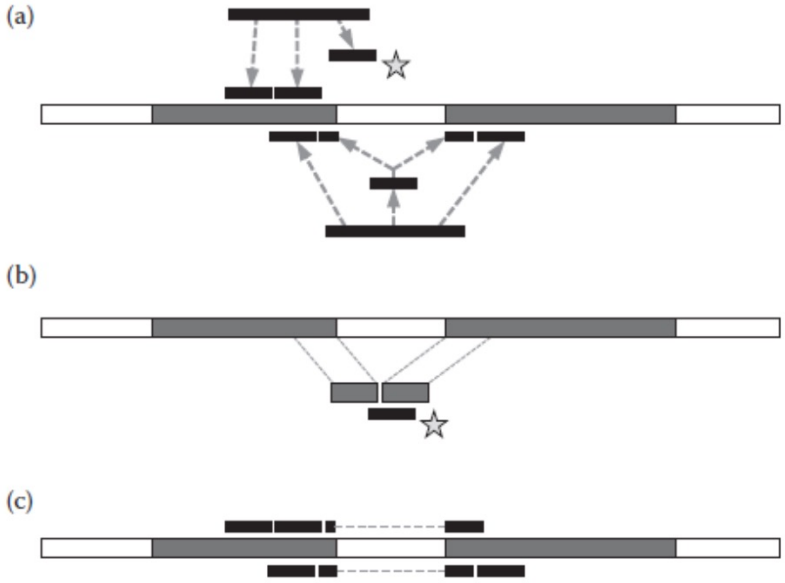
Align: STAR



Sahraeian SME, Mohiyuddin M, Sebra R, et al. ***Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis.*** *Nat Commun*. 2017;8(1):59. Published 2017 Jul 5. doi:10.1038/s41467-017-00050-4

Align: STAR&Hisat2 vs. Bowtie





Align: STAR

- 教程：https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf

- 网站/下载：https://github.com/alexdobin/STAR

STAR:Generating genome indexes

- STAR --runMode genomeGenerate --genomeDir ~/genome/mouse/index/ --runThreadN 8 --genomeFastaFiles ~/genome/mouse/dna/Mus_musculus.GRCm38.dna_rm.primary_assembly.fa --sjdbGTFfile ~/genome/mouse/ann/Mus_musculus.GRCm38.102.gtf --sjdbOverhang 148

WHY NEED index?
为了控制比对的时间在一个合理的范围之内，reads不会直接和基因组fasta文件直接进行比对。而是基于BWT？算法生成index，再将reads和index进行比对，当然也可以去hisat2官网下载
这里我更偏向于自己构建index
you can rename the chromosome names in the **chrName.txt** keeping the order of the chromosomes in the file: the names from this file will be used in all output files (e.g. SAM/BAM).
以避免后续流程出现的 ‘chr’ 的问题

STAR:Generating genome indexes

- STAR --runMode genomeGenerate --genomeDir ~/genome/mouse/index/ --runThreadN 8 --genomeFastaFiles ~/genome/mouse/dna/Mus_musculus.GRCm38.dna_rm.primary_assembly.fa --sjdbGTFfile ~/genome/mouse/ann/Mus_musculus.GRCm38.102.gtf --sjdbOverhang 148

< sjdbOverhang **>:** specifies the length of the genomic sequence around the annotated junction to be used in constructing the splice junctions database. Ideally, this length should be equal to the ReadLength-1, where ReadLength is the length of the reads. For instance, for Illumina 2x100b paired-end reads, the ideal value is 100-1=99. In case of reads of varying length, the ideal value is max(ReadLength)-1. **In most cases, the default value of 100 will work as well as the ideal value.**

STAR: Mapping

- cat ~/test_wdc.txt |while read id; do
- cd ~/download/wdc/;
- STAR --runMode alignReads --runThreadN 8 \
- --readFilesIn ${id}.R1.fastq ${id}.R2.fastq \
- --genomeDir /home/zhushu/zhouty19990625/genome/mouse/index \
- --outFileNamePrefix ~/download/wdc_data/${id}/${id}.\
- --outBAMsortingThreadN 8 \

--readFilesCommand zcat

- ${star_p};
- done

STAR: Mapping

- max_intron_size=1000000 #哺乳动物一般设这个值就可以，植物或者其他需要改动
- star_p=" --outFilterType BySJout --outSAMattributes NH HI AS NM MD \
- --outFilterMultimapNmax 20 --alignSJoverhangMin 8
- --alignSJDBoverhangMin 1 \
- --alignIntronMin 20 --alignIntronMax ${max_intron_size} \
- --alignMatesGapMax ${max_intron_size} \
- --outFilterMatchNminOverLread 0.66 --outFilterScoreMinOverLread 0.66 \
- --winAnchorMultimapNmax 70 --seedSearchStartLmax 45 \
- --outSAMattrIHstart 0 --outSAMstrandField intronMotif \
- --genomeLoad NoSharedMemory --outReadsUnmapped Fastx \
- --outSAMtype BAM SortedByCoordinate --quantMode TranscriptomeSAM GeneCounts"
- **ENCODE standard options**

STAR: Mapping

- max_intron_size=1000000 #哺乳动物一般设这个值就可以，植物或者其他需要改动
- star_p=" --outFilterType BySJout --outSAMattributes NH HI AS NM MD \
- --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 \
- --alignIntronMin 20 --alignIntronMax ${max_intron_size} \
- --alignMatesGapMax ${max_intron_size} \
- --outFilterMatchNminOverLread 0.66 --outFilterScoreMinOverLread 0.66 \
- --winAnchorMultimapNmax 70 --seedSearchStartLmax 45 \
- --outSAMattrIHstart 0 --outSAMstrandField intronMotif \
- --genomeLoad NoSharedMemory --outReadsUnmapped Fastx \
- --**outSAMtype BAM SortedByCoordinate** --quantMode TranscriptomeSAM GeneCounts"

STAR: Results

```
#总用量 8.3G
#-rw-r--r-- 1 zhouty19990625 zhushu  2.3G 2月  10 10:46 S316WT_L2_338X38.Aligned.sortedByCoord.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  4.1G 2月  10 10:45 S316WT_L2_338X38.Aligned.toTranscriptome.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.final.out
#-rw-r--r-- 1 zhouty19990625 zhushu   17K 2月  10 10:46 S316WT_L2_338X38.Log.out
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.progress.out
#-rw-r--r-- 1 zhouty19990625 zhushu  1.4M 2月  10 10:45 S316WT_L2_338X38.ReadsPerGene.out.tab
#-rw-r--r-- 1 zhouty19990625 zhushu  6.3M 2月  10 10:45 S316WT_L2_338X38.SJ.out.tab
#drwx------ 3 zhouty19990625 zhushu  4.0K 2月  10 10:46 S316WT_L2_338X38._STARtmp
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate1
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate2
```

--outSAMtype BAM SortedByCoordinate
output sorted by coordinate Aligned.sortedByCoord.out.bam file, similar to samtools sort command.

SJ.out.tab contains high confidence collapsed splice junctions in tab-delimited format.

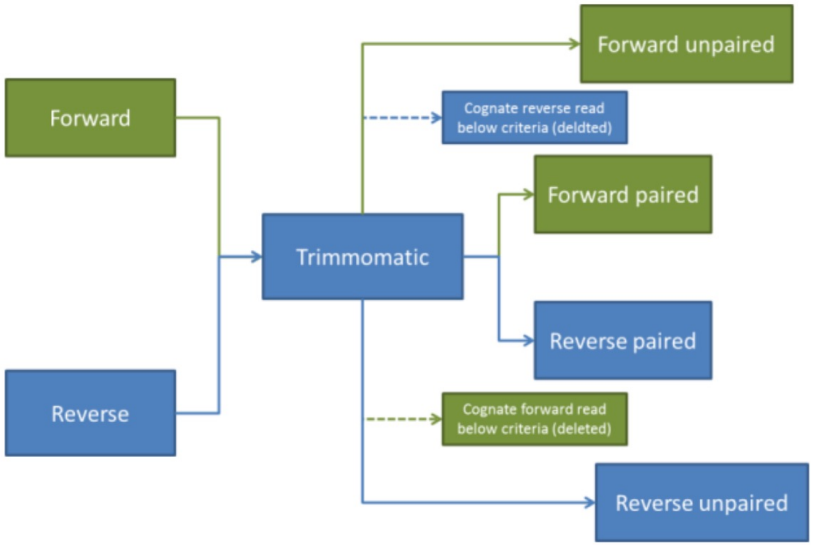
STAR: Results

```
#总用量 8.3G
#-rw-r--r-- 1 zhouty19990625 zhushu  2.3G 2月  10 10:46 S316WT_L2_338X38.Aligned.sortedByCoord.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  4.1G 2月  10 10:45 S316WT_L2_338X38.Aligned.toTranscriptome.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.final.out
#-rw-r--r-- 1 zhouty19990625 zhushu   17K 2月  10 10:46 S316WT_L2_338X38.Log.out
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.progress.out
#-rw-r--r-- 1 zhouty19990625 zhushu  1.4M 2月  10 10:45 S316WT_L2_338X38.ReadsPerGene.out.tab
#-rw-r--r-- 1 zhouty19990625 zhushu  6.3M 2月  10 10:45 S316WT_L2_338X38.SJ.out.tab
#drwx------ 3 zhouty19990625 zhushu  4.0K 2月  10 10:46 S316WT_L2_338X38._STARtmp
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate1
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate2
```

