实验一 测序数据的质量控制与预处理

实验目的

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

实验材料

1. sample.fastq
2. sample2.fastq

实验工具

1. Galaxy
2. FastQC
3. FASTX-Toolkit

实验步骤

1. 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; sample.fastq: Illumina 1.9）
        - 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
2. 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？

3. 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？

4. Preprocessing with FASTX-Toolkit
    1. 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？
    2. 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？

5. Clean adapter containing reads from FASTQ data
    1. 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？
    2. Clean adapter containing reads
        - 工具："NGS: QC and manipulation" ---> "Trim Galore!"
        - 数据：sample2.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。
    3. Checking read quality after cleaning adaper
        - 工具："NGS: QC and manipulation" ---> "FastQC"
        - 数据：sample2_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？
6. 探索"NGS: QC and manipulation"中的其他工具

参考资料

- FastQC Help
- fastqc_sweave.pdf
- QC results

实验二 外显子组测序数据的处理

实验目的

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

实验材料

1. NA8524_chr21.fq: human(hg19), fastqsanger

实验工具

1. Galaxy
2. BWA
3. SAMtools
4. SnpEff

实验步骤

1. Upload data to Galaxy（略；参看实验一）
2. Checking read quality with FastQC（略；参看实验一）
3. Preprocessing（略；参看实验一）
4. 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

5. Statistics with SAMtools
    1. Convert SAM to BAM
        - 工具："NGS: SAMtools" ---> "SAM-to-BAM"
    2. Print descriptive information for a BAM dataset
        - 工具："NGS: SAMtools" ---> "Flagstat"
        - 思考：尝试"Stats"、"IdxStats"等质控工具
6. Call variants
    1. 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）
    2. Variant detection with Varscan
        - 工具："NGS: Variant Analysis" ---> "Varscan"
        - 思考：尝试调整"Varscan"的参数
7. 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
8. 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
9. 补充：实际的数据处理过程中还需要对比对结果（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'

实验三 RNA-Seq的数据处理

实验目的

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

实验材料

1. h1-hESC_Sample_Dataset.fastqsanger: human(hg19), fastqsanger
2. GM12878_Sample_Dataset.fastqsanger: human(hg19), fastqsanger
3. UCSC_Main_on_Human_*refGene_chr19*BED.bed
4. UCSC_Main_on_Human_*refGene_chr19*GTF.gtf

实验工具

1. Galaxy
2. Bowtie
3. TopHat
4. Cufflinks

实验步骤

1. Upload data to Galaxy（略；参看实验一）
2. 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
3. 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）
4. 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

5. 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
6. Identify transcripts that are differentially expressed
    1. 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
    2. 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
7. Visualization with CummeRbund（略）

参考资料

- RNA-Seq using Galaxy
- /training/Glossina_annotation/RNA-Seq_files