RNA-SEQ

Group 2 刘曦瑞 吴仁杰 徐震洋 谢天辰

2022.6.23

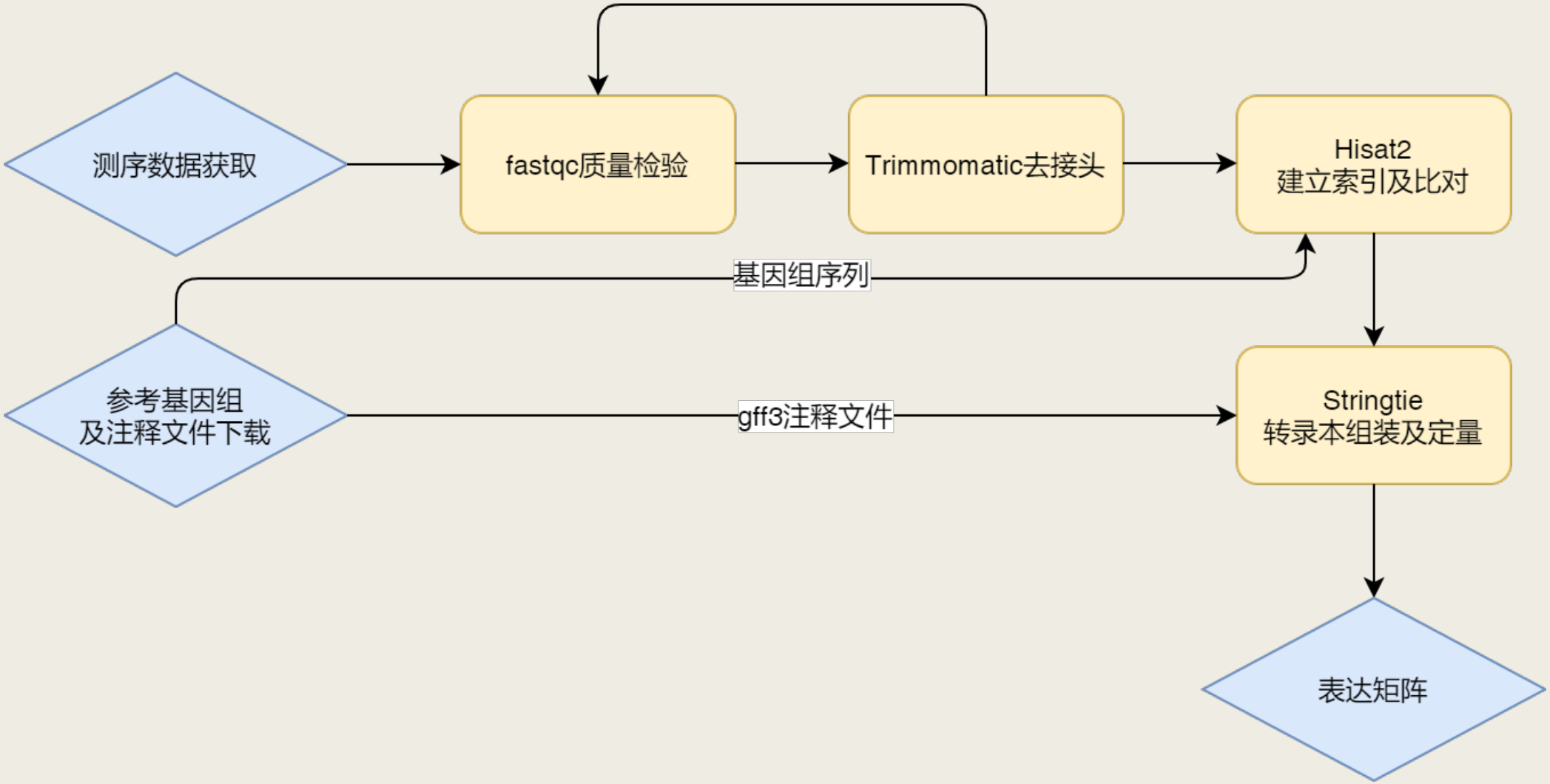
什么是RNA seq

利用转录组数据来识别转录本和表达定量，从而对造成细胞/组织/个体间**不同状态的差异**的内部原因进行诊断分析，挖掘关键基因：

1. 同一物种，不同组织：研究基因在不同组织的表达情况，找到细胞组织特异性的基因；
2. 同一物种，同一组织：研究基因在不同处理或条件下的表达变化，挖掘特异的功能基因，指导后续物种改良、疾病诊断等；
3. 同一组织，不同物种：研究基因的进化关系；
4. 时间序列实验：基因在不同时期的表达情况与其发育的关系，找到发育阶段特异性的基因；