--quantMode TranscriptomeSAM GeneCounts
With --quantMode GeneCounts option STAR will count number reads per gene while mapping. A read is counted if it overlaps (1nt or more) one and only one gene. Both ends of the paired- end read are checked for overlaps. **The counts coincide with those produced by htseq-count with default parameters.**

结果存储于ReadsPerGene.out.tab

STAR: Questions:解决trimmomatic unpaired




```
Feb 05 20:40:07 ...... FATAL ERROR, exiting

EXITING because of FATAL ERROR in reads input: short read sequence line: 0
Read Name=@A00838:309:HKMKLDSXY:2:1355:7916:14450
Read Sequence====
DEF_readNameLengthMax=50000
DEF_readSeqLengthMax=650

Feb 05 20:51:33 ...... FATAL ERROR, exiting

EXITING because of FATAL ERROR in reads input: short read sequence line: 0
Read Name=@A00838:309:HKMKLDSXY:2:1350:7771:19366
Read Sequence====
DEF_readNameLengthMax=50000
DEF_readSeqLengthMax=650

Feb 05 21:04:35 ...... FATAL ERROR, exiting

EXITING because of FATAL ERROR in reads input: short read sequence line: 0
Read Name=@A00838:309:HKMKLDSXY:2:1347:31340:31657
Read Sequence====
DEF_readNameLengthMax=50000
DEF_readSeqLengthMax=650
```

- mySampleFiltered_1P.fq.gz - for paired forward reads
- mySampleFiltered_1U.fq.gz - for unpaired forward reads
- mySampleFiltered_2P.fq.gz - for paired reverse reads
- mySampleFiltered_2U.fq.gz - for unpaired reverse reads

SAMTOOLS

cat ~/test_wdc.txt|while read id; do cd /home/zhushu/zhouty19990625/download/wdc_data/${id}/ ;samtools index ${id}.Aligned.sortedByCoord.out.bam ;done

http://samtools.github.io/hts-specs/SAMv1.pdf

SAMTOOLS

cat ~/test_wdc.txt|while read id; do cd /home/zhushu/zhouty19990625/download/wdc_data/${id}/ ;samtools index ${id}.Aligned.sortedByCoord.out.bam ;done

Samtool三板斧

samtools view -S SRR35899${i}.sam -b > SRR35899${i}.bam
samtools sort SRR35899${i}.bam -o SRR35899${i}_sorted.bam
samtools index SRR35899${i}_sorted.bam

SAMTOOLS

cat ~/test_wdc.txt|while read id; do cd /home/zhushu/zhouty19990625/download/wdc_data/${id}/ ;samtools index ${id}.Aligned.sortedByCoord.out.bam ;done

.sam vs .bam

SAM(Sequence Alignment Map)格式是一种通用的比对格式，用来存储reads到参考序列的比对信息。
SAM是一种序列比对格式标准，由sanger制定，是以TAB为分割符的文本格式。主要应用于测序序列mapping到基因组上的结果表示，当然也可以表示任意的多重比对结果。
SAM分为两部分，注释信息（header section）和**比对结果部分（alignment section）**

Bam文件内容同sam,但是他是二进制的，占用空间小，使用samtools view可以查看。

SAMTOOLS

cat ~/test_wdc.txt|while read id; do cd /home/zhushu/zhouty19990625/download/wdc_data/${id}/ ;samtools index ${id}.Aligned.sortedByCoord.out.bam ;done

.sam vs .bam

行：除注释外，每一行是一个read

https://www.jianshu.com/p/9c99e09630da

SAMTOOLS

cat ~/test_wdc.txt|while read id; do cd /home/zhushu/zhouty19990625/download/wdc_data/${id}/ ;samtools index ${id}.Aligned.sortedByCoord.out.bam ;done

Samtool三板斧

samtools view -S SRR35899${i}.sam -b > SRR35899${i}.bam
samtools sort SRR35899${i}.bam -o SRR35899${i}_sorted.bam
samtools index SRR35899${i}_sorted.bam

SAMTOOLS

cat ~/test_wdc.txt|while read id; do cd /home/zhushu/zhouty19990625/download/wdc_data/${id}/ ;samtools index ${id}.Aligned.sortedByCoord.out.bam ;done

samtools sort SRR35899${i}.bam -o SRR35899${i}_sorted.bam

在测序的时候序列是随机打断的，所以reads也是随机测序记录的，进行比对的时候，产生的结果自然也是乱序的，**为了后续分析的便利，将bam文件进行排序（按坐标，按reads名称）**。事实上，后续很多分析都建立在已经排完序的前提下。

*注意自己排序的时候使用的方法，比如你用的sort by coordinate，后面计数（如htseq-counts）要选择对应的选项。 **（当然现在这个问题其实可以不用考虑了）**

STAR

```
#总用量 8.3G
#-rw-r--r-- 1 zhouty19990625 zhushu  2.3G 2月  10 10:46 S316WT_L2_338X38.Aligned.sortedByCoord.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  4.1G 2月  10 10:45 S316WT_L2_338X38.Aligned.toTranscriptome.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.final.out
#-rw-r--r-- 1 zhouty19990625 zhushu   17K 2月  10 10:46 S316WT_L2_338X38.Log.out
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.progress.out
#-rw-r--r-- 1 zhouty19990625 zhushu  1.4M 2月  10 10:45 S316WT_L2_338X38.ReadsPerGene.out.tab
#-rw-r--r-- 1 zhouty19990625 zhushu  6.3M 2月  10 10:45 S316WT_L2_338X38.SJ.out.tab
#drwx------ 3 zhouty19990625 zhushu  4.0K 2月  10 10:46 S316WT_L2_338X38._STARtmp
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate1
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate2
```

column 1: gene ID

column 2: counts for unstranded RNA-seq

column 3: counts for the 1st read strand aligned with RNA (htseq-count option -s yes)

column 4: counts for the 2nd read strand aligned with RNA (htseq-count option -s reverse)

```
cat ~/test_wdc.txt|while read id; do htseq-count -r pos -f bam -n 2
~/download/wdc_data/${id}/${id}.Aligned.sortedByCoord.out.bam
/home/zhushu/zhouty19990625/genome/mouse/ann/Mus_musculus.GRCm38.102.gtf >
~/download/wdc_data/output2/${id}.count;done
```

SAMTOOLS index

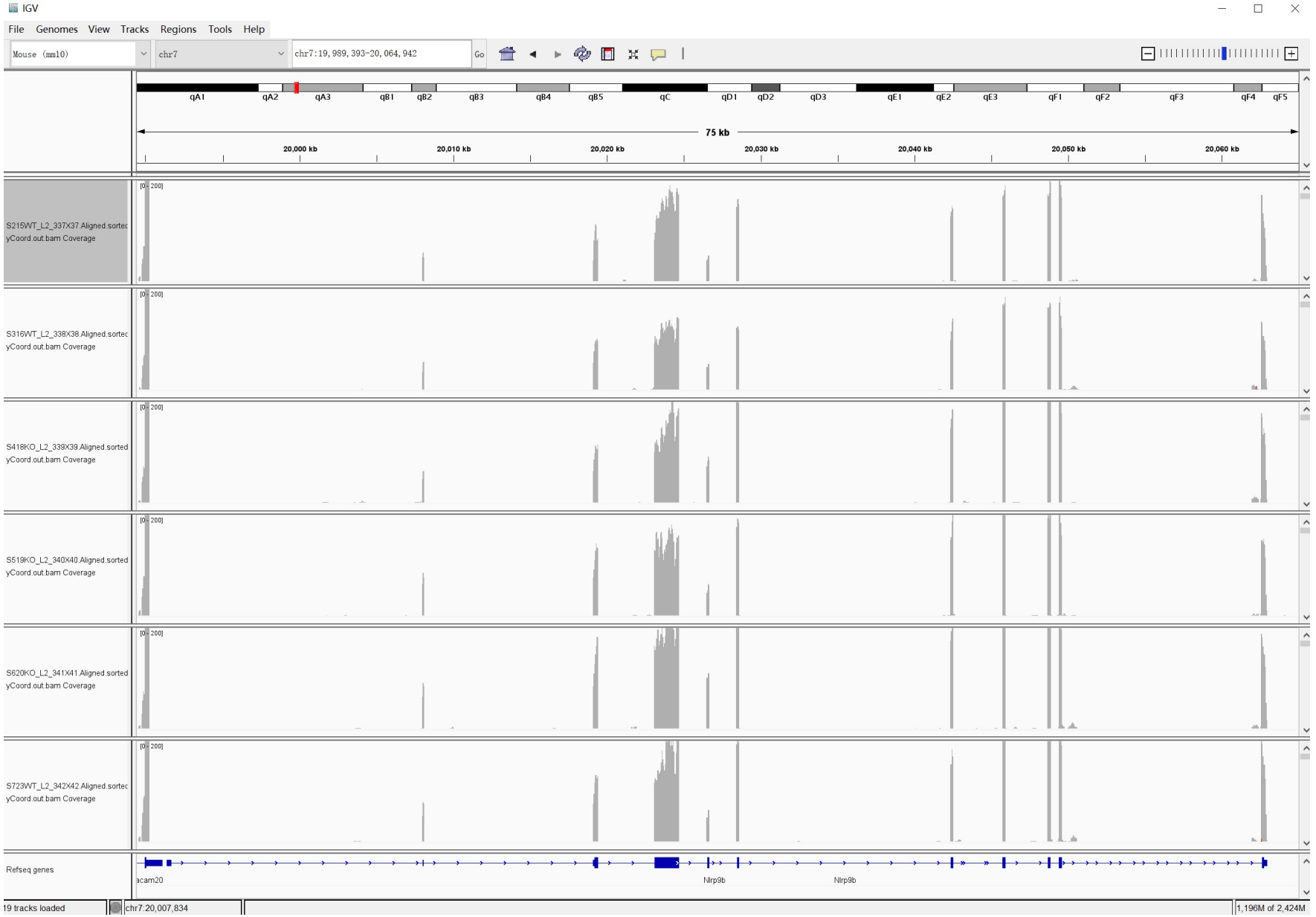
cat ~/test_wdc.txt|while read id; do cd /home/zhushu/zhouty19990625/download/wdc_data/${id}/ ;samtools **index** ${id}.Aligned.sortedByCoord.out.bam ;done

