

# 天津医科大学理论课教案首页

(共 4 页、第 1 页)

课程名称：生物信息学 课程内容/章节：第五章 (5.3) 基因组功能的高级注释

教师姓名：伊现富 职称：讲师 教学日期：2016 年 10 月 10 日 10:00-12:00

授课对象：生物医学工程与技术学院 2014 级生信班 (本) 听课人数：30

授课方式：理论讲授 学时数：2 教材版本：生物信息学：基础及应用

教学目的与要求 (分掌握、熟悉、了解、自学四个层次)：

- 掌握序列标识的含义和制作工具。
- 熟悉变异位点注释结果的解析；基因集的富集分析及其结果解析；box plot 的含义及其绘制。
- 了解变异位点注释的内容和常用工具；基因集富集分析的常用工具。
- 自学变异位点注释、基因集富集分析、序列标识制作等工具的使用方法。

授课内容及学时分配：

- (5') 回顾与导入：回顾基因组注释的基础知识，介绍功能注释的主要内容。
- (25') 变异位点的注释：介绍变异位点注释的内容、步骤及相关的注释工具，讲解对注释结果的解析。
- (20') 基因集富集分析：介绍基因集富集分析的用途，讲解常用的 DAVID 工具及其结果的解析。
- (25') 序列标识：讲解序列标识的含义，介绍常用的 WebLogo 及其使用方法并讲解对结果的解析。
- (20') box plot：介绍 box plot 及相关概念，讲解绘制 box plot 的主要步骤。
- (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：序列标识的含义；解决策略：通过制作过程的演示和对结果的解读来加深学生的理解。
- 难点：注释分析结果的解析；解决策略：通过对实例的分析帮助学生掌握解析结果的基本原则和主要步骤。

专业外语词汇或术语：

单核苷酸变异 (SNV)

基因集 (gene set)

GO (gene ontology)

富集分析 (enrichment analysis)

序列标识 (sequence logo)

箱线图 (box plot)

辅助教学情况：

- 多媒体：变异位点注释、基因集富集分析的实例；序列标识和 box plot 的示意图；DAVID、WebLogo 等工具的界面；绘制 box plot 的演示视频。
- 板书：box plot 的主要绘制步骤。
- 操作演示：序列标识的制作。

复习思考题：

- 以变异位点的注释结果为例，论述如何解析一张表。
- 以 DAVID 富集分析结果为例，论述如何解析一张表。
- 简述序列标识的含义，能解读实际的序列标识图。
- 以 box plot 为例，论述如何解析一张图。

参考资料：

- 李霞，李亦学，廖飞。生物信息学，人民卫生出版社，2010 年。
- 朱玉贤，李毅，郑晓峰。现代分子生物学 (第 3 版)，高等教育出版社，2007。
- 维基百科

主任签字：

年 月 日

教务处制

## 一、回顾与导入 (5 分钟)

回顾基因组注释的基础知识, 介绍高级注释的内容, 强调基础知识在高级注释中无处不在。

## 1. 基因组注释的基础知识

- 基因组的组装版本: hg19 与 GRCh37, mm10 与 GRCm38
- 两种坐标系统: 0-based, 1-based
- 四种常用格式: FASTA, BED, GFF, VCF
- 逻辑运算模式: intersect, subtract, join, ...

## 2. 基因组功能的高级注释

- 变异位点的注释: SNVs、非同义多态性的注释
- 基因集富集分析: GO, KEGG, DAVID
- 序列标识: WebLogo

Pos	Alt SNP	Ref SNP	Ref SNP Codon	Alt SNP Codon	Ref SNP AA	Alt SNP AA	Anno Type
3	G	A	---	---	---	---	5'UTR
5	A	C	CAT	CCT	His	Pro	Non_Synonymous
13	G	T	CCG	CCT	Pro	Pro	Synonymous
25	C	A	TAC	TAA	Tyr	Stop	Stop Loss
43	A	C	---	---	---	---	Splice Site
44	G	T	---	---	---	---	Intronic
50	C	A	---	---	---	---	Essential Splice Site
53	A	C	ACC	CCC	Thr	Pro	Non_Synonymous
66	C	T	---	---	---	---	3'UTR
72	A	C	---	---	---	---	Downstream

## 二、变异位点的注释 (25 分钟)

重点讲解对注释结果的解析及其在功能注释流程中承上启下的作用。

## 1. 单核苷酸变异的注释

- 注释内容: 附加相关的基因组注释信息 (数据库 ID, 基因名, 变异功能类别, ...)
- 注释工具: SeattleSeq Annotation, variant tools, SnpEff
- 【难点】结果解析: SeattleSeq Annotation 的注释结果 (通过实例解读注释结果; 对注释结果过滤筛选后可继续进行非同义多态性的注释)