https://www.jianshu.com/p/dce7ed2cc2d0

实验流程


测序数据获取
fastqc质量检验
Trimmomatic去接头
Hisat2
建立索引及比对
基因组序列
参考基因组
及注释文件下载
gff3注释文件
Stringtie
转录本组装及定量
表达矩阵

获取数据

- 测序数据来自*Arabidopsis* Duodecuple Mutant of PYL ABA Receptors Reveals PYL Repression of ABA-Independent SnRK2 Activity。

13	SRR7160928	SAMN09205655	5.27 G	1.84 Gb	SRX4080143	wildtype	GSM3140728	GSM3140728	300 mM mannitol for 24 hours
14	SRR7160929	SAMN09205654	4.80 G	1.70 Gb	SRX4080144	wildtype	GSM3140729	GSM3140729	300 mM mannitol for 24 hours
15	SRR7160930	SAMN09205653	5.03 G	1.73 Gb	SRX4080145	wildtype	GSM3140730	GSM3140730	300 mM mannitol for 24 hours
16	SRR7160931	SAMN09205652	4.58 G	1.60 Gb	SRX4080146	wildtype	GSM3140731	GSM3140731	1/2 MS for 24 hours
17	SRR7160932	SAMN09205651	5.00 G	1.72 Gb	SRX4080147	wildtype	GSM3140732	GSM3140732	1/2 MS for 24 hours
18	SRR7160933	SAMN09205650	4.71 G	1.64 Gb	SRX4080148	wildtype	GSM3140733	GSM3140733	1/2 MS for 24 hours

- 通过通过 NCBI 数据库及文章后列出的 GEO 收录号 GSE114379 查找得到 SRA 序列号，在安装sra-tools后通过prefetch进行下载。

```
1 prefetch SRR7160928
2 输出：SRR7160928.sra
```

获取数据

- 所得sra文件再通过 fastq-dump 或 fasterq-dump 软件解压得到双端测序fastq文件。

```
fasterq-dump -e 4
     -p #显示进度条
     -3(--split-3) #分开两个read
     --gzip #输出压缩文件
     -o <outfile>
     SRR7160928.sra #之前下载的sra数据
输出：SRR7160928_1.fastq SRR7160928_2.fastq
```

获取数据

- 参考基因组选自TAIR www.arabidopsis.org/Genes。
- 下载的数据包括拟南芥基因组 gff3/gtf 注释文件与基因组序列 .fas。

```
#基因组注释文件，提供剪接位点、外显子与Stringtie所需注释信息
wget https://www.arabidopsis.org/download_files/Genes/Araport11_genome_release/Araport11_GFF3_genes_transposons.May2022.gff.gz
wget https://www.arabidopsis.org/download_files/Genes/Araport11_genome_release/Araport11_GTF_genes_transposons.May2022.gtf.gz

#依据基因组序列文件创建Hisat2索引
wget https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_chromosome_files/TAIR10_chr_all.fas
```

去接头——Trimmomatic

- 在进行后续比对定量过程前，首先要去除测序过程中的接头与污染序列：包括PCR引物、adapter、FlowCell序列等。Trimmomatic 安装时自带 Illumina 测序可能的接头/污染序列文件。

注意事项：

1. 注意选择单端(SE)或双端(PE)测序模式
2. 根据测序种类选取对应的文件：GAII 选择 TruSeq2 文件；HiSeq 与 MiSeq 选择 TruSeq3 文件。或在Illumina网站上直接查询。
3. 使用时还需注意测序数据 .fastq文件所使用的质量值体系，分为 phred33 与 phred64。目前常见的测序结果均为 phred33 格式。

