

天津医科大学理论课教案首页

(共 4 页、第 1 页)

课程名称：系统生物学 课程内容/章节：基因组学（测序数据分析） / 第 2 章

教师姓名：伊现富 职称：讲师 教学日期：2016 年 9 月 20 日 13:30-15:30

授课对象：生物医学工程与技术学院 2013 级生信班（本） 听课人数：28

授课方式：理论讲授 学时数：2 教材版本：系统生物学，第 1 版

教学目的与要求（分掌握、熟悉、了解、自学四个层次）：

- 掌握 FASTQ、BED、GFF 等数据格式，第二代测序数据分析的基本流程，外显子组测序的分析步骤。
- 熟悉 SAM、VCF 等数据格式，测序深度、覆盖度等术语，测序数据分析的常用工具。
- 了解 SRA、GEO 等数据库，外显子组测序的流程和应用。
- 自学 SRA、GEO 等数据库的使用，测序数据分析常用工具的使用。

授课内容及学时分配：

- (5') 引言与导入：回顾第二代测序的主要技术和基本流程。
- (30') 数据库与数据格式：介绍 SRA 和 GEO 等与第二代测序相关的数据库，讲解第二代测序数据分析中常见的 FASTQ、SAM、BED、GFF 和 VCF 等数据格式。
- (40') 测序数据分析：讲解测序深度和覆盖度等术语，总结测序数据分析的主要流程，介绍分析流程中每个步骤的作用、常用工具和注意事项。
- (20') 外显子组测序：介绍外显子组测序的基本流程，讲解外显子组测序数据分析的基本步骤，通过实例介绍外显子组测序的应用。
- (5') 总结与答疑：总结授课内容中的知识点与技能，解答学生疑问。

教学重点、难点及解决策略：

- 重点：FASTQ、BED、GFF 等数据格式，第二代测序数据的分析流程。
- 难点：SAM、VCF 等数据格式。
- 解决策略：通过实例讲解和比较类比帮助学生理解、记忆。

专业外语词汇或术语：

深度 (depth)	外显子组测序 (WES, whole exome sequencing)
覆盖度 (coverage)	
质量控制 (QC, quality control)	全基因组测序 (WGS, whole genome sequencing)
预处理 (preprocessing)	

辅助教学情况：

- 多媒体：与第二代测序相关的数据库和数据格式。
- 板书：第二代测序数据的分析流程。

复习思考题：

- 举例说明 FASTQ、BED 和 GFF 数据格式。
- 总结测序数据分析的基本流程。
- 解释测序深度和覆盖度。
- 列举测序数据分析中的常用工具。

参考资料：

- 维基百科等网络资源。

主任签字：

年 月 日

教务处制

一、引言与导入 (5 分钟)

1. 测序历史: 第一代 (Sanger) \Rightarrow 第二代 (高通量) \Rightarrow 第三代 (单分子)
2. 二代测序: Roche/454, Illumina/Solexa (桥式扩增 + 边合成边测序), ABI/SOLiD

二、数据库与数据格式 (30 分钟)

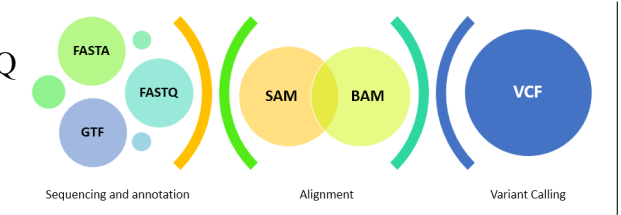
1. 数据库

- 测序数据: SRA (Sequence Read Archive), GEO (Gene Expression Omnibus)
- 肿瘤相关: TCGA (Cancer Genome Atlas), ICGC (International Cancer Consortium)
- 其他: 1000 Genomes Project

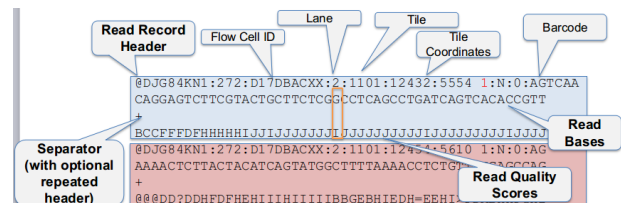
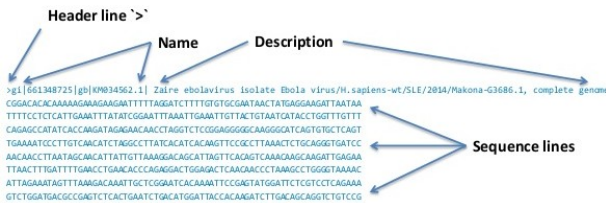
2. 数据格式