File:

data://pos-as-7.1.Final/gvBathOutput/SeattleSeqAnnotation137.1individual.294000400650.txt

Title:

Individual

Counts:

HapMapRefType HapMapRefInlinor

polyRefType polyRefInScore

Count missense SNPs = 8

Count stop SNPs = 0

Count SNPs in splice sites = 0

Count SNPs in coding synonymous = 8

Count SNPs in coding (not mod 3) = 0

Count SNPs in a UTR = 0

Count SNPs near a gene = 0

Count SNPs in introns = 0

Count intergenic SNPs = 0

number SNPs in microRNAs = 0

number accessions coding-synonymous NCBI = 19

number accessions missense NCBI = 15

number accessions stop NCBI = 0

number accessions splice-site NCBI = 0

number SNPs in dbSNP = 16

number SNPs not in dbSNP = 0

number SNPs total = 16

Add or Remove Columns:

☒ Sample Alleles
 ☒ Alleles in dbSNP
 ☒ GVS Function
 ☒ dbSNP Function
 ☒ Chimp Allele
 ☒ Copy Number Variations
 ☒ HapMap Rare-Allele Frequencies
 ☒ dbSNP Validation
 ☒ RepeatMasker
 ☒ Translational Reads
 ☒ microRNAs
 ☒ Grantham Score
 ☒ cDNA Position
 ☒ PolyPhen Prediction
 ☒ Clinical Association
 ☒ Distance to Nearest Splice Site
 ☒ NHLBI ESP Allele Counts

Sort by Column Value:

☒ Original Order
 ☐ dbSNP Function
 ☐ GVS Function
 ☐ Conservation Score phastCons
 ☐ Conservation Score GERP
 ☐ In dbSNP

Sort Direction:

☒ Forward
 ☐ Reverse

Filter:

☐ Only missense, nonsense, splice, frameshift (GVS)
 ☐ Only synonymous SNPs or coding (not frameshift) indels (GVS)
 ☐ Only intron (GVS)
 ☐ Only variations not in dbSNP
 ☐ Only variations with clinical association

Table

reset

16 SNP locations 38 accession lines page 1 of 1

dbSNP ID	chr	pos	ref	alt	sample	sampleGenotype	sampleAlleles	alleles/DBSNP	accession	functionGVS	functionDBSNP	rsID	aminoAcids	proteinPos
dbSNP_130	10	1126383	A	R	AVG	AVG	AVG	NM_014023.3	coding-synonymous	synonymous-codon	73578538	none	121495	
dbSNP_86	10	3150973	C	Y	C/T	C/T	C/T	NM_001242339.1	coding-synonymous	synonymous-codon	1132172	none	309777	
dbSNP_86	10	3150973	C	Y	C/T	C/T	C/T	NM_002627.4	coding-synonymous	synonymous-codon	1132173	none	317785	

## 2. 非同义多态性的注释

- 注释内容: 对蛋白质产物结构和功能的影响
- 注释工具: SIFT, PolyPhen-2, SNPs3D, PROVEAN
- 【难点】结果解析: SIFT 的注释结果 (通过实例解读注释结果; 承接 SNVs 的注释, 对结果过滤筛选后可继续进行基因集的富集分析)

Transcript ID	Protein ID	Substitution	Region	dbSNP ID	SNP Type	Prediction	SIFT Score
ENST00000294724	ENSP00000294724	R1487G	EXON CDS	rs12118058:G	Nonsynonymous	TOLERATED	0.46
ENST00000294724	ENSP00000294724	E1405G	EXON CDS	rs28730708:G	Nonsynonymous	DAMAGING	0.01
ENST00000294724	ENSP00000294724	R1487R	EXON CDS	rs12118058:G	Synonymous	TOLERATED	0.64
ENST00000330029	ENSP00000332887	E49A	EXON CDS	novel	Nonsynonymous	DAMAGING	0.02
ENST00000371564	ENSP00000360619	T612N	EXON CDS	rs6067785:T	Nonsynonymous	DAMAGING	0
ENST00000283943	ENSP00000283943	Q1910*	EXON CDS	rs1803846:A	Nonsynonymous	N/A	N/A
ENST00000341772	ENSP00000345229	P433L	EXON CDS	rs17853365:A	Nonsynonymous	DAMAGING	0.02

