生物信息学

天津医科大学
生物医学工程与技术学院

2016-2017 学年下学期（春）
14 级基础班

第五章 基因组功能注释分析

伊现富（Yi Xianfu）

天津医科大学（TIJMU）
生物医学工程与技术学院

2017 年 5 月

章节内容概览

5.1,5.2 基因组功能注释分析基础

1. 基础知识：组装版本，坐标系统，常用格式，逻辑运算模式
2. 准备工作：坐标转换，格式转换，逻辑运算
3. 扩展知识：文本文件与文本编辑器

5.3 基因组功能的高级注释

1. 高级注释：变异位点注释，富集分析，序列标识
2. 扩展知识：box plot，解析图表

5.3 Galaxy 分析平台

1. Galaxy 分析平台：简介，使用
2. 扩展知识：数据处理三段论

教学提纲

1. 引言
2. 基因组组装版本
3. 基因组坐标系统
4. 基因组注释常用格式
5. 文本文件与文本编辑器
6. 基因组坐标的逻辑运算
7. 总结与答疑
8. 引言
9. 变异位点的注释
10. 基因集富集分析
11. 序列标识
12. box plot
13. 解析图表
14. 总结与答疑
15. 引言
16. Galaxy 分析平台
17. Galaxy 使用演示
18. 数据处理三段论
19. 总结与答疑
20. 复习思考题

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

基因组注释（genome annotation）

从原始的基因组核酸序列中挖掘有用的生物学信息并阐释其生物学含义，包括基因组结构注释和基因组功能注释两大部分。

基因组结构注释（structural annotation）

在基因组序列中寻找基因等功能元件并明确其基本结构。

基因组功能注释（functional annotation）

在结构注释的基础上，将进化保守性（evolutionary conservation）和基因本体论（gene ontology）等元数据（meta-data）与功能元件对应起来，找到其生物学功能。

基因组注释（genome annotation）

从原始的基因组核酸序列中挖掘有用的生物学信息并阐释其生物学含义，包括基因组结构注释和基因组功能注释两大部分。

基因组结构注释（structural annotation）

在基因组序列中寻找基因等功能元件并明确其基本结构。

基因组功能注释（functional annotation）

在结构注释的基础上，将进化保守性（evolutionary conservation）和基因本体论（gene ontology）等元数据（meta-data）与功能元件对应起来，找到其生物学功能。

基因组注释（genome annotation）

从原始的基因组核酸序列中挖掘有用的生物学信息并阐释其生物学含义，包括基因组结构注释和基因组功能注释两大部分。

基因组结构注释（structural annotation）

在基因组序列中寻找基因等功能元件并明确其基本结构。

基因组功能注释（functional annotation）

在结构注释的基础上，将进化保守性（evolutionary conservation）和基因本体论（gene ontology）等元数据（meta-data）与功能元件对应起来，找到其生物学功能。

基因组注释

- 结构注释 ⇐ 实验手段，单个基因
  - 限制性酶切位点分析、开放阅读框分析、启动子分析、CpG 岛识别
  - 重复序列分析、基因识别
  - mRNA 选择性剪接分析
- 功能注释 ⇐ 组学时代，复杂疾病
  - 变异位点的注释
  - 基因集富集分析
  - 生物学通路分析
  - 相互作用网络分析
  - 分子进化分析

引言 | 基础知识

- 基因组组装版本
- 基因组坐标系统
- 注释常用格式
- 文本编辑器
- 坐标的逻辑运算

Sequence to Variation Workflow
Raw Data
FastQ
FastX Tookit
FastQ
BWA/ BWASW
BWA/ BWASW
SAM
SAM Tools
BAM
Samtools mPileup
Pileup
BCF Tools
Raw VCF
BCF Tools
Filter VCF
GFF
IGV

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

组装版本 | 疑惑

- These sequences were mapped to human and mouse genomes sequences (hg18 and mm9, respectively) using BLASTN.
- We used DNA sequences from the human and mouse genome assemblies hg18 and mm9.
- Currently there are 25,000 genes annotated in the human (hg18) and mouse (mm9) genome, which comprise less than 3% of the genome (UCSC genome browser; http://genome.ucsc.edu/).
- The GRCh37/hg19 and GRCm38/mm10 assemblies at the UCSC genome browser (http://genome.ucsc.edu/) were used for mapping the chromosomal defect and gene annotations.
- The genome assemblies from which the sequences obtained were Dec 2011 (GRCm38/mm10), Feb 2009 (GRCh37/hg19) and Nov 2004 (Baylor3.4/rn4) for mouse, human and rat respectively.