去接头——Trimmomatic

```
trimmomatic
    PE data/${input_file}_1.fastq data/${input_file}_2.fastq
        #双端测序
    -phred33 #phred33质量值体系
    -baseout workdir/$input_file/$input_file
    -threads 4
    ILLUMINACLIP:./genome_annotation/TruSeq-PE.fa:2:30:10
        #HiSeq测序仪，选择TruSeq3文件
    LEADING:5 TRAILING:5 SLIDINGWINDOW:5:20
    MINLEN:36 AVGQUAL:20 #质量控制参数
```

- 输出1P、1U、2P、2U四个标准文件。1U与2U为保留单端的读段，1P与2P分别为正向/反向均保留的读段。

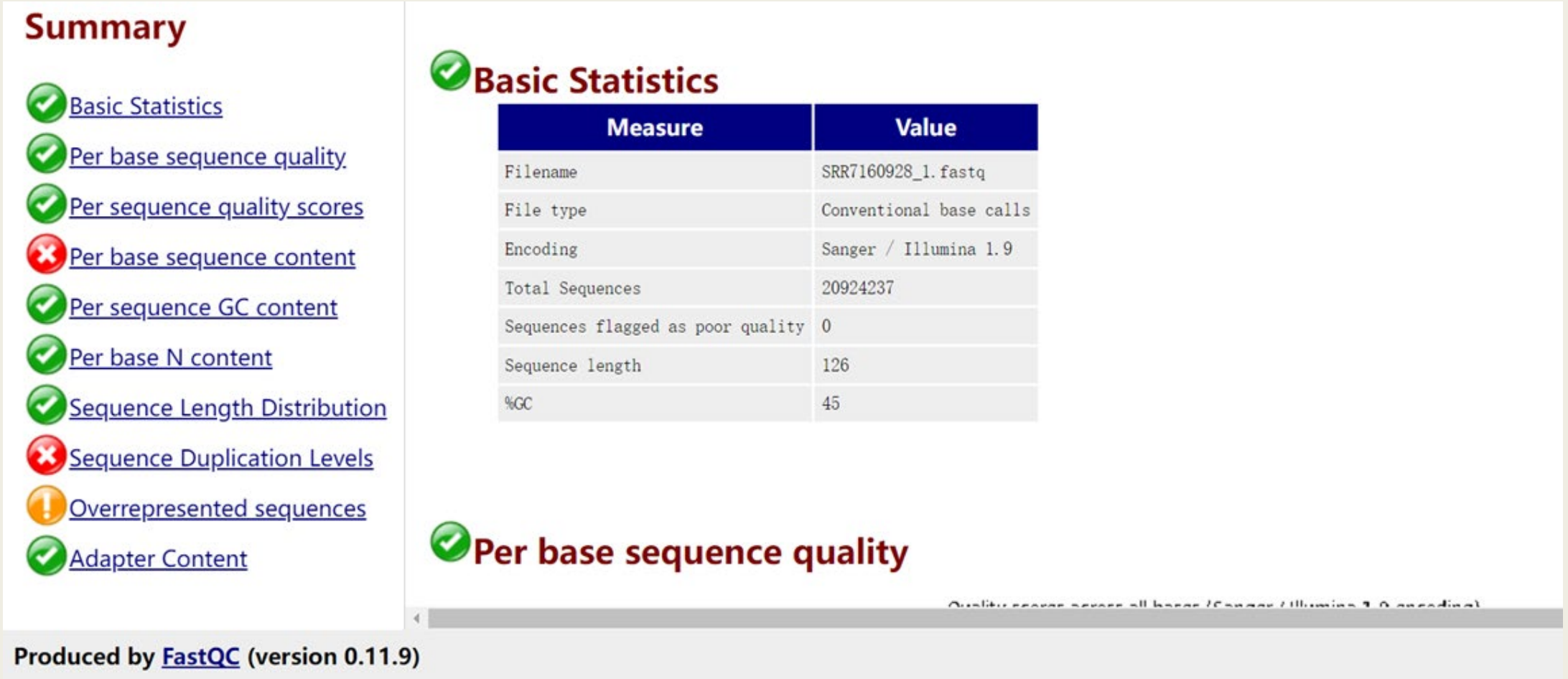
质量控制检验——FastQC

- 在去接头前/后，我们需要对测序数据质量进行检验，以评估是否能达到后续处理的要求，进一步决定是否要更换预处理方法或参数等等/重新提交材料测序。
- FastQC是一个依赖Java环境的高通量序列数据的质量检测工具，用于快速了解fastq数据是否存在问题。

```
fastqc -t 4 \
       -o ${workdir} \
       ${workdir}/${input_file}_1P \
       ${workdir}/${input_file}_2P \
# 输入文件为上一步trimmomatic去接头后得到的两个paired fastq文件
# -t 指定线程数
# -o 指定输出文件存放位置
```