## 三、基因集富集分析 (20 分钟)

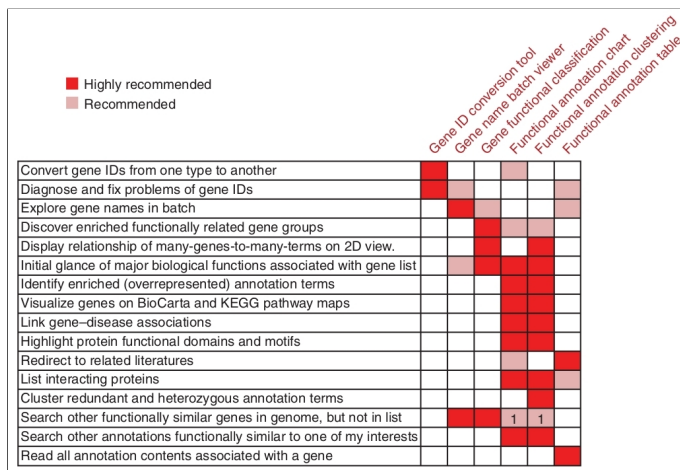
## 1. 基因集富集分析 (承接变异位点的注释)

- 富集分析: 基因集, GO, KEGG
- GO (Gene Ontology)
  - biological process
  - molecular function
  - cellular component
- 结果解析 (解析使用 DAVID 进行 GO 富集分析的结果)
  - 富集显著性
  - 多重检验校正

Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_CC_FAT	extracellular region	RT		40	25.8	6.9E-6	1.5E-3
GOTERM_CC_FAT	extracellular region part	RT		24	15.5	3.8E-5	4.0E-3
GOTERM_MF_FAT	oxygen binding	RT		6	3.9	3.8E-5	1.4E-2
GOTERM_CC_FAT	extracellular space	RT		19	12.3	9.4E-5	6.5E-3
GOTERM_MF_FAT	heme binding	RT		8	5.2	1.0E-4	1.9E-2
GOTERM_BP_FAT	defense response	RT		18	11.6	1.3E-4	1.7E-1
GOTERM_BP_FAT	response to bacterium	RT		10	6.5	1.4E-4	9.1E-2
GOTERM_MF_FAT	tetrapyrrole binding	RT		8	5.2	1.5E-4	1.9E-2
GOTERM_MF_FAT	iron ion binding	RT		11	7.1	4.3E-4	3.9E-2
GOTERM_BP_FAT	defense response to bacterium	RT		7	4.5	8.9E-4	3.4E-1
GOTERM_BP_FAT	response to drug	RT		9	5.8	1.5E-3	4.0E-1
GOTERM_BP_FAT	regulation of response to external stimulus	RT		7	4.5	5.2E-3	7.7E-1
GOTERM_BP_FAT	taxis	RT		7	4.5	5.4E-3	7.2E-1
GOTERM_BP_FAT	chemotaxis	RT		7	4.5	5.4E-3	7.2E-1
GOTERM_CC_FAT	hemoglobin complex	RT		3	1.9	5.7E-3	2.6E-1
GOTERM_MF_FAT	oxygen transporter activity	RT		3	1.9	5.8E-3	3.5E-1

## 2. DAVID 分析工具 (根据任务选择工具)

- Gene Name Batch Viewer
- Gene ID Conversion Tool
- Gene Functional Classification Tool
- Functional Annotation Tool
  - Functional Annotation Clustering: 根据注释信息聚类注释项目
  - Functional Annotation Chart: 根据注释信息进行富集分析
  - Functional Annotation Table: 以表格形式呈现注释信息



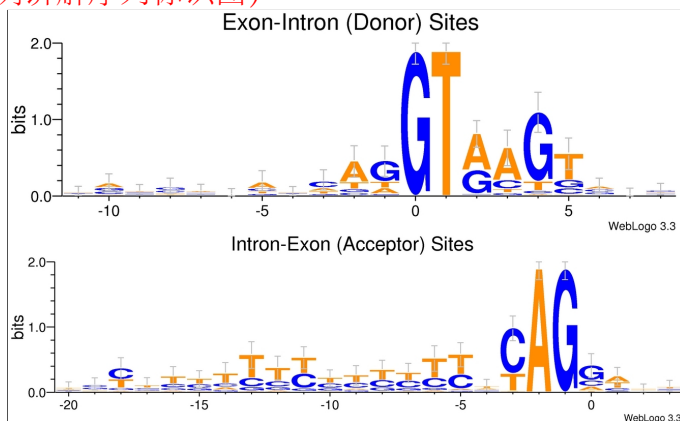
## 四、序列标识 (25 分钟)