samtools index SRR35899${i}_sorted.bam

建立索引，减少扫描数量，是下一步使用htseq-count进行计数或者把bam文件导入igv进行查看的必要步骤

igv： **Integrative Genomics Viewer**

https://software.broadinstitute.org/software/igv/


IGV
File Genomes View Tracks Regions Tools Help
Mouse (mm10)
chr7
chr7:19,989,393-20,064,942
Go
qA1 qA2 qA3 qB1 qB2 qB3 qB4 qB5 qC qD1 qD2 qD3 qE1 qE2 qE3 qF1 qF2 qF3 qF4 qF5
75 kb
20,000 kb
20,010 kb
20,020 kb
20,030 kb
20,040 kb
20,050 kb
20,060 kb
S215WT_L2_337X37.Aligned.sortedyCoord.out.bam Coverage
S316WT_L2_338X38.Aligned.sortedyCoord.out.bam Coverage
S418KO_L2_339X39.Aligned.sortedyCoord.out.bam Coverage
S519KO_L2_340X40.Aligned.sortedyCoord.out.bam Coverage
S620KO_L2_341X41.Aligned.sortedyCoord.out.bam Coverage
S723WT_L2_342X42.Aligned.sortedyCoord.out.bam Coverage
[0-200]
Refseq genes
acam20
Nlrp9b
Nlrp9b
19 tracks loaded
chr7:20,007,834
1,196M of 2,424M

RSeQC

- 对比对之后的BAM进行质控
- 网站&教学：http://rseqc.sourceforge.net

这一步质控的意义：

比对的情况怎么样？多少比对上了？基因组是不是选对了？比对的质量高不高？

是不是存在PCR扩增导致的偏倚？

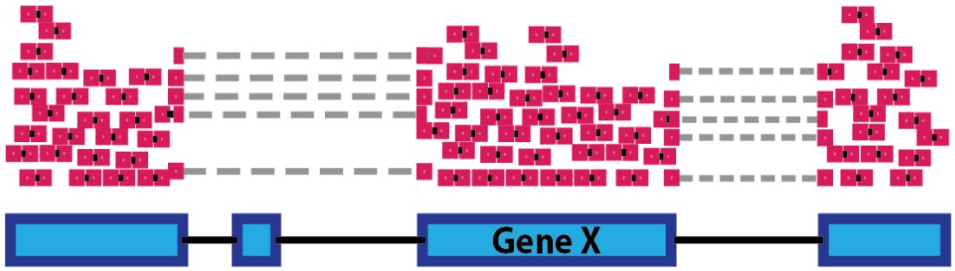
RSeQC

- geneBody_coverage.py -r /home/zhushu/zhouty19990625/genome/mouse/ann_bed/Mus_musculus.GRCm38.model.gtf.bed12 -i ~/download/wdc_data/bam_path.txt -o ~/download/wdc_data/output/

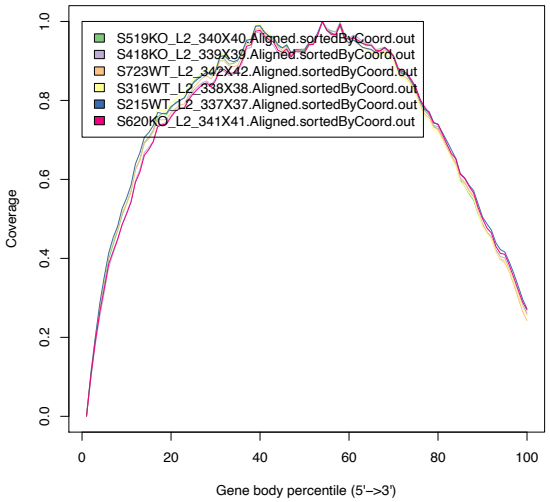
- cat ~/test_wdc.txt|while read id; do echo ${id}; cd ~/download/wdc_data/${id}/; bam_stat.py -i ${id}.Aligned.sortedByCoord.out.bam; read_distribution.py -i ${id}.Aligned.sortedByCoord.out.bam -r /home/zhushu/zhouty19990625/genome/mouse/ann_bed/Mus_musculus.GRCm38.102.gtf.bed12; RPKM_saturation.py -i ${id}.Aligned.sortedByCoord.out.bam -r /home/zhushu/zhouty19990625/genome/mouse/ann_bed/Mus_musculus.GRCm38.102.gtf.bed12 -s 10 -q 0 -o ~/download/wdc_data/output/${id}.RPKM_saturation;done

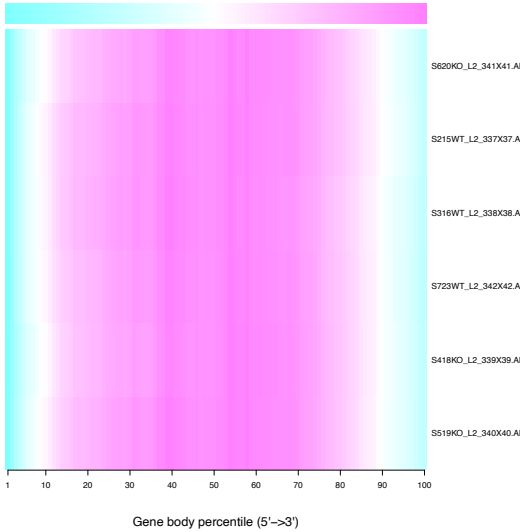
RSeQC

- geneBody_coverage.py -r /home/zhushu/zhouty19990625/genome/mouse/ann_bed/Mus_musculus.GRCm38.model.gtf.bed12 -i ~/download/wdc_data/bam_path.txt -o ~/download/wdc_data/output/

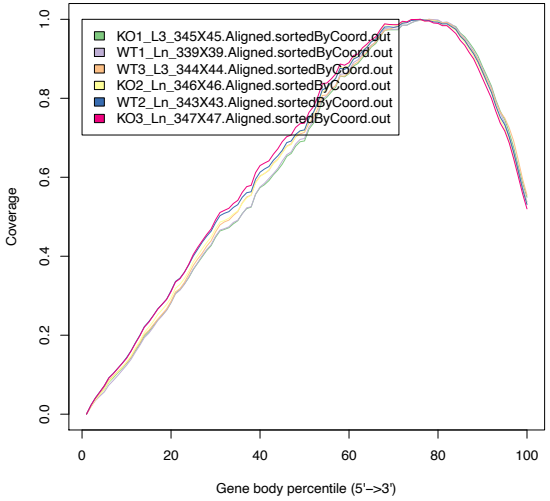


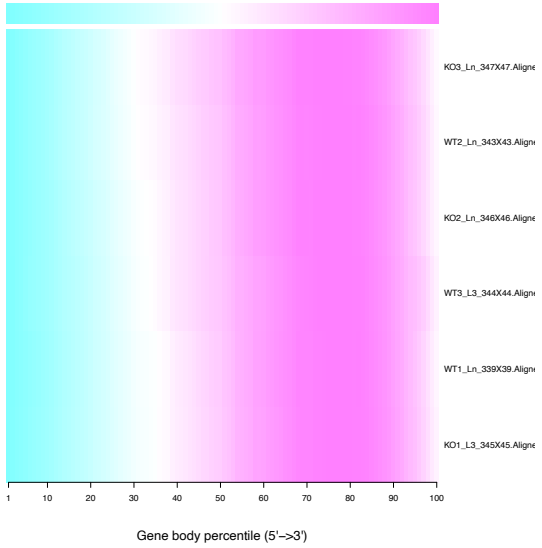
RXX : DHX30





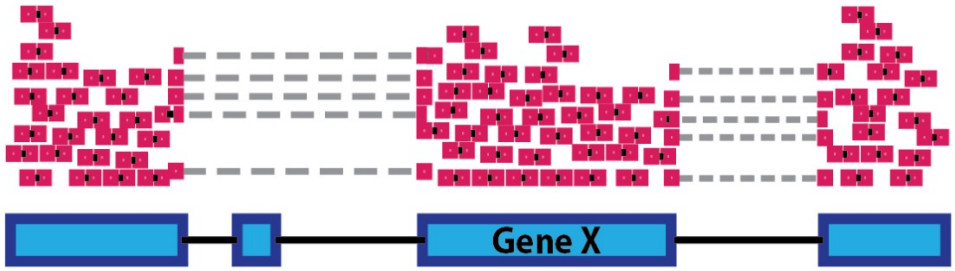
WDC : N9





RSeQC

- geneBody_coverage.py -r /home/zhushu/zhouty19990625/genome/mouse/ann_bed/Mus_musculus.GRCm38.model.gtf.**bed12** -i ~/download/wdc_data/bam_path.txt -o ~/download/wdc_data/output/



