

# 实验一 测序数据的质量控制与预处理

## 实验目的

- i. 掌握测序数据的FASTQ格式。
- ii. 熟悉FastQC、FASTX-Toolkit等质量控制工具的使用方法。
- iii. 熟悉Galaxy的使用方法。
- iv. 了解FastQC输出结果的含义。

## 实验材料

- i. [sample.fastq](#)
- ii. [sample2.fastq](#)

## 实验工具

- i. [Galaxy](#)
- ii. [FastQC](#)
- iii. [FASTX-Toolkit](#)

## 实验步骤

- i. Upload data to Galaxy
  - 工具：“Get data” —> “Upload File”
  - 数据：sample.fastq, sample2.fastq
  - 注意：
    - 既可以通过链接获取数据，也可以直接上传本地数据（推荐前者）
    - 选择正确的数据格式（提示：Fastq；sample.fastq: Illumina 1.5; sample2.fastq: Illumina 1.5）
    - 因为基因组版本在本实验中无关紧要，所以随便选择一个即可（比如：hg19）
  - 思考：在实际的数据处理中，如何
    - 获取测序数据
    - 拿到数据格式（Fastq的质量编码类型）
    - 选择基因组版本
- ii. Checking read quality with FastQC
  - 工具：“NGS: QC and manipulation” —> “FastQC”
  - 数据：sample.fastq
  - 注意：理解输出报告中每一部分结果的含义
  - 思考：
    - 如何查找FastQC的使用说明？
    - “Basic Statistics”中的Encoding说明什么？
    - 从“Per base sequence quality”中能得到什么信息？
    - 从“Per base sequence content”中能得到什么信息？
- iii. Convert FASTQ quality to sanger
  - 工具：“NGS: QC and manipulation” —> “FASTQ Groomer”
  - 数据：sample.fastq
  - 注意：指定正确的输入数据的质量编码类型（提示：Illumina 1.5）
  - 思考：为什么首先要把Fastq的质量编码转换成Sanger，之后才进行后续的处理？
- iv. Preprocessing with FASTX-Toolkit
  - i. Remove reads with lower quality
    - 工具：“NGS: QC and manipulation” —> “Filter by quality”

- 数据：sample\_sanger.fastq
- 注意：设定正确的参数（要求：keeping only reads that have at least 75% of bases with a quality score of 20 or more）
- 思考：
  - 总的输入、最终输出、丢掉的reads数目各是多少？
  - 在实际的数据处理中，如何选择**合适**的参数？
- ii. Trim the bases with sequence bias from reads
  - 工具：“NGS: QC and manipulation” —> “Trim sequences”
  - 数据：sample\_sanger\_filtered.fastq
  - 注意：设定正确的参数
    - 参考FastQC输出报告中的“Per base sequence content”设定参数“First base to keep”
    - 参考FastQC输出报告中的“Basic Statistics”或者“Sequence Length Distribution”设定参数“Last base to keep”
  - 思考：在实际的数据处理中，如何选择**合适**的参数？
- v. Clean adapter containing reads from FASTQ data
  - i. Checking read quality with FastQC
    - 工具：“NGS: QC and manipulation” —> “FastQC”
    - 数据：sample2.fastq
    - 思考：
      - “Basic Statistics”中的Encoding说明什么？
      - 从“Overrepresented sequences”中能得到什么信息？
  - ii. Convert FASTQ quality to sanger
    - 工具：“NGS: QC and manipulation” —> “FASTQ Groomer”
    - 数据：sample2.fastq
    - 注意：指定正确的输入数据的质量编码类型（提示：Illumina 1.5）
  - iii. Clean adapter containing reads
    - 工具：“NGS: QC and manipulation” —> “Trim Galore!”
    - 数据：sample2\_sanger.fastq
    - 注意：设定正确的参数，要求如下
      - Throw away processed reads shorter than 20 bases
      - The level of error tolerance is adjusted by specifying a maximum 10% error rate
    - 思考：
      - 如何指定adapter的序列？（提示：FastQC输出报告中的“Overrepresented sequences”）
      - 总的输入、带有adapter的reads数目各是多少？
      - 尝试使用“NGS: QC and manipulation” —> “Clip”去除adapter，并比较两种工具的结果。
  - iv. Checking read quality after cleaning adapter
    - 工具：“NGS: QC and manipulation” —> “FastQC”
    - 数据：sample2\_sanger\_trim.fastq
    - 思考：
      - 比较去除adapter前后的FastQC输出报告。
      - 不去除adapter的话对后续的处理有没有影响？
  - vi. 探索“NGS: QC and manipulation”中的其他工具