(1) 简介

- 序列与读段: FASTA、2bit, FASTQ
- 读段比对: SAM、BAM
- 特征注释: BED、GTF/GFF
- 变异信息: VCF、BCF



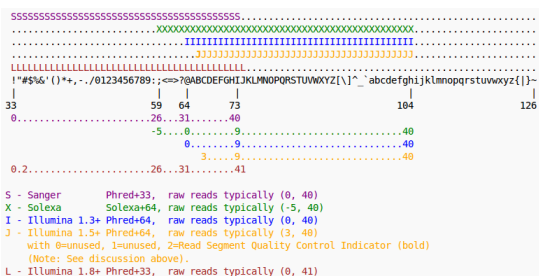
(2) 序列格式: FASTA 和 FASTQ



- FASTQ = FASTA + PHRED (quality score)
- 常用后缀: .fq, .fastq
- ID: /1 和/2 \Rightarrow PE
- PHRED: $Q_{sanger} = -10\log_{10}p$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

\rightarrow 1% error rate



@SQ SN:chr12 LN:133851895
 @RG ID:Sample_ID LB:Sample_Library PL:ILLUMINA SM:Sample_Name PU:Platform_Unit

Read name Flag Chr 5' pos MAPQ Cigar paired 5' pos of the mate Insert size

ERR166338.1 99 chr12 82670685 23 101M = 82670850 266

CCCCCTGGGGATGTTTGCACCAAGCCACTGTCTCCAGCTGG sequence

BBC@GIIHGCFCIEHEAIEIFFGEONDNJFINIONHNGJNNNNKKNJN Base quality

RG:Z:Sample_ID KTA:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 XO:i:0 MD:Z:100 XA:Z tags

Group affiliation

(3) 比对格式: SAM

- SAM: Sequence Alignment/Map format, human readable
- BAM: Binary version of SAM, compress \Rightarrow smaller, index \Rightarrow random access

(4) 注释格式: BED 和 GTF/GFF

- BED \Rightarrow bigBed, bedGraph
- GTF: GFF2.5, Gene Transfer Format
- GFF: GFF3, General Feature Format
- 坐标系统
 - BED: [0-based), length=stop-start
 - GTF/GFF: [1-based], length=stop-start+1

chr1	817371	819837	ENSG00000177757.2_FAM87B_lincRNA	0+
chr1	826206	827522	ENSG00000225880.5_LINC00115_lincRNA	0-
chr1	827608	859446	ENSG00000228794.5_LINC01128_processed_transcript	0+
chr1	868071	876903	ENSG00000230368.2_FAM41C_lincRNA	0-
chr1	873292	874349	ENSG00000234711.1_TUBB8P11_unprocessed_pseudogene	0+
chr1	904834	915976	ENSG00000272438.1_RP11-54O7.16_lincRNA	0+
chr1	911435	914948	ENSG00000230699.2_RP11-54O7.1_lincRNA	0+
chr1	914171	914971	ENSG00000241180.1_RP11-54O7.2_lincRNA	0+
chr1	916865	921016	ENSG00000223764.2_RP11-54O7.3_lincRNA	0-
chr1	924880	944581	ENSG00000187634.7_SAMD11_protein_coding	0+

(共4页、第3页)

(5) 变异格式: VCF

```

#MGTF format
381 Twinscan exon 150 200 . + . gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon 300 401 . + . gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan CDS 380 401 . + 0 gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon 501 650 . + . gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan CDS 501 650 . + 2 gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon 700 800 . + . gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon 700 707 . + 2 gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon 900 1000 . + . gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan start_codon 380 382 . + 0 gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan stop_codon 708 710 . + 0 gene_id "381.000"; transcript_id "381.000.1";

#hgff-version 3
#sequence-region ctg123 1 1497228
ctg123 Prokka gene 1000 9000 . + . ID=gene0001;Name=E2EN
ctg123 Prokka TF_binding_site 1000 1012 . + . ID=fbas0001;Parent=gene0001
ctg123 Prokka mRNA 1050 9000 . + . ID=mRNA0001;Parent=gene0001;Name=E2EN.1
ctg123 Prokka mRNA 1050 9000 . + . ID=mRNA0002;Parent=gene0001;Name=E2EN.2
ctg123 Prokka mRNA 1300 9000 . + . ID=mRNA0003;Parent=gene0001;Name=E2EN.3
ctg123 Prokka exon 1300 1500 . + . ID=exon0001;Parent=mRNA0003
ctg123 Prokka exon 1050 1500 . + . ID=exon0002;Parent=mRNA0001,mRNA0002
ctg123 Prokka exon 3000 3902 . + . ID=exon0003;Parent=mRNA0001,mRNA0003
ctg123 Prokka exon 5000 5500 . + . ID=exon0004;Parent=mRNA0001,mRNA0002,mRNA0003
ctg123 Prokka exon 7000 9000 . + . ID=exon0005;Parent=mRNA0001,mRNA0002,mRNA0003
ctg123 Prokka CDS 1200 1500 . + 0 ID=cds0001;Parent=mRNA0001;Name=denprotein.1
ctg123 Prokka CDS 3000 3902 . + 0 ID=cds0001;Parent=mRNA0001;Name=denprotein.1
ctg123 Prokka CDS 5000 5500 . + 0 ID=cds0001;Parent=mRNA0001;Name=denprotein.1
ctg123 Prokka CDS 7000 7600 . + 0 ID=cds0001;Parent=mRNA0001;Name=denprotein.1
ctg123 Prokka CDS 1201 1500 . + 0 ID=cds0002;Parent=mRNA0002;Name=denprotein.2
ctg123 Prokka CDS 5000 5500 . + 0 ID=cds0002;Parent=mRNA0002;Name=denprotein.2
ctg123 Prokka CDS 7000 7600 . + 0 ID=cds0002;Parent=mRNA0002;Name=denprotein.2
ctg123 Prokka CDS 3301 3902 . + 0 ID=cds0003;Parent=mRNA0003;Name=denprotein.3
ctg123 Prokka CDS 5000 5500 . + 1 ID=cds0003;Parent=mRNA0003;Name=denprotein.3
ctg123 Prokka CDS 7000 7600 . + 1 ID=cds0003;Parent=mRNA0003;Name=denprotein.3