为什么PCR会导致3’ 偏倚？？

GTFtoBED

- gtfToGenePred -ignoreGroupsWithoutExons ~/genome/mouse/ann/Mus_musculus.GRCm38.102.gtf ~/genome/mouse/ann_bed/Mus_musculus.GRCm38.102.gtf.50505050.pred
- genePredToBed ~/genome/mouse/ann_bed/Mus_musculus.GRCm38.102.gtf.50505050.pred ~/genome/mouse/ann_bed/Mus_musculus.GRCm38.102.gtf.bed12
- awk ‘$3-$2>1000 && $3-$2<2000 ‘ Mus_musculus.GRCm38.102.gtf.bed12 >Mus_musculus.GRCm38.model.gtf.bed12

RSeQC

- cat ~/test_wdc.txt|while read id; do echo ${id}; cd ~/download/wdc_data/${id}/; **bam_stat.py** -i ${id}.Aligned.sortedByCoord.out.bam; **read_distribution.py** -i ${id}.Aligned.sortedByCoord.out.bam -r /home/zhushu/zhouty19990625/genome/mouse/ann_bed/Mus_musculus.GRCm38.102.gtf.bed12; **RPKM_saturation.py** -i ${id}.Aligned.sortedByCoord.out.bam -r /home/zhushu/zhouty19990625/genome/mouse/ann_bed/Mus_musculus.GRCm38.102.gtf.bed12 -s 10 -q 0 -o ~/download/wdc_data/output/${id}.RPKM_saturation;done

RSeQC

- **bam_stat.py**
- 'bam_stat.py' is used to check the mapping statistics of reads that areQC failed, unique mapped, splice mapped, mapped in proper pair, etc

```
bam_stat.py  -i Pairend_nonStrandSpecific_36mer_Human_hg19.bam

#Output (all numbers are read count)
#==================================================
Total records:                                 41465027
QC failed:                                     0
Optical/PCR duplicate:                         0
Non Primary Hits                               8720455
Unmapped reads:                                0

mapq < mapq_cut (non-unique):                  3127757
mapq >= mapq_cut (unique):                     29616815
Read-1:                                        14841738
Read-2:                                        14775077
Reads map to '+':                              14805391
Reads map to '-':                              14811424
Non-splice reads:                              25455360
Splice reads:                                  4161455
Reads mapped in proper pairs:                  21856264
Proper-paired reads map to different chrom:    7648
```

https://www.researchgate.net/publication/228087325_RSeQC_quality_control_of_RNA-seq_experiments

RSeQC

- **read_distribution.py**

- ‘read_distribution.py’ calculates the fraction of reads mapped tocoding exons, 5’-untranslated region (UTR) exons, 3’-UTR exons,introns and intergenic regions based on the gene model provided.This module roughly reflects the uniformity of coverage; for example,reads are generally over-represented in 3’-UTR for the polyA+RNA-seq protocol. One can also apply this module to estimate thebackground noise level.

Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

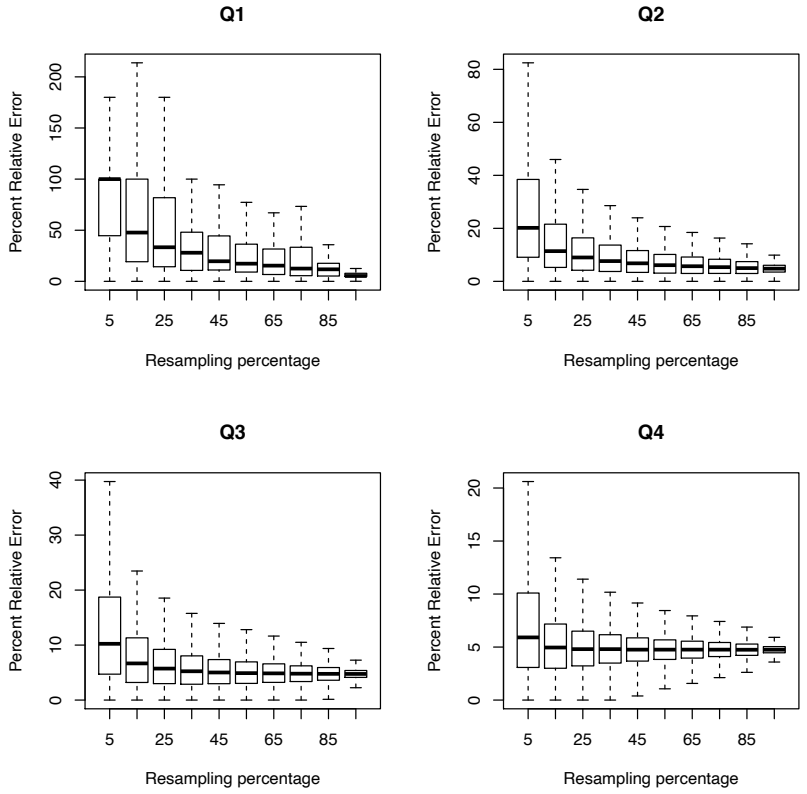
RSeQC

- **RPKM_saturation.py**

'RPKM_saturation.py' determines the precision of estimated RPKMs at the current sequencing depth by resampling (jackknifing) the total mapped reads. We use percent relative error ($100 \times |RPKM_{obs} - RPKM_{real}|/RPKM_{real}$) to measure the precision of estimated RPKM (Fig. 1B). In practice, it is impossible to evaluate $RPKM_{real}$, and we use RPKM estimated from total reads to approximate $RPKM_{real}$.

$$\text{Percent Relative Error} = \frac{|RPKM_{obs} - RPKM_{real}|}{RPKM_{real}} \times 100$$

样本统计（RPKM）的精度受样本大小（测序深度）的影响，**重抽样（resample）**是使用部分数据来评估样本统计量的精度的方法。这个模块从总的RNA reads中重抽样并计算每次的RPKM值，通过这样我们就能**检测当前测序深度是不是够**的(如果测序深度不够RPKM的值将不稳定,如果测序深度足够则RPKM值将稳定)。*默认情况下,这个模块将计算20个RPKM值(分别是对个转录本使用5%,10%,…,95%的总reads*



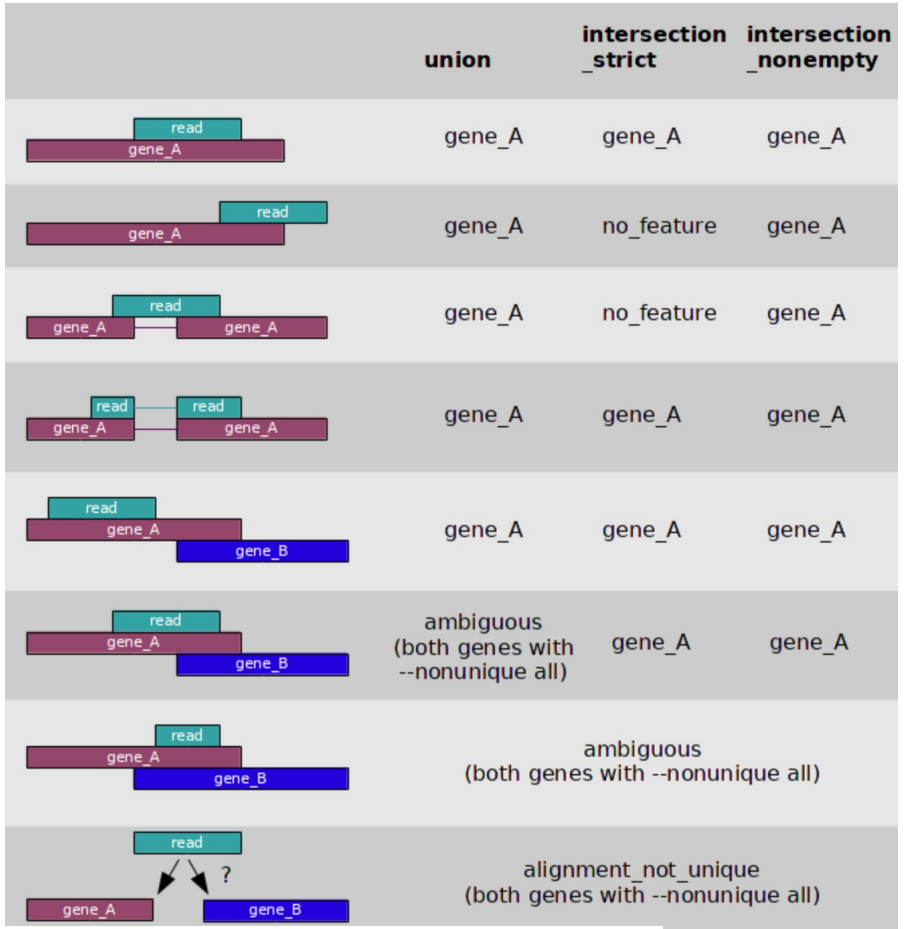
Htseq-counts

- cat ~/test_wdc.txt|while read id; do htseq-count -r pos -f bam -n 2 ~/download/wdc_data/${id}/${id}.Aligned.sortedByCoord.out.bam /home/zhushu/zhouty19990625/genome/mouse/ann/Mus_musculus.GRCm38.102.gtf > ~/download/wdc_data/output2/${id}.count;done