### 1. 【重点】图形含义 (以“GT-AG 规则”为例讲解序列标识图)

- 数据：多序列比对信息
- 横轴：序列的坐标位置
- 纵轴：比特，计量单位
- 字符堆叠的总高度：此位置的保守性
- 每个字符的高度：出现的相对频率

### 2. 制作工具 (演示 WebLogo 的使用)

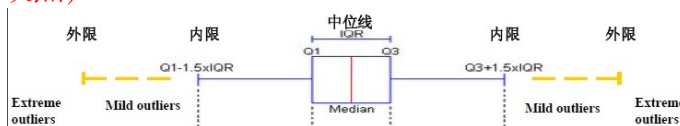
- WebLogo
- enoLOGOS
- Skylogn



## 五、box plot (20 分钟)

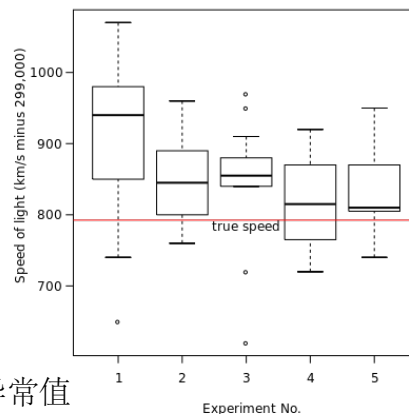
### 1. box plot 简介 (通过实例和示意图讲解其优缺点)

- box plot, Box-whisker Plot, 箱线图
- 1977, 美国, 约翰·图基 (John Tukey)
- 显示一组数据分散情况的统计图
- 可以粗略看出数据分布的离散程度
- 适合用于几个样本的比较
- 不能提供数据分布偏态的精确度量



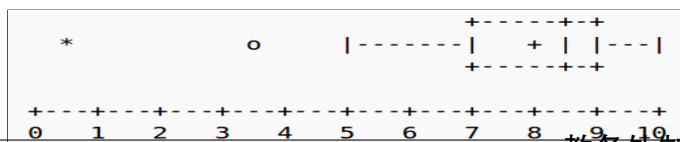
### 2. 相关概念 (通过实例帮助学生理解记忆)

- 最小值 min, 最大值 max, 中位数 median
- 下四分位数 Q1, 上四分位数 Q3
- 四分位数差  $IQR = Q3 - Q1$
- 内限:  $Q3 + 1.5IQR$ ,  $Q1 - 1.5IQR$
- 外限:  $Q3 + 3IQR$ ,  $Q1 - 3IQR$
- 异常值 (outliers): 处于内限以外的数据
- 温和的异常值 (mild outliers): 在内限与外限之间的异常值
- 极端的异常值 (extreme outliers): 在外限以外的异常值



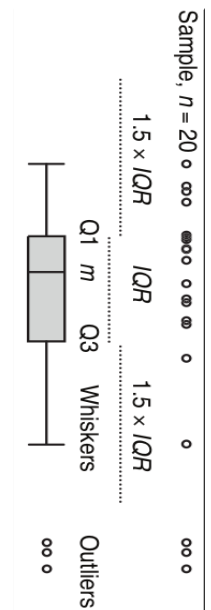
### 3. 图解概念

- $min = 0.5$ ,  $max = 10$
- $Q1 = 7$ ,  $Q3 = 9$ ,  $IQR = 2$
- $median = 8.5$ ,  $mean = 8$



## 4. 绘图步骤 (通过观看视频学习绘图的具体步骤)

- 绘制数轴。
- 计算上四分位数 ( $Q3$ )，中位数，下四分位数 ( $Q1$ )。
- 计算四分位数差 ( $IQR$ )。
- 绘制箱线图的矩形，上限为  $Q3$ ，下限为  $Q1$ 。在矩形内部中位数的位置画一条横线 (中位线)。
- 在  $Q3 + 1.5IQR$  和  $Q1 - 1.5IQR$  处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在  $Q3 + 3IQR$  和  $Q1 - 3IQR$  处画两条线段，称为外限。
- 在非异常值的数据中，最靠近上边缘和下边缘 (即内限) 的两个数值处画横线，作为箱线图的触须。
- 从矩形的两端向外各画一条线段直到不是异常值的最远点 (即上一步的触须)，表示该批数据正常值的分布区间。
- 温和的异常值用空心圆表示；极端的异常值用实心点 (一说用星号  $*$ ) 表示。



## 5. 绘图工具: BoxPlotR, ECplot, R, ...

## 六、总结与答疑 (5 分钟)

### 1. 知识点

- 变异位点的注释: 用途, 注释工具
- 基因集富集分析: 功能, 分析工具
- 序列标识: 含义, 制作工具
- box plot: 理解, 绘制

### 2. 技能

- 解析表格: 行列, 缩写, 数值
- 解析图片: 数据, 横纵轴, 图元素, 元素大小、颜色