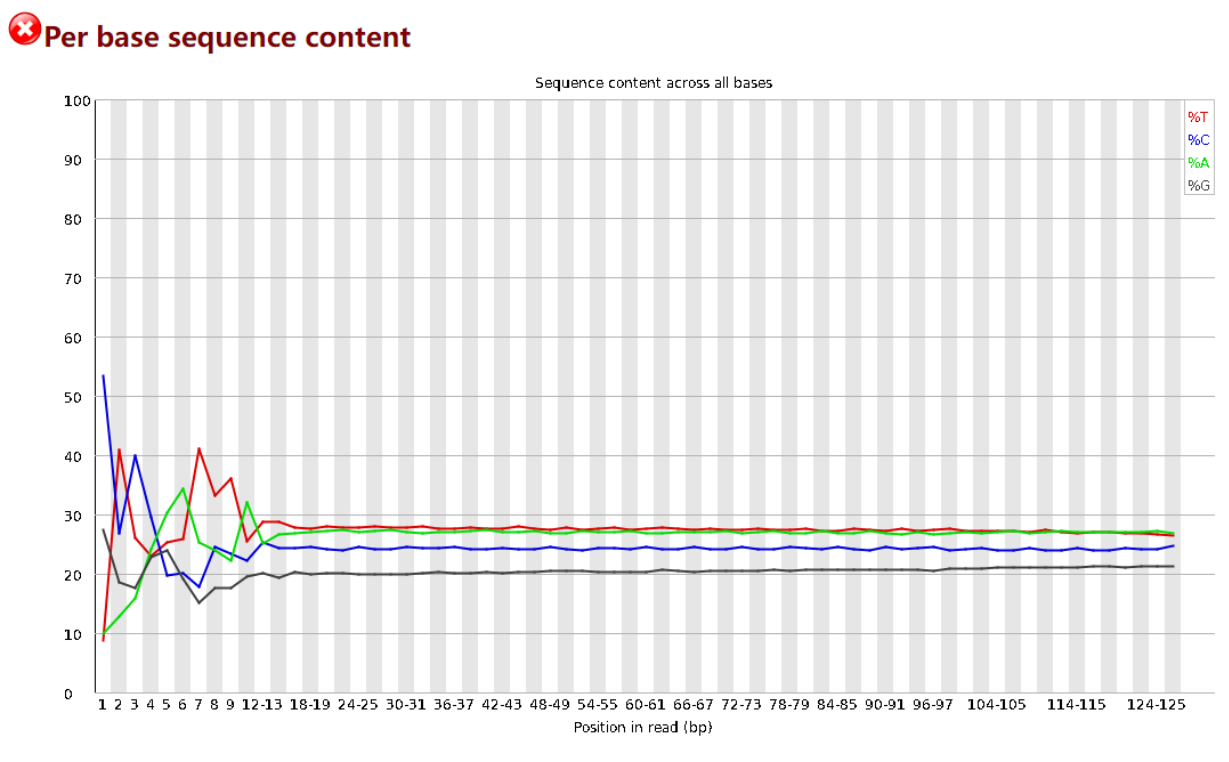
质量控制检验——FastQC

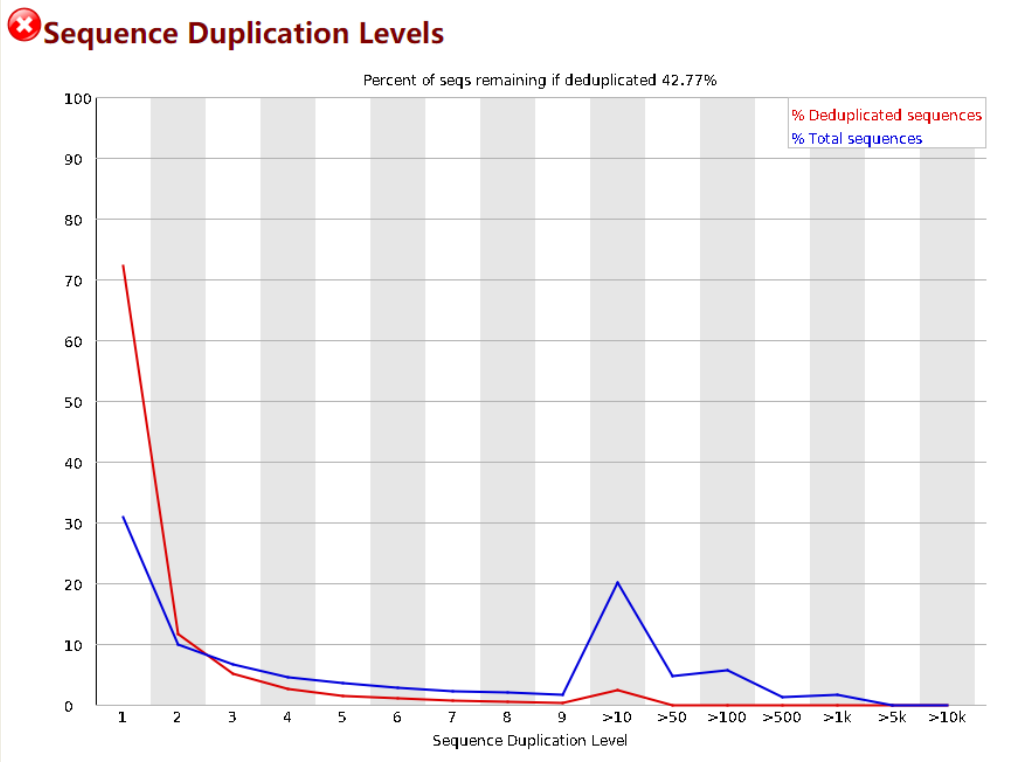
- 打开html报告可以看到直观的质检结果。左栏的Summary中结果分为绿色的PASS，黄色的WARN和红色的FAIL。



质量控制检验——FastQC

- 去接头后Per base sequence content与Sequence Duplication Levels仍报告FAIL。这可能是建库过程存在偏差或文库遭到污染所造成的，并非去接头能改变的。