组装版本 | XP vs. Win7

基因组序列不是确定的吗？也需要版本升级？

SPECIES	UCSC	DATE	NCBI
Human	hg19	Feb. 2009	Genome Reference Consortium GRCh37
	hg18	Mar. 2006	NCBI Build 36.1
	hg17	May 2004	NCBI Build 35
	hg16	Jul. 2003	NCBI Build 34
Mouse	mm10	Dec. 2011	Genome Reference Consortium GRCm38
	mm9	Jul. 2007	NCBI Build 37
	mm8	Feb. 2006	NCBI Build 36
	mm7	Aug. 2005	NCBI Build 35

human: *Homo sapiens*; mouse: *Mus musculus*
hg: human genome; GRC: Genome Reference Consortium

基因组序列不是确定的吗？也需要版本升级？

SPECIES	UCSC	DATE	NCBI
Human	hg19	Feb. 2009	Genome Reference Consortium
			GRCh37
	hg18	Mar. 2006	NCBI Build 36.1
	hg17	May 2004	NCBI Build 35
	hg16	Jul. 2003	NCBI Build 34
Mouse	mm10	Dec. 2011	Genome Reference Consortium
			GRCm38
	mm9	Jul. 2007	NCBI Build 37
	mm8	Feb. 2006	NCBI Build 36
	mm7	Aug. 2005	NCBI Build 35

human: *Homo sapiens*; mouse: *Mus musculus*
hg: human genome; GRC: Genome Reference Consortium

基因组序列不是确定的吗？也需要版本升级？

SPECIES	UCSC	DATE	NCBI
Human	hg19	Feb. 2009	Genome Reference Consortium
			GRCh37
	hg18	Mar. 2006	NCBI Build 36.1
	hg17	May 2004	NCBI Build 35
	hg16	Jul. 2003	NCBI Build 34
Mouse	mm10	Dec. 2011	Genome Reference Consortium
			GRCm38
	mm9	Jul. 2007	NCBI Build 37
	mm8	Feb. 2006	NCBI Build 36
	mm7	Aug. 2005	NCBI Build 35

human: *Homo sapiens*; mouse: *Mus musculus*

hg: human genome; GRC: Genome Reference Consortium

- UCSC liftOver tool：支持 BED 和“chrN:start-end”格式的输入
- Galaxy（基于 UCSC liftOver tool）：支持 BED、GFF 和 GTF 格式的输入
- CrossMap：支持 SAM/BAM、Wiggle/BigWig、BED、GFF/GTF 和 VCF 格式的输入，输出对应格式
- NCBI Remap：支持 BED、GFF、GTF 和 VCF 等格式的输入
- Ensembl assembly converter（2015 年退休，CrossMap 继位）：支持 BED、GFF、GFT 和 PSL 格式的输入，但输出都是 GFF 格式的
- pyliftover：仅支持点坐标（point coordinates）的转换，无法对区段（ranges）坐标进行转换

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

坐标系统 | 坐标轴

坐标系统 | 坐标轴

y-axis
x-axis
II
I
III
IV
P(3, 5)
(0, 0)
origin
10
5
−5
−10
−10
−5
5
10

mm
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

坐标系统 | 坐标轴

mm
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Start
Stop
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

坐标系统 | 坐标轴

hg19

- SNP, rs1800468: ”chr19 41860587”; ”chr19:41860587”
- gene, *SAMD11*: ”chr1 861121 879961”; ”chr1:861121-879961”

坐标系统 | 两大系统

序列

0-based index	0	1	2	3	4	5	6	7
Sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

TG 的坐标

- 0-based，half-open：[3,5)
- 1-based，fully-closed：[4,5]

实例

- 0-based：BED、BAM、PSL，dbSNP、Table Browser
- 1-based：GFF、VCF、SAM、Wiggle，DAS、Genome Browser

坐标系统 | 两大系统

序列

0-based index	0	1	2	3	4	5	6	7
Sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

TG 的坐标

- 0-based，half-open：[3,5)
- 1-based，fully-closed：[4,5]

实例

- 0-based：BED、BAM、PSL，dbSNP、Table Browser
- 1-based：GFF、VCF、SAM、Wiggle，DAS、Genome Browser

坐标系统 | 两大系统

序列

0-based index	0	1	2	3	4	5	6	7
Sequence	A	A	T	T	G	G	C	C
1-based index	1	2	3	4	5	6	7	8

TG 的坐标

- 0-based，half-open：[3,5)
- 1-based，fully-closed：[4,5]

实例

- 0-based：BED、BAM、PSL，dbSNP、Table Browser
- 1-based：GFF、VCF、SAM、Wiggle，DAS、Genome Browser