## 参考资料

---

- [FastQC Help](#)
- [fastqc\\_sweave.pdf](#)
- [QC results](#)

## 备注

---

- 除了Main public site上的[Galaxy](#)以外，还可以尝试使用其他Public servers，比如[Erasmus MC Bioinformatics](#)

---

## 实验二 外显子组测序数据的处理

---

### 实验目的

- i. 掌握外显子组测序数据的分析流程。
- ii. 熟悉BWA、SAMtools、SnpEff等工具的使用方法。
- iii. 熟悉Galaxy的使用方法。
- iv. 了解存储变异信息的VCF格式。

### 实验材料

- i. NA8524\_chr21.fq: human(hg19), fastqsanger

### 实验工具

- i. [Galaxy](#)
- ii. [BWA](#)
- iii. [SAMtools](#)
- iv. [SnpEff](#)

### 实验步骤

- i. Upload data to Galaxy（略；参看实验一）
- ii. Checking read quality with FastQC（略；参看实验一）
- iii. Preprocessing（略；参看实验一）
- iv. Map with BWA
  - 工具：“NGS: Mapping” → “Map with BWA for Illumina”
  - 数据：NA8524\_chr21.fq
  - 注意：指定合适的基因组组装版本
  - 思考：尝试“Map with Bowtie for Illumina”、“Map with BWA”等工具
- v. Statistics with SAMtools
  - i. Convert SAM to BAM
    - 工具：“NGS: SAMtools” → “SAM-to-BAM”
  - ii. Print descriptive information for a BAM dataset
    - 工具：“NGS: SAMtools” → “Flagstat”
    - 思考：尝试“Stats”、“IdxStats”等质控工具
- vi. Call variants
  - i. Call variants with MPileup
    - 工具：“NGS: SAMtools” → “MPileup”
    - 注意：设定正确的参数
      - 选择和先前一致的基因组组装版本
      - 设定参数“Genotype Likelihood Computation”为“Do not perform genotype likelihood computation (output pileup)”
    - 思考：尝试直接使用MPileup提取变异（跳过后面的Varscan）
  - ii. Variant detection with Varscan
    - 工具：“NGS: Variant Analysis” → “Varscan”
    - 思考：尝试调整“Varscan”的参数
- vii. Annotate variants
  - 工具：“NGS: Variant Analysis” → “SnpEff”

- 注意：设定合适的参数
  - 思考：尝试“ANNOVAR Annotate VCF”等其他注释工具
    - 提示：“ANNOVAR Annotate VCF”可能无法正常使用
- viii. Filter variants
- 工具：“NGS: VCF Manipulation” —> “VCFfilter”
  - 注意：根据自己的需要构建表达式
    - 提示：“VCFfilter”可能无法正常使用
  - 思考：尝试使用“Text Manipulation”和“Filter and Sort”中的工具处理VCF
- ix. 补充：实际的数据处理过程中还需要对对比结果（BAM文件）和变异数据（VCF文件）进行以下处理
- Mark/Remove PCR Duplicates
  - Local Realignments Around Indels
  - Quality Recalibration

## 参考资料

---

- [Galaxy Workflow 'Exome Analysis'](#)

## 备注

---

- 除了Main public site上的[Galaxy](#)以外，还可以尝试使用其他Public servers，比如[Erasmus MC Bioinformatics Galaxy Server](#)等。