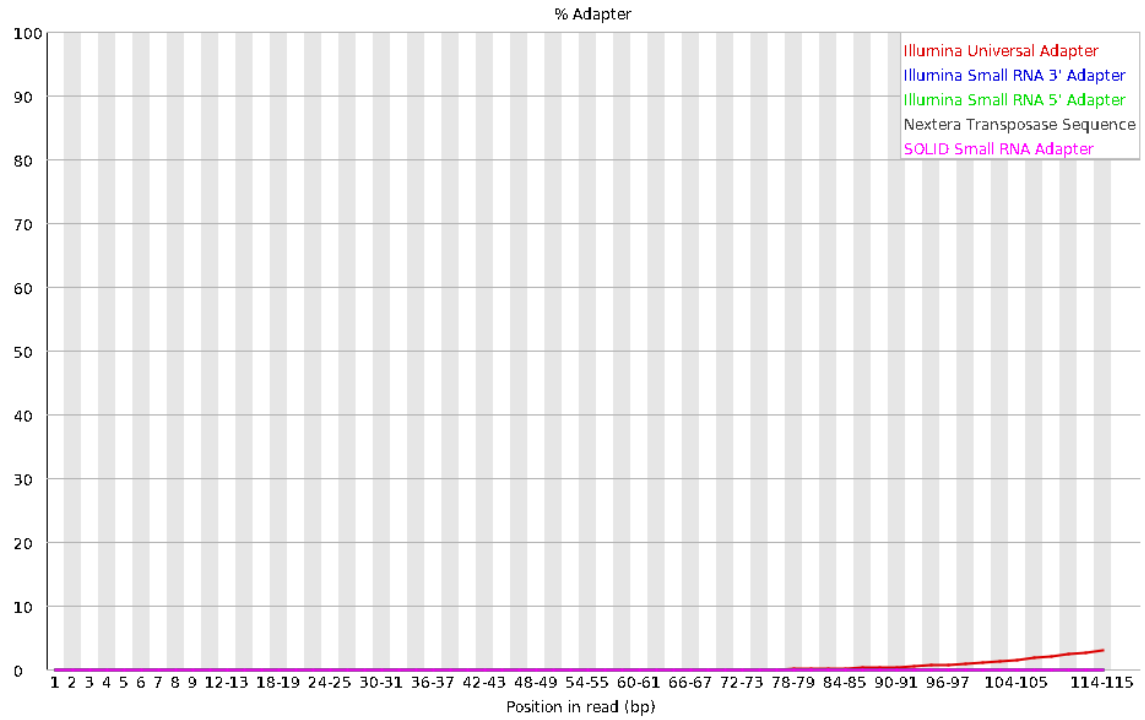


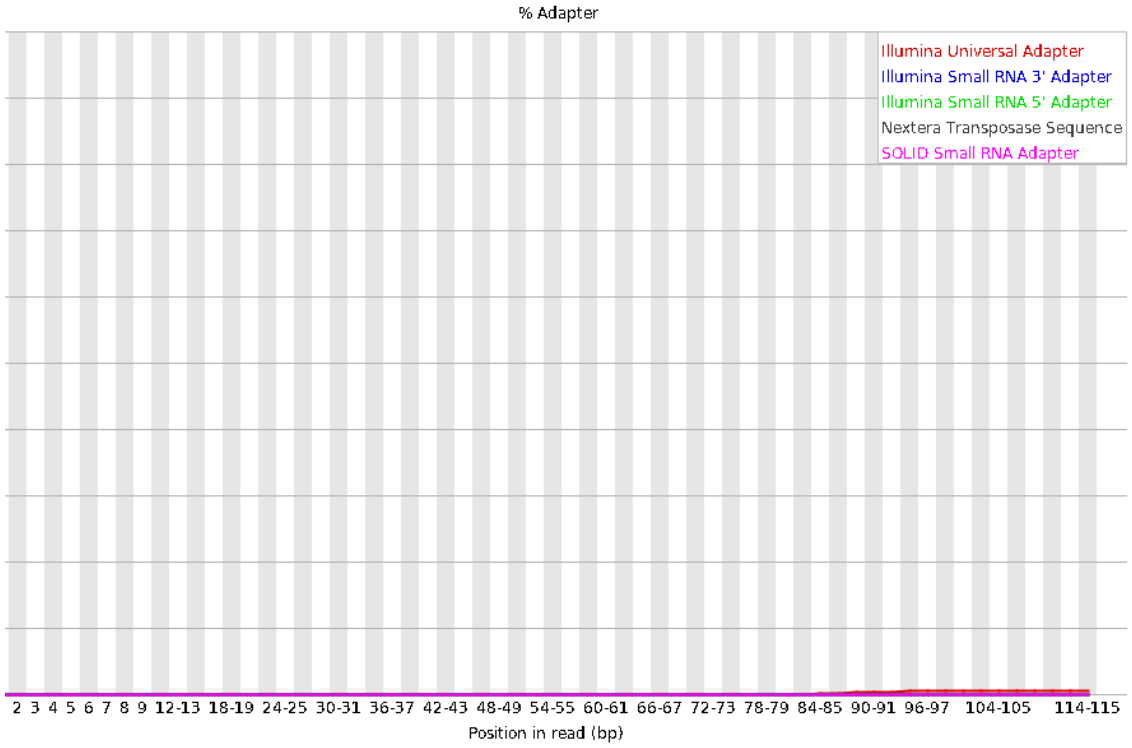
质量控制检验——FastQC

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
ATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCC	23029	0.11005897132593175	TruSeq Adapter, Index 10 (100% over 50bp)

Adapter Content





比对——Hisat2

- 去除接头后，需要将cDNA reads比对的参考基因组上，以便确定不同基因的表达量。Hisat2比对拥有高敏感性和比对速度快的优点，同时方便后续使用Stringtie继续处理。

- 在使用Hisat2比对前，需要获得参考基因组的index文件。先对注释文件进行处理以得到外显子和剪接位点信息。使用Hisat2提供的python脚本文件完成这一操作：

```
1 extract_exons.py Araport11_GTF_genes_transposons.Apr2022.gtf > genome.exon.gtf
2 extract_splice_sites.py Araport11_GTF_genes_transposons.Apr2022.gtf > splice_site.txt
```