- 教程： https://htseq.readthedocs.io/en/master/count.html

Htseq-counts

- --mode:
- The three overlap resolution modes
- cat ~/test_wdc.txt|while read id; do htseq-count -r pos -f bam -n 2 ~/download/wdc_data/${id}/${id}.Aligned.sortedByCoord.out.bam /home/zhushu/zhouty19990625/genome/mouse/ann/Mus_musculus.GRCm38.102.gtf > ~/download/wdc_data/output2/${id}.count;done



- the union of all the sets *S(i)* for mode `union`. This mode is recommended for most use cases.
- the intersection of all the sets *S(i)* for mode `intersection-strict`.
- the intersection of all non–empty sets *S(i)* for mode `intersection-nonempty`.

Htseq-counts

column 1: gene ID

column 2: counts for unstranded RNA-seq

column 3: counts for the 1st read strand aligned with RNA (htseq-count option -s yes)

column 4: counts for the 2nd read strand aligned with RNA (htseq-count option -s reverse)

-s <yes/no/reverse>, **--stranded**=<yes/no/reverse>
Whether the data is from a strand-specific assay (default: `yes`)

For `stranded=no`, a read is considered overlapping with a feature regardless of whether it is mapped to the same or the opposite strand as the feature. For `stranded=yes` and single-end reads, the read has to be mapped to the same strand as the feature. For paired-end reads, the first read has to be on the same strand and the second read on the opposite strand. For `stranded=reverse`, these rules are reversed.

Htseq-counts

- cat ~/test_wdc.txt|while read id; do htseq-count **-r** pos **-f** bam -n 2 ~/download/wdc_data/${id}/${id}.Aligned.sortedByCoord.out.bam /home/zhushu/zhouty19990625/genome/mouse/ann/Mus_musculus.GRCm38.102.gtf > ~/download/wdc_data/output2/${id}.count;done

-f `<format>`, **--format**=`<format>`

Format of the input data. Possible values are `sam` (for text SAM files) and `bam` (for binary BAM files). Default is `sam`.

DEPRECATED: Modern versions of samtools/htslibs, which HTSeq uses to access SAM/BAM/CRAM files, have automatic file type detection. This flag will be removed in future versions of htseq-count.

-r `<order>`, **--order**=`<order>`

For paired-end data, the alignment have to be sorted either by read name or by alignment position. If your data is not sorted, use the `samtools sort` function of `samtools` to sort it. Use this option, with `name` or `pos` for `<order>` to indicate how the input data has been sorted. The default is `name`.

If `name` is indicated, `htseq-count` expects all the alignments for the reads of a given read pair to appear in adjacent records in the input data. For `pos`, this is not expected; rather, read alignments whose mate alignment have not yet been seen are kept in a buffer in memory until the mate is found. While, strictly speaking, the latter will also work with unsorted data, sorting ensures that most alignment mates appear close to each other in the data and hence the buffer is much less likely to overflow.

整合生成表达矩阵

```
ls ~/download/zwq/mid_data/|while read id;do
cd ${id};
awk '{print $1,$3}' ${id}.ReadsPerGene.out.tab > ~/download/zwq/mid_count/${id}.count;
cd ~/download/zwq/mid_data/;
done
```

```
(base) [zhouty19990625@mgt non_count]$ head   SRR2921588.count  SRR2921611.count  SRR2921634.count  SRR2
921667.count  SRR2921690.count
==> SRR2921588.count <==
N_unmapped 447601
N_multimapping 92362
N_noFeature 711759
N_ambiguous 21442
ENSMUSG00000102693 0
ENSMUSG00000064842 0
ENSMUSG00000051951 0
ENSMUSG00000102851 0
ENSMUSG00000103377 0
ENSMUSG00000104017 0

==> SRR2921611.count <==
N_unmapped 327062
N_multimapping 92435
N_noFeature 581599
N_ambiguous 16831
ENSMUSG00000102693 0
ENSMUSG00000064842 0
ENSMUSG00000051951 0
ENSMUSG00000102851 0
ENSMUSG00000103377 0
ENSMUSG00000104017 0
```

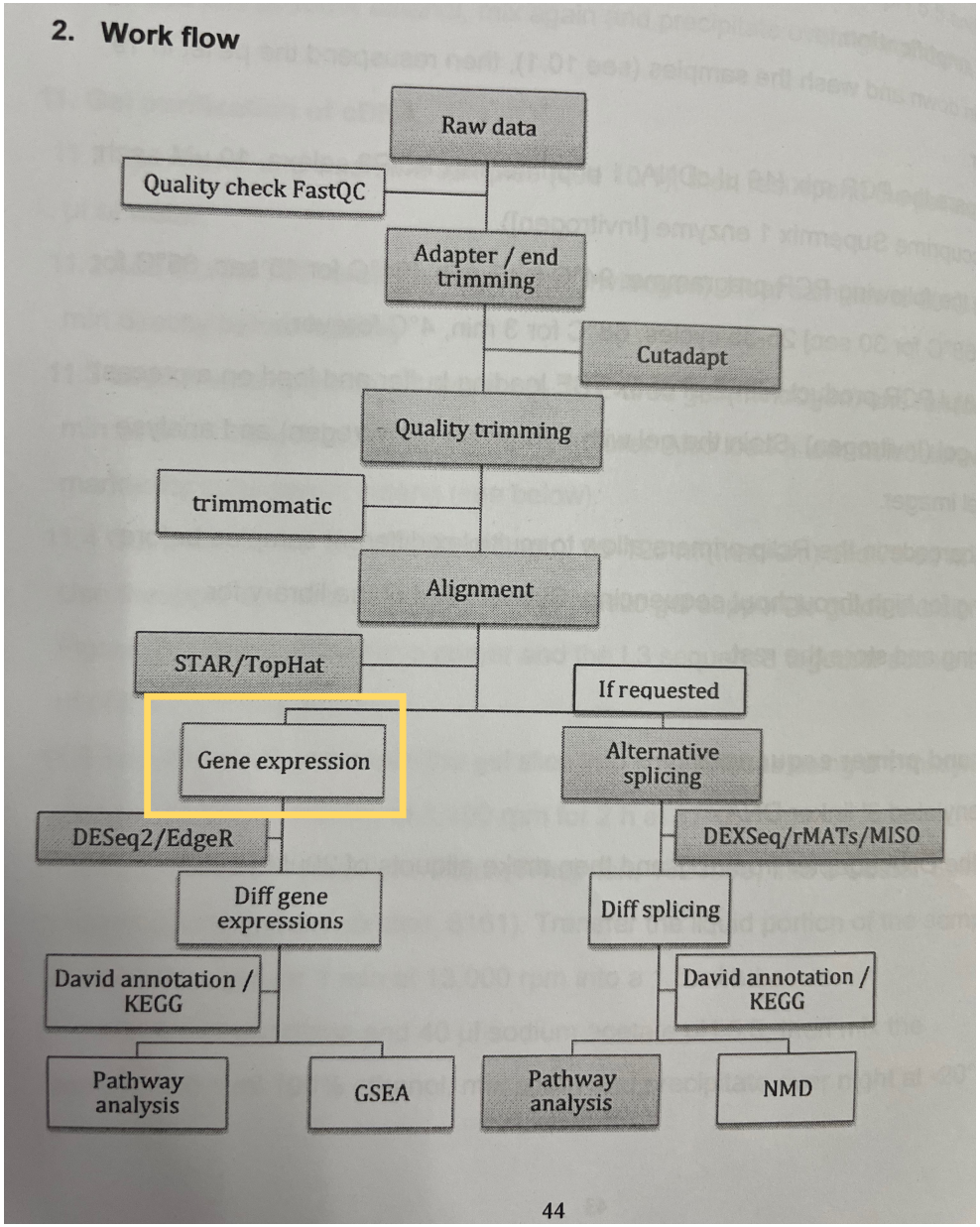
整合生成表达矩阵

```
$perl -lne 'if ($ARGV=~/(.*).count/){print "$1\t$_"}' *.count >matrix.count
$R
>non<-read.csv('/Users/tingyue/Desktop/download/non_matrix.count',header = F,sep= ' \t ')
>colnames(non)=c('sample','gene','count' )
>non_counts=dcast(non,formula=gene~sample)
>non_counts<-non_counts[-c(55488:55491 ),]
>non_counts<-merge(ann3,non_counts,by='gene' )
>non_counts<-as.matrix(non_counts)
>rownames(non_counts)<-non_counts[,2]
>non_counts<-non_counts[,-c(1,2)]
>non<-apply(non_counts,2,as.numeric)
>rownames(non)<-rownames(non_counts)
>library(Matrix)
>nonSM <- as(as.matrix(non), "dgCMatrix")
```