---

# 实验三 RNA-Seq的数据处理

---

## 实验目的

---

- i. 掌握RNA-Seq测序数据的分析流程。
- ii. 熟悉Tuxedo套件的使用方法。
- iii. 熟悉Galaxy的使用方法。
- iv. 了解存储注释信息的GTF/GFF格式。

## 实验材料

---

- i. [h1-hESC\\_Sample\\_Dataset.fastqsanger](#): human(hg19), fastqsanger
- ii. [GM12878\\_Sample\\_Dataset.fastqsanger](#): human(hg19), fastqsanger
- iii. [UCSC\\_Main\\_on\\_Human\\_refGene\\_chr19\\_BED.bed](#)
- iv. [UCSC\\_Main\\_on\\_Human\\_refGene\\_chr19\\_GTF.gtf](#)

## 实验工具

---

- i. [Galaxy](#)
- ii. [Bowtie](#)
- iii. [TopHat](#)
- iv. [Cufflinks](#)

## 实验步骤

---

- i. Upload data to Galaxy（略；参看实验一）
- ii. Checking read quality with FastQC（略；参看实验一）
  - 思考：可以尝试“NGS: QC and manipulation”中的“FastQC”、“Build base quality distribution”、“Draw quality score boxplot”、“Compute quality statistics”、“FASTQ Summary Statistics”（结合“Graph/Display Data”中

- 的“Boxplot”使用）等工具
- iii. Preprocessing（略；参看实验一）
    - 思考
      - 标准：remove base positions that have a median quality score of below 15
      - Is trimming needed for the datasets?
      - If necessary, trim the reads!（可以尝试“NGS: QC and manipulation”中的“Trim Galore!”、“Trimmomatic”、“Trim sequences”、“FASTQ Trimmer”、“FASTQ Quality Trimmer”等工具）
  - iv. Map reads with TopHat
    - 工具：“NGS: RNA Analysis” → “TopHat”
    - 数据：\*.fastq
    - 注意：指定合适的基因组组装版本
    - 思考
      - 可以尝试利用“UCSC\_Main\_on\_Human\_\_refGene\_chr19\_BED.bed”对TopHat的结果进行可视化
      - 理解TopHat主要参数的含义
      - 理解TopHat每个输出文件的含义
  - v. Assemble and analyze transcripts
    - 工具：“NGS: RNA Analysis” → “Cufflinks”
    - 数据：accepted\_hits.bam
    - 注意：设置合适的参数（此处默认即可）
      - 理解Cufflinks主要参数的含义
      - 理解Cufflinks每个输出文件的含义
  - vi. Identify transcripts that are differentially expressed
    - i. Compare assembled transcripts
      - 工具：“NGS: RNA Analysis” → “Cuffcompare”
      - 数据：(assembled transcripts) X 2 + “UCSC\_Main\_on\_Human\_\_refGene\_chr19\_GTF.gtf”
      - 注意：设置合适的参数
        - 理解Cuffcompare主要参数的含义
        - 理解Cuffcompare每个输出文件的含义
    - ii. Find significant changes in transcript expression
      - 工具：“NGS: RNA Analysis” → “Cuffdiff”
      - 数据：combined transcripts + (accepted\_hits.bam) X 2
      - 注意：设置合适的参数
        - 指定“Condition Name”
        - 理解Cuffdiff主要参数的含义
        - 理解Cuffdiff每个输出文件的含义
      - 思考：分别从“transcript differential expression testing”和“gene differential expression testing”中提取显著差异表达的转录本和基因
  - vii. Visualization with CummeRbund（略）

## 参考资料

---

- [RNA-Seq using Galaxy](#)
- [/training/Glossina\\_annotation/RNA-Seq\\_files](#)

## 备注

---

- 除了Main public site上的[Galaxy](#)以外，还可以尝试使用其他Public servers，比如[Erasmus MC Bioinformatics Galaxy Server](#)等。
-