```

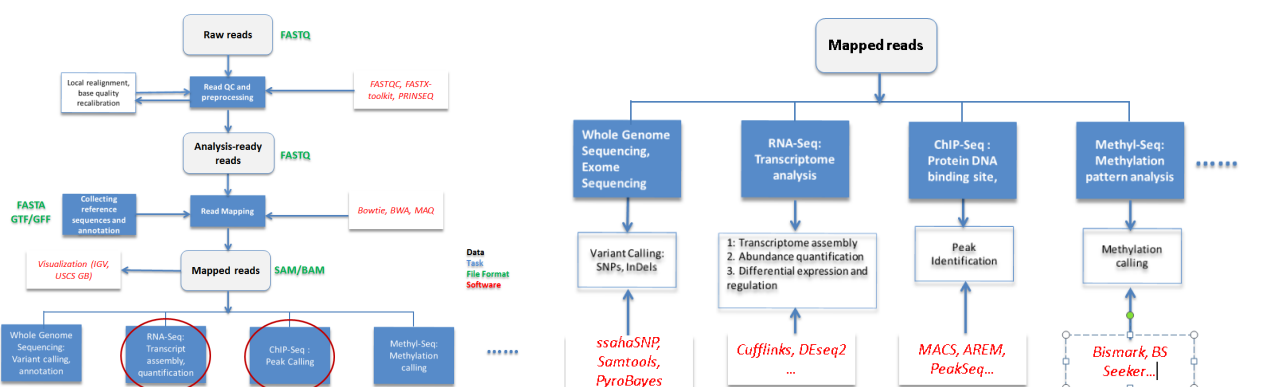
[illegible]

三、测序数据分析 (40 分钟)

1. 常見術語

- 深度: 测序得到的总碱基数与待测基因组大小的比值
- 覆盖度: 测序获得的序列占整个基因组的比例
- SE(single end) vs. PE(paired end)
- PE: insert vs. inner mate distance

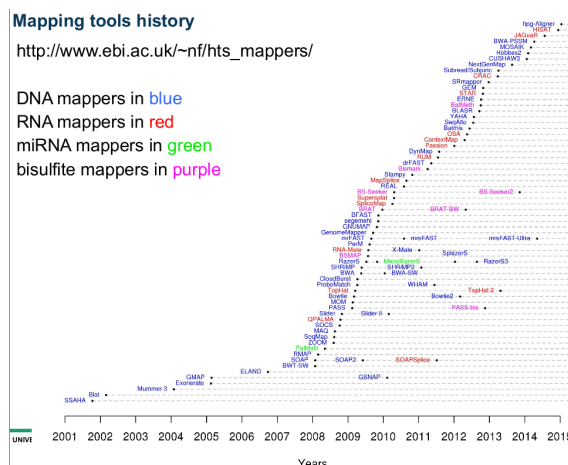
2. 分析流程



- (1) Quality Control: FastQC, NGS QC Toolkit, SolexaQA
- (2) Preprocessing(trim & filter): FASTX-Toolkit, PRINSEQ
- (3) Mapping: BWA, Bowtie, SOAP
- (4) Calling Variants: Samtools, GATK, VarScan
- (5) Variant Annotation: SnpEff, ANNOVAR, SeattleSeq Annotation, SIFT, PolyPhen-2
- (6) Visualization: Genome Browser, IGV, Tablet, Circos
- (7) Others: Galaxy, Picard, bedtools, BEDOPS, csykit

3. 补遗

- Removal of PCR duplicates
- Indel Realignment
- Base quality recalibration
- Ohters: Replicates(biological vs. technical), depth, length, SE vs. PE



四、 外显子组测序 (20 分钟)

1. 基本概念

- exome: genome \Rightarrow 1%, 30Mb
- WES: exome \Rightarrow sequencing
- WGS: genome \Rightarrow sequencing

2. 【重点】 流程：实验 + 分析

3. 应用实例

五、 总结与答疑 (5 分钟)

1. 知识点

- 数据库: SRA、GEO
- 数据格式: FASTQ、BED、GTF/GFF, SAM、VCF
- 测序术语: 深度、覆盖度、PE
- 分析流程: QC、preprocessing, mapping、variants、annotation, visualization
- 外显子组测序: 实验与分析流程

2. 技能

- 测序数据分析软件: 使用方法
- 全基因组测序: 数据分析
- 外显子组测序: 数据分析

