

天津医科大学实验课教案首页

(共 2 页、第 1 页)

课程名称：系统生物学	实验名称：实验二 外显子组测序数据的处理
教师姓名：伊现富	职称：讲师
授课对象：生物医学工程与技术学院 2014 级生信班（本）	实验人数：30
实验类型（验证型、综合型、设计型、创新型）：验证型	实验分组：一人一机
学时数：3	教材版本：系统生物学实验讲义（自编教材）

实验目的与要求：

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

实验内容及学时分配：

- (10') 分析流程：回顾外显子组测序数据分析的基本流程。
- (10') 常用工具：回顾总结外显子组测序数据分析中的常用工具。
- (10') VCF 格式：回顾存储变异信息的 VCF 格式。
- (120') 实验操作：从单端测序的外显子组测序数据中提取变异并进行注释。

主要仪器和实验材料：

- 实验材料：以 FASTQ 格式存储的单端外显子组测序数据。
- 主要仪器：联网的计算机。
- 分析工具：Galaxy, BWA, SAMtools, SnpEff。

实验重点、难点及解决策略：

- 难点：VCF 格式；解决策略：通过实例进行讲解。
- 重点：BWA、SAMtools 和 SnpEff 的使用；解决策略：根据资料进行学习，通过练习熟练掌握。

思考题：

- 总结外显子组测序数据的分析流程。
- 列举外显子组测序数据分析中的常用工具。
- 解释存储变异信息的 VCF 格式。

参考资料：

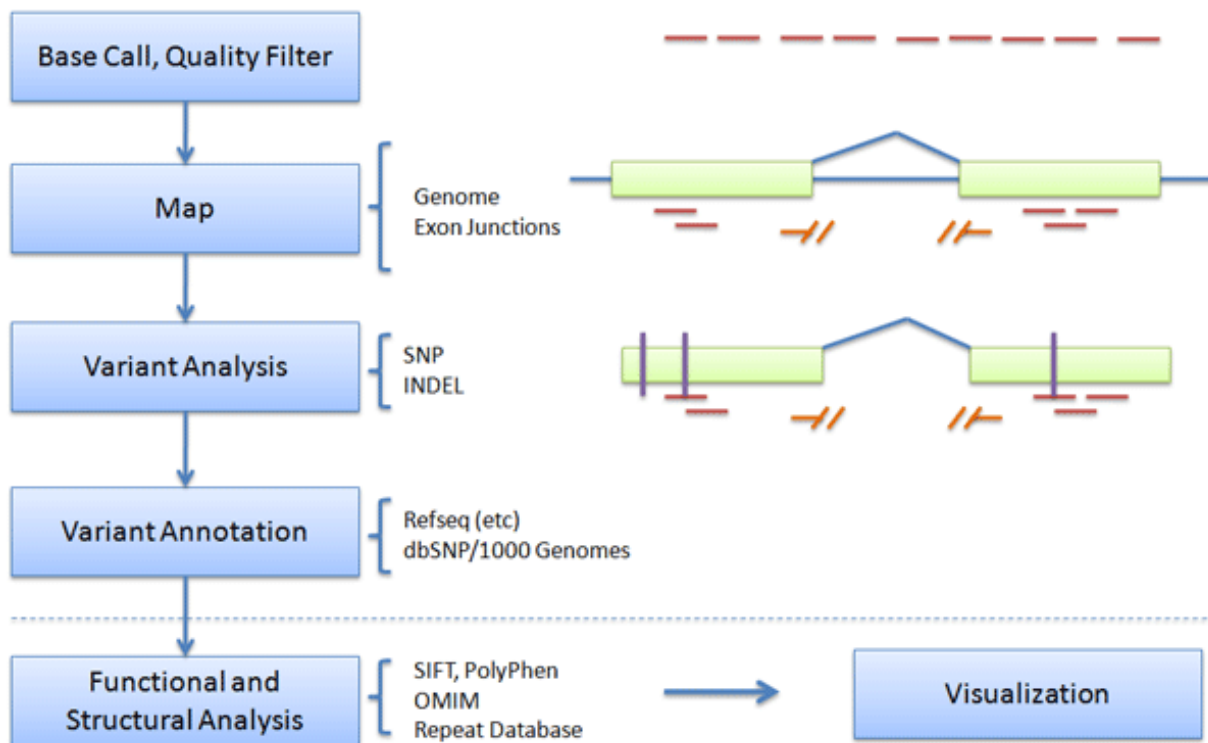
- BWA
- SAMtools
- SnpEff
- Galaxy

主任签字：

年 月 日

教务处制

一、分析流程 (10 分钟)



二、常用工具 (10 分钟)

- BWA: mapping reads against a reference genome
- Bowtie: an ultrafast, memory-efficient short read aligner
- SAMtools: interacting with high-throughput sequencing data
- VarScan: detect variants in NGS data
- SnpEff: genetic variant annotation and effect prediction toolbox
- ANNOVAR: functionally annotate genetic variants

VCF header

```

##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">

```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0/1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1/0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Annotations:

- Mandatory header lines:** ##fileformat=VCFv4.0
- Optional header lines (meta-data about the annotations in the VCF body):** ##INFO, ##FORMAT, ##ALT, ##INFO
- Reference alleles (GT=0):** A, AT, T, CT, G, C
- Alternate alleles (GT>0 is an index to the ALT column):** T, CT, G, C,
- Phased data (G and C above are on the same chromosome):** 1/1:12:3
- Deletion:**
- SNP:** rs1
- Large SV:** SVTYPE=DEL;END=300
- Insertion:** T, CT
- Other event:** H2;AA=T

三、VCF 格式 (10 分钟)

四、实验操作 (120 分钟)

1. Upload data to Galaxy (比较导入数据的不同方法; 注意参数的设定)
2. Checking read quality with FastQC; Preprocessing (参照实验一)
3. Map with BWA (可以尝试一下其他类似工具)
4. Statistics with SAMtools (可以尝试一下其他类似工具)
5. Call variants (比较 MPileup 和 Varscan 的结果)
6. Annotate variants (尝试不同的工具并比较结果)
7. Filter variants (摸索阈值, 尝试工具)