索引构建——Hisat2

- 使用Hisat2建立索引。这一步需要用到之前生成的两个分别包含外显子和剪切位点信息的文件，以及在下载的数据包中包括的.fas基因组序列文件。

```
hisat2-build \
    -p 16 \
    --ss splice_site.txt --exon genome.exon.gtf \
    TAIR10_chr_all.fas \
    ./genome #生成索引的前缀
```

- 运行完成后生成8个.ht(genome.*.ht) 索引文件。

比对——Hisat2 & Samtools

- 输入去除接头后的数据，输出比对到基因组的sam文件，由于sam文件占用空间较大，此步骤可以与bam压缩一同完成。

```
hisat2 \
    -p 16 \
    -x genome_annotation/genome #Hisat2索引地址
    --known-splicesite-infile genome_annotation/splice_sites.txt
        #上一步骤生成的剪接位点文件
    --rna-strandness RF #fr-firststrand建库方式（d-UTP）
    --dta #适用于转录本组装(Stringtie)的输出形式
    -1 SRRnumber_1P -2 SRRnumber_2P > hisat_out_sam
```

```
(samtools sort #注意stringtie只接受排序后/按染色体索引好的bam文件
    -@ 16 -l 9 #最大等级压缩
    -o bam -T tmp hisat_out_sam > hisat_out_bam) > hisat_map_info 2>&1
```

转录本组装定量——Stringtie

- 最后需要根据比对到基因组的reads进行统计，确定对应的基因/转录本的表达量，这一过程被称作转录本组装与定量。Stringtie有较高的灵敏度和准确度，且运行速度远快于cufflinks等，适合本次实验。
- 注意：此步骤使用gff3格式注释，否则gtf会因为包含重复而报错。

```
stringtie ${workdir}/${input_file}.bam \
    -p 4 --rf \ #fr-firststrand建库方式
    -e \ #仅处理与-G给出的参考转录本匹配的部分，加速
    -o ${workdir}/${input_file}.gtf \
    -G genome_annotation/Araport11_GFF3_genes_transposons.May2022.gff \
    -b ${workdir} 2> ${workdir}/stringLog.txt
```

- 输出可用于Ballgown进行差异表达分析的五个.ctab文件，与包含转录本信息的transcript.gtf文件。

转录本组装定量——Stringtie

- 最后如需使用其他软件进行差异表达分析，可使用Stringtie提供的prepDE.py3脚本将gtf输出转换为csv格式的表达矩阵：

```
prepDE.py \
    -i smlist \ #包含所有SRR序列号与.gtf输出的地址
    -l 126 #read平均长度
```

- 注意：read平均长度作为计数的参数之一。实验中去接头后read长度为126，如果使用默认值75会导致所有基因的计数等比例缩小。

	SRR7160928	SRR7160929	SRR7160930	SRR7160931	SRR7160932	SRR7160933
AT1G01010|2200934,UniProt=Q0WV96	643	445	523	285	335	293
AT1G01020	336	344	333	354	330	407
AT1G01020|2200939,UniProt=Q5MK24	116	76	108	35	123	70
AT1G03987.1	0	0	0	0	0	0
AT1G01030|2200949,UniProt=Q9MAN1	93	79	123	92	100	116
AT1G01030	21	43	0	14	0	0
AT1G03997	0	0	0	0	0	0
AT1G01040	1683	1614	1601	1661	1881	1744
AT1G03993	1	1	1	1	1	0
AT1G01046	42	34	43	43	52	51
STRG.6	1	1	1	1	1	1
AT1G01050	2292	2456	2361	2291	2494	2371
AT1G01060	306	416	493	605	882	455
AT1G01060|2200969,UniProt=Q6R0H1	319	206	118	290	87	292
AT1G01060|1005715162,UniProt=Q6R0H1	48	20	44	32	75	107
AT1G01070	614	545	528	207	204	179
AT1G01080	511	265	316	914	1122	880
AT1G01080|2200974,UniProt=Q8W592	259	517	502	612	418	1021
AT1G01080|1009021145,UniProt=F4HQH8	232	38	25	351	533	219

差异表达分析

- 差异表达分析通过本地R中DESeq2包来完成。下图为部分结果展示：