整合生成表达矩阵

	SRR2921562	SRR2921563	SRR2921564	SRR2921565	SRR2921566	SRR2921567	SRR2921568	SRR2921569	SRR2921570	SRR2921571
Gnai3	441	403	443	0	600	138	21	196	247	941
Pbsn	0	0	0	0	0	0	0	0	0	0
Cdc45	115	163	0	116	143	726	20	52	31	1317
H19	0	0	0	0	0	0	0	0	0	0
Scml2	0	5	0	0	0	0	0	0	0	0
Apoh	0	0	0	0	0	0	0	0	0	0
Narf	0	0	210	0	0	104	0	0	563	7
Cav2	0	0	0	0	0	0	0	0	0	0
Klf6	209	2117	639	352	976	1353	72	1144	1414	72
Scmh1	0	0	0	17	0	0	0	0	304	0
Cox5a	569	642	774	194	126	302	456	234	163	424
Tbx2	0	0	0	0	0	0	0	0	0	0
Tbx4	0	0	0	0	0	0	0	0	0	0
Zfy2	0	0	0	0	0	0	0	0	0	0
Ngfr	0	0	0	0	0	0	0	0	0	0

2. Work flow


Raw data
Quality check FastQC
Adapter / end trimming
Cutadapt
Quality trimming
trimmomatic
Alignment
STAR/TopHat
If requested
Gene expression
Alternative splicing
DESeq2/EdgeR
DEXSeq/rMATs/MISO
Diff gene expressions
Diff splicing
David annotation / KEGG
David annotation / KEGG
Pathway analysis
GSEA
Pathway analysis
NMD

DESeq2

下载（a R package）：
https://bioconductor.org/packages/release/bioc/html/DESeq2.html

- 教程：
- 中文教程&使用教程： https://www.jianshu.com/p/2689e9a1d10c
- 说明书&原理：
  https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

DESeq2 --input

- dds<-DESeqDataSetFromMatrix(mRNA_expr_for_DESeq,colData = condition_table,design = ~feature)
- View(mRNA_expr_for_DESeq);View(condition_table)

	MEF.KO.EMCV.6h.1	MEF.KO.EMCV.6h.2	MEF.WT.EMCV.6h.1	MEF.WT.EMCV.6h.2
ENSMUSG00000000001	3836	4224	1210	3904
ENSMUSG00000000028	665	851	760	937
ENSMUSG00000000031	99508	125430	14337	85620
ENSMUSG00000000037	19	46	10	9
ENSMUSG00000000049	7	10	2	8
ENSMUSG00000000056	382	465	68	431
ENSMUSG00000000058	659	766	96	537
ENSMUSG00000000078	2930	3412	2931	3329
ENSMUSG00000000085	424	455	190	436
ENSMUSG00000000088	2196	2807	7969	2743
ENSMUSG00000000093	203	259	195	294
ENSMUSG00000000094	188	295	3	118
ENSMUSG00000000120	895	1116	53	706
ENSMUSG00000000126	211	267	83	225
ENSMUSG00000000127	868	938	150	934
ENSMUSG00000000131	1919	2364	529	2550
ENSMUSG00000000134	1205	1447	823	1480

	feature
MEF.KO.EMCV.6h.1	KO
MEF.KO.EMCV.6h.2	KO
MEF.WT.EMCV.6h.1	WT
MEF.WT.EMCV.6h.2	WT

DESeq2 --condition_table

- condition_table$feature<-as.factor(condition_table$feature)
- condition_table$feature<-relevel(condition_table$feature,ref = 'WT')

	feature
MEF.KO.EMCV.6h.1	KO
MEF.KO.EMCV.6h.2	KO
MEF.WT.EMCV.6h.1	WT
MEF.WT.EMCV.6h.2	WT

DESeq2 --input

	MEF.KO.EMCV.6h.1	MEF.KO.EMCV.6h.2	MEF.WT.EMCV.6h.1	MEF.WT.EMCV.6h.2
ENSMUSG00000000001	3836	4224	1210	3904
ENSMUSG00000000028	665	851	760	937
ENSMUSG00000000031	99508	125430	14337	85620
ENSMUSG00000000037	19	46	10	9
ENSMUSG00000000049	7	10	2	8
ENSMUSG00000000056	382	465	68	431
ENSMUSG00000000058	659	766	96	537
ENSMUSG00000000078	2930	3412	2931	3329
ENSMUSG00000000085	424	455	190	436
ENSMUSG00000000088	2196	2807	7969	2743
ENSMUSG00000000093	203	259	195	294
ENSMUSG00000000094	188	295	3	118
ENSMUSG00000000120	895	1116	53	706
ENSMUSG00000000126	211	267	83	225
ENSMUSG00000000127	868	938	150	934
ENSMUSG00000000131	1919	2364	529	2550
ENSMUSG00000000134	1205	1447	823	1480

	SRR2921562	SRR2921563	SRR2921564	SRR2921565	SRR2921566	SRR2921567	SRR2921568	SRR2921569	SRR2921570	SRR2921571
Gnai3	441	403	443	0	600	138	21	196	247	941
Pbsn	0	0	0	0	0	0	0	0	0	0
Cdc45	115	163	0	116	143	726	20	52	31	1317
H19	0	0	0	0	0	0	0	0	0	0
Scml2	0	5	0	0	0	0	0	0	0	0
Apoh	0	0	0	0	0	0	0	0	0	0
Narf	0	0	210	0	0	104	0	0	563	7
Cav2	0	0	0	0	0	0	0	0	0	0
Klf6	209	2117	639	352	976	1353	72	1144	1414	72
Scmh1	0	0	0	17	0	0	0	0	304	0
Cox5a	569	642	774	194	126	302	456	234	163	424
Tbx2	0	0	0	0	0	0	0	0	0	0
Tbx4	0	0	0	0	0	0	0	0	0	0
Zfy2	0	0	0	0	0	0	0	0	0	0
Ngfr	0	0	0	0	0	0	0	0	0	0

Why un-normalized counts?

As input, the DESeq2 package expects count data as obtained, e.g., from RNA-seq or another high-throughput sequencing experiment, in the form of a matrix of integer values. The value in the *i*-th row and the *j*-th column of the matrix tells how many reads can be assigned to gene *i* in sample *j*. Analogously, for other types of assays, the rows of the matrix might correspond e.g. to binding regions (with ChIP-Seq) or peptide sequences (with quantitative mass spectrometry). We will list method for obtaining count matrices in sections below.

The values in the matrix should be un-normalized counts or estimated counts of sequencing reads (for single-end RNA-seq) or fragments (for paired-end RNA-seq). The RNA-seq workflow describes multiple techniques for preparing such count matrices. It is important to provide count matrices as input for DESeq2's statistical model (Love, Huber, and Anders 2014) to hold, as only the count values allow assessing the measurement precision correctly. The DESeq2 model internally corrects for library size, so transformed or normalized values such as counts scaled by library size should not be used as input.

DESeq2

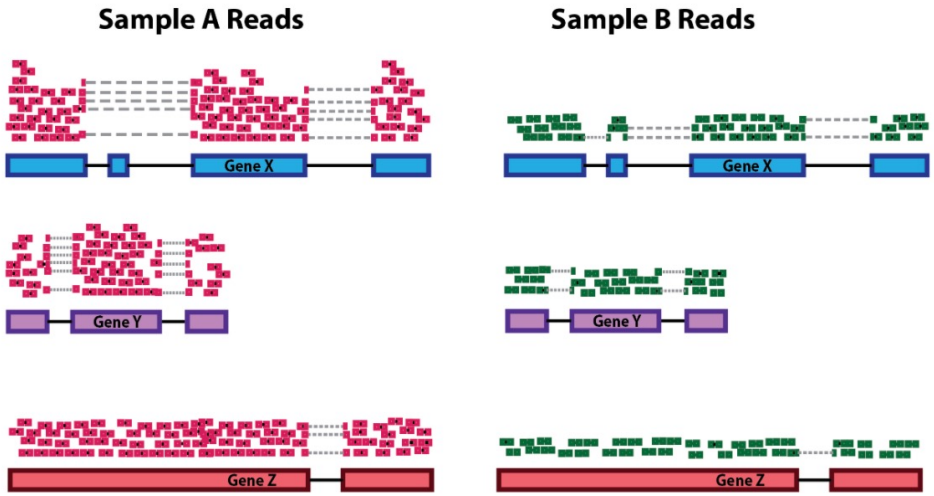
- DESeq2只能Read counts作为输入，而其他经过normolize的数据不行
- Explore different types of normalization method:
- https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

- CPM, TPM, FPKM/RPKM, TMM(edgeR), DESeq2 **median of ratios ……**

???

Why normalize?

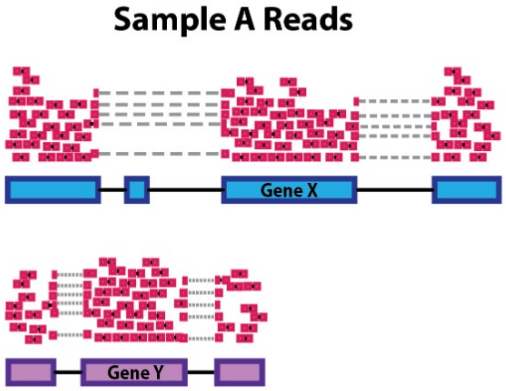
- **Sequencing depth:** Accounting for sequencing depth is necessary for comparison of gene expression between samples. In the example below, each gene appears to have doubled in expression in *Sample A* relative to *Sample B*, however this is a consequence of *Sample A* having double the sequencing depth.



> ***NOTE:*** *In the figure above, each pink and green rectangle represents a read aligned to a gene. Reads connected by dashed lines connect a read spanning an intron.*

Why normalize?

- **Gene length:** Accounting for gene length is necessary for comparing expression between different genes within the same sample. In the example, *Gene X* and *Gene Y* have similar levels of expression, but the number of reads mapped to *Gene X* would be many more than the number mapped to *Gene Y* because *Gene X* is longer.

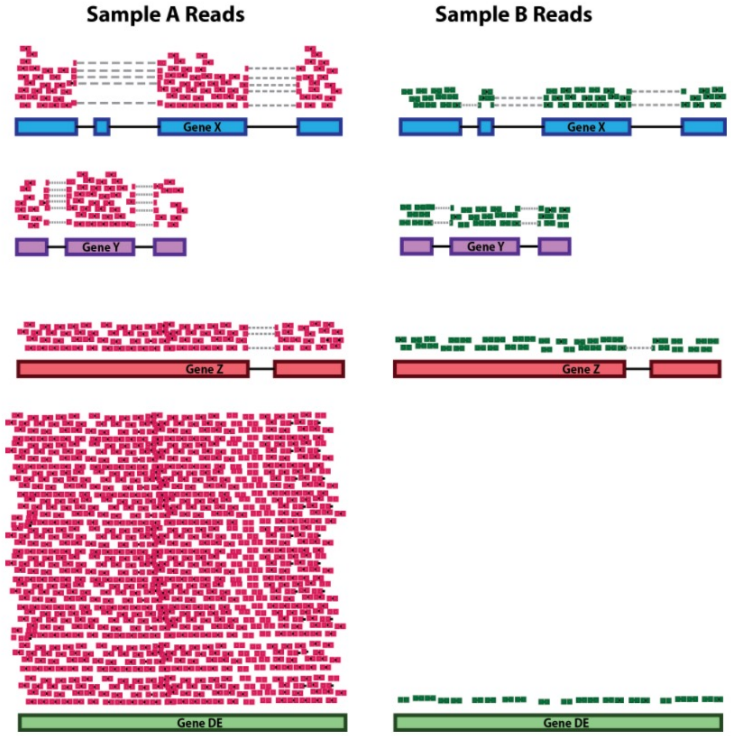


Why normalize?

While normalization is essential for differential expression analyses, it is also necessary for exploratory data analysis, visualization of data, and whenever you are exploring or comparing counts between or within samples.

- **RNA composition:** A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods. Accounting for RNA composition is recommended for accurate comparison of expression between samples, and is particularly important when performing differential expression analyses [1].

In the example, if we were to divide each sample by the total number of counts to normalize, the counts would be greatly skewed by the DE gene, which takes up most of the counts for *Sample A*, but not *Sample B*. Most other genes for *Sample A* would be divided by the larger number of total counts and appear to be less expressed than those same genes in *Sample B*.



Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

DESeq2’ s median of ratios

- dds <- estimateSizeFactors(dds)
- sizeFactors(dds)
- nor_non<-counts(dds,normalized=TRUE)

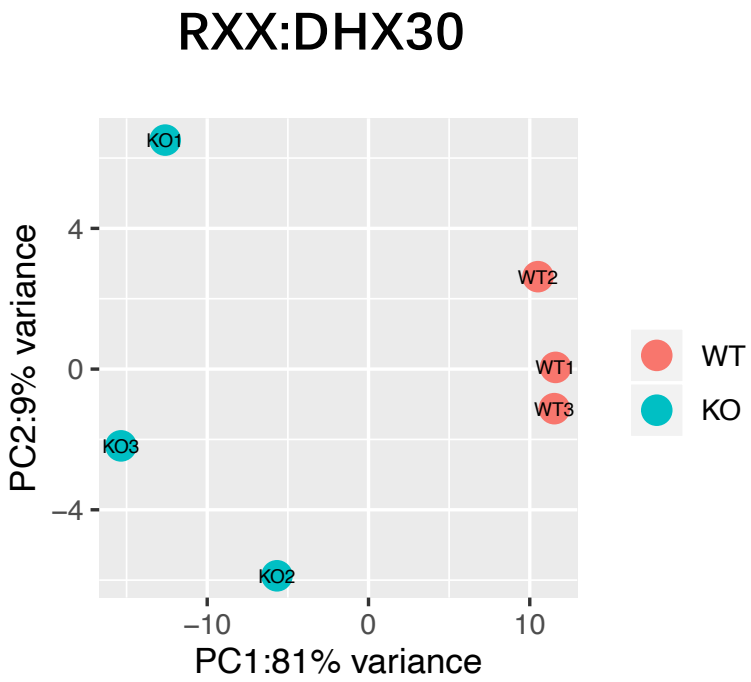
- **Step 1: creates a pseudo-reference sample (row-wise geometric mean)**
- **Step 2: calculates ratio of each sample to the reference**
- **Step 3: calculate the normalization factor for each sample (size factor)**
- **Step 4: calculate the normalized count values using the normalization factor**

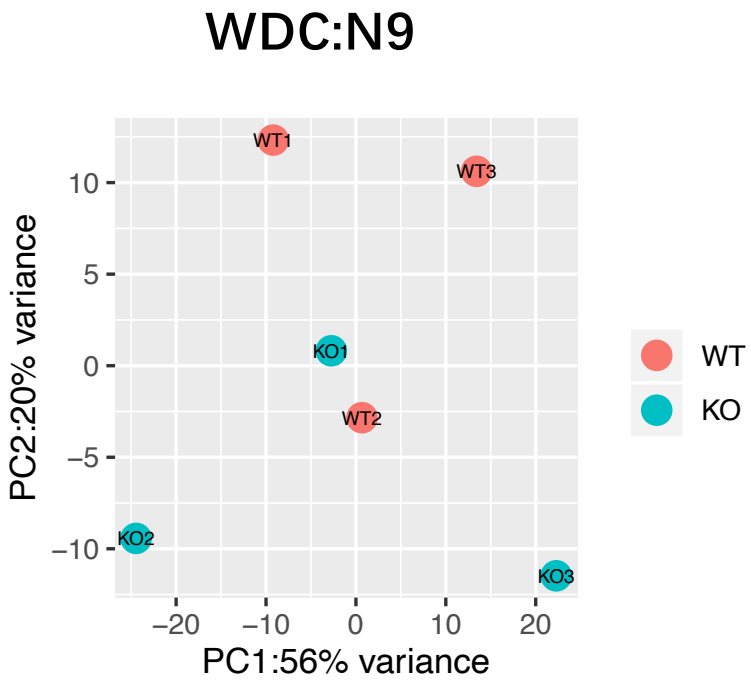
*无法组内比较基因表达的差异！（e.g. 没有考虑基因的长度）

DESeq2 --PCA

- res<-vst(dds_DE,blind = F)
- exp<-assay(res)
- par(cex = 0.7)
- par(mar=c(1,3,1,1))
- n.sample=ncol(exp)
- cols <- rainbow(n.sample*1.2)
- boxplot(exp,las=2,col=cols,)
- PCA_data<-plotPCA(res,intgroup='feature',returnData=T)
- plotPCA(res,intgroup='feature')

DESeq2 --PCA


RXX:DHX30
KO1
WT2
WT1
WT3
KO3
KO2
4
0
−4
−10
0
10
PC2:9% variance
PC1:81% variance
WT
KO


WDC:N9
WT1
WT3
KO1
WT2
KO2
KO3
10
5
0
−5
−10
−20
−10
0
10
20
PC2:20% variance
PC1:56% variance
WT
KO

DESeq2 --MDS, PCoA……

- mdsdata <- data.frame(cmdscale(sampleDistMatrix))
  cmdscale(classical multidimensional scaling)
- MDSmds <- cbind(mdsdata,as.data.frame(colData(vsd)))
  ggplot(data=mds,aes(X1,X2,color=cell,shape=dex)) +
  geom_point(size=3)

DESeq2

- dds<-DESeqDataSetFromMatrix(mRNA_expr_for_DESeq,colData = condition_table,design = ~feature)
- dds_DE<-DESeq(dds)
- res_DE<-results(dds_DE,alpha = 0.05,contrast = c('feature','KO','WT'))

DESeq2

- 返回一个数据框res，包含6列：**baseMean、log2FC、lfcSE、stat、pvalue**、baseMean表示所有样本经过归一化系数矫正的read counts（counts/sizeFactor）**padj**
- 的均值。baseMean = apply(normalized_counts, 1, mean)。
- log2Foldchange表示该基因的表达发生了多大的变化，是估计的效应大小effect size。对差异表达的倍数取以2为底的对数，log2FC反映的是不同分组间表达量的差异，这个差异由两部分构成，一种是样本间本身的差异，比如生物学重复样本间基因的表达量就有一定程度的差异，另外一部分就是我们真正感兴趣的，由于分组不同或者实验条件不同造成的差异。用归一化之后的数值直接计算出的log2FC包含了以上两种差异，而我们真正感兴趣的只有分组不同造成的差异，DESeq2在差异分析的过程中已经考虑到了样本本身的差异，其最终提供的log2FC只包含了分组间的差异，所以会与手动计算的不同）。
- lfcSE(logfoldchange Standard Error)是对于log2Foldchange估计的标准误差估计，效应大小估计有不确定性。stat是Wald统计量，它是由log2Foldchange除以标准差所得。
- pvalue和padj分别代表原始的p值以及经过校正后的p值。

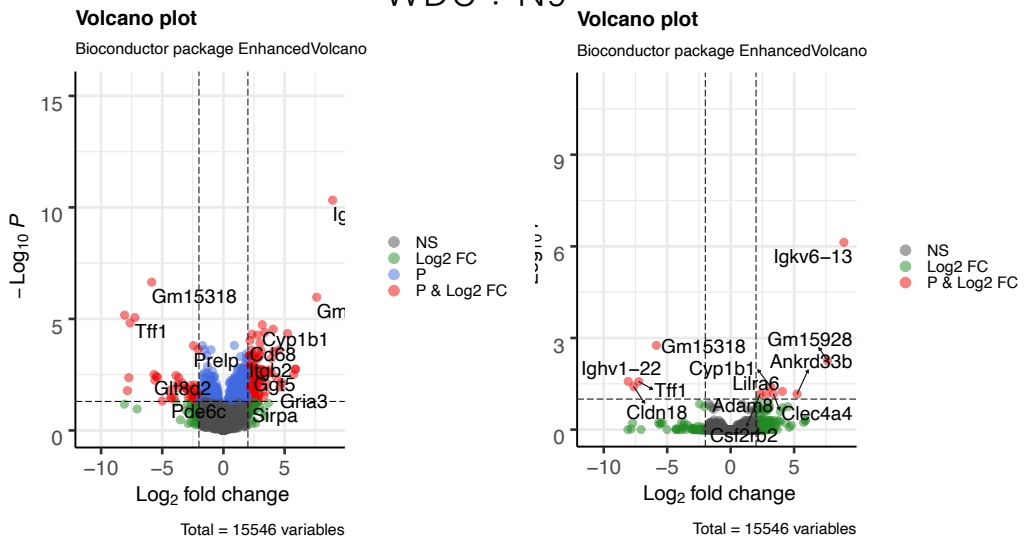
DESeq2

- dds<-DESeqDataSetFromMatrix(mRNA_expr_for_DESeq,colData = condition_table,design = ~feature)
- dds_DE<-DESeq(dds)
- res_DE<-results(dds_DE,alpha = 0.05,contrast = c('feature','KO','WT'))
- for_volcano<-data.frame('log2FoldChange'=res_DE$log2FoldChange,
- 'padj'=res_DE$padj,
- 'pvalue'=res_DE$pvalue,
- 'descrip'<-rep('no',length(res_DE$log2FoldChange)),
- 'gene_name'<-rownames(res_DE))

DESeq2_ Visualization

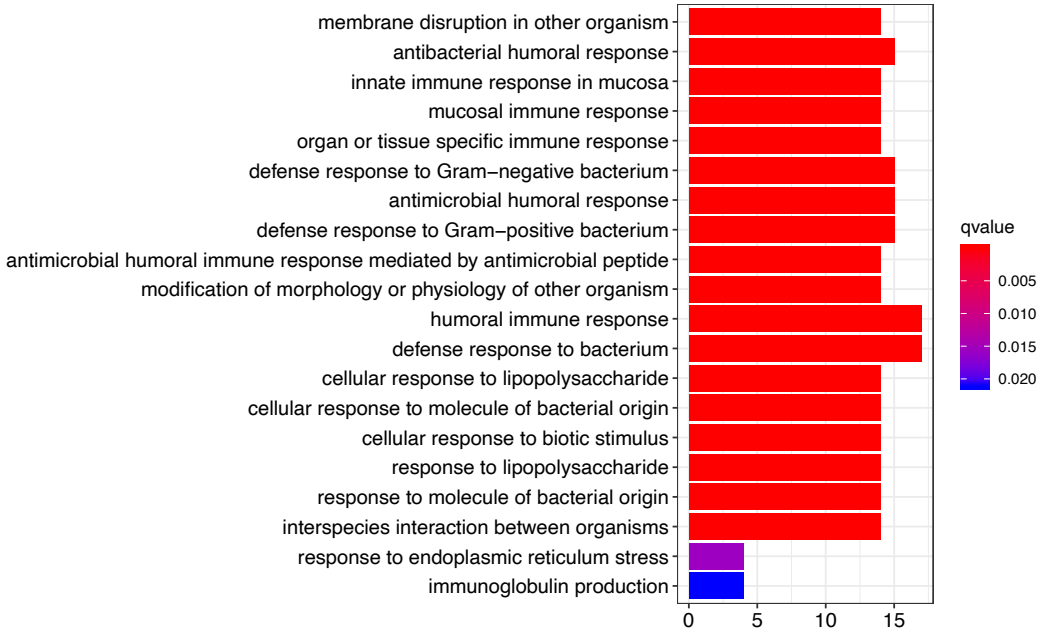
- rownames(for_volcano)<-rownames(res_DE)
- names(for_volcano)[4]<-'descrip'
- up_sig<-intersect(which(for_volcano$log2FoldChange>2),which(for_volcano$padj<0.1)) #满足这些要求的行名
- down_sig<-intersect(which(for_volcano$log2FoldChange<(-2)),which(for_volcano$padj<0.1)) #
- for_volcano$descrip<-as.character(for_volcano$descrip)
- for_volcano[up_sig,'descrip']<-'Up'
- for_volcano[down_sig,'descrip']<-'Down'
- for_volcano$descrip<-as.factor(for_volcano$descrip)
- colnames(for_volcano)[5]<-'gene_id'
- for_volcano<-merge(for_volcano,ann3,by='gene_id')
- **EnhancedVolcano(for_volcano,**
- **lab = for_volcano$gene_symbol,**
- **x = 'log2FoldChange',**
- **y = 'padj')**

WDC：N9



DESeq2_ Visualization

- ego<-enrichGO(gene = **up_gene$ENTREZID**,
- OrgDb = 'org.Mm.eg.db',
- ont = 'BP',
- pAdjustMethod = 'BH',
- pvalueCutoff = 0.05,
- qvalueCutoff = 0.1,
- readable = T)
- ggo<-gseGO(geneList = glist,
- OrgDb = 'org.Mm.eg.db',
- ont = 'BP',
- minGSSize = 10,
- maxGSSize = 500,
- pvalueCutoff = 0.5)



Batch effects

- ds<-DESeqDataSetFromMatrix(non,colData = mid_meta,design = ~batch)
- View(counts(dds))
- dds <- estimateSizeFactors(dds)
- sizeFactors(dds)
- nor_non<-counts(dds,normalized=TRUE)
- modcombat = model.matrix(~1, data = mid_meta)
- batch = mid_meta$batch
- combat_edata = ComBat(dat=nor_non, batch=batch, mod=modcombat,par.prior=TRUE, prior.plots=FALSE)

STAR: Results

```
#总用量 8.3G
#-rw-r--r-- 1 zhouty19990625 zhushu  2.3G 2月  10 10:46 S316WT_L2_338X38.Aligned.sortedByCoord.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  4.1G 2月  10 10:45 S316WT_L2_338X38.Aligned.toTranscriptome.out.bam
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.final.out
#-rw-r--r-- 1 zhouty19990625 zhushu   17K 2月  10 10:46 S316WT_L2_338X38.Log.out
#-rw-r--r-- 1 zhouty19990625 zhushu  2.0K 2月  10 10:46 S316WT_L2_338X38.Log.progress.out
#-rw-r--r-- 1 zhouty19990625 zhushu  1.4M 2月  10 10:45 S316WT_L2_338X38.ReadsPerGene.out.tab
#-rw-r--r-- 1 zhouty19990625 zhushu  6.3M 2月  10 10:45 S316WT_L2_338X38.SJ.out.tab
#drwx------ 3 zhouty19990625 zhushu  4.0K 2月  10 10:46 S316WT_L2_338X38._STARtmp
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate1
#-rw-r--r-- 1 zhouty19990625 zhushu 1023M 2月  10 10:45 S316WT_L2_338X38.Unmapped.out.mate2
```

STAR: Mapping

- --outFilterType BySJout

STAR: Mapping

STAR: Mapping

STAR: Mapping

STAR: Mapping

STAR: Mapping