坐标系统 | 类比

first floor

坐标系统 | 类比

first floor

数组

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

格式 | 文件格式

格式 | FASTA

```
>gi|183121|gb|M29645.1|HUMGFIII Human insulin-like growth factor II mRNA, complete cds
CAGGGGCCGAAGAGTCACCACCGAGCTTGTGTGGGAGGAGGTGGATTCCAGCCCCCAGCCCCAGGGCTCT
GAATCGCTGCCAGCTCAGCCCCCTGCCCAGCCTGCCCCACAGCCTGAGCCCCAGCAGGCCAGAGAGCCCA
GTCCTGAGGTGAGCTGCTGTGGCCTGTGGCCAGGCGACCCCAGCGCTCCCAGAACTGAGGCTGGCAGCCA
GCCCCAGCCTCAGCCCCAACTGCGAGGCAGAGAGACACCAATGGGAATGCCAATGGGGAAGTCGATGCTG
GTGCTTCTCACCTTCTTGGCCTTCGCCTCGTGCTGCATTGCTGCTTACCGCCCCAGTGAGACCCTGTGCG
GCGGGGAGCTGGTGGACACCCTCCAGTTCGTCTGTGGGGACCGCGGCTTCTACTTCAGCAGGCCCGCAAG
CCGTGTGAGCCGTCGCAGCCGTGGCATCGTTGAGGAGTGCTGTTTCCGCAGCTGTGACCTGGCCCTCCTG
GAGACGTACTGTGCTACCCCCGCCAAGTCCGAGAGGGACGTGTCGACCCCTCCGACCGTGCTTCCGGACA
ACTTCCCCAGATACCCCGTGGGCAAGTTCTTCCAATATGACACCTGGAAGCAGTCCACCCAGCGCCTGCG
CAGGGGCCTGCCTGCCCTCCTGCGTGCCCGCCGGGGTCACGTGCTCGCCAAGGAGCTCGAGGCGTTCAGG
GAGGCCAAACGTCACCGTCCCCTGATTGCTCTACCCACCCAAGACCCCGCCCACGGGGGCGCCCCCCCAG
AGATGGCCAGCAATCGGAAGTGAGCAAAACTGCCGCAAGTCTGCAGCCCGGCGCCACCATCCTGCAGCCT
CCTCCTGACCACGGACGTTTCCATCAGGTTCCATCCCGAAATCTCTCGGTTCCACGTCCCCCTGGGGCTT
CTCCTGACCCAGTCCCCGTGCCCCGCCTCCCCGAAACAGGCTACTCTCCTCGGCCCCCTCCATCGGGCTG
AGGAAGCACAGCAGCATCTTCAAACATGTACAAAATCGATTGGCTTTAAACACCTTCACATACCT
```

格式 | FASTA | 注意事项

- 每一行最好不要超过 80 个字符
- 序列中的换行符不会影响序列的连续性
- 使用标准的 IUB/IUPAC 核酸代码和氨基酸代码
- 允许小写字母的存在，但会转换成大写
- 单个“-”代表不明长度的空位
- 在氨基酸序列中允许出现“U”和“*”
- 任何数字都应该被去掉或转换成字母
- 不明核酸和氨基酸分别用“N”和“X”表示

格式 | FASTA | IUB/IUPAC 核酸

格式 | FASTA | IUB/IUPAC 核酸

Code	Meaning	Code	Meaning
A	Adenine	Y	Pyrimidine (C, T, or U)
C	Cytosine	K	T, U, or G (keto)
G	Guanine	W	T, U, or A (weak)
T	Thymine	B	C, T, U, or G (not A)
U	Uracil	D	A, T, U, or G (not C)
R	Purine (A or G)	H	A, T, U, or C (not G)
S	C or G (strong)	V	A, C, or G (not T, not U)
M	C or A (amino)	N	Any base (A, C, G, T, or U)
X	masked	-	gap of indeterminate length

格式 | FASTA | IUB/IUPAC 氨基酸

1	3	Meaning	1	3	Meaning
A	Ala	Alanine	B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine	D	Asp	Aspartic acid
E	Glu	Glutamic acid	F	Phe	Phenylalanine
G	Gly	Glycine	H	His	Histidin
I	Ile	Isoleucine	K	Lys	Lysine
L	Leu	Leucine	M	Met	Methionine
N	Asn	Asparagine	P	Pro	Proline
Q	Gln	Glutamine	R	Arg	Arginine
S	Ser	Serine	T	Thr	Threonine
U	Sec	Selenocysteine	V	Val	Valine
W	Trp	Tryptophan	X	Xaa	Any amino acid
Y	Tyr	Tyrosine	Z	Glx	Glutamine or Glutamic acid
*		translation stop	-		gap of indeterminate length
O	Pyl	Pyrrolysine			

格式 | FASTA | FASTA vs. Sequence

FASTA

```
>gi|183121|gb|M29645.1|HUMGFIII Human insulin-like growth factor II mRNA, complete cds
CAGGGGCCGAAGAGTCACCACCGAGCTTGTGTGGGAGGAGGTGGATTCCAGCCCCCAGCCCCAGGGCTCT
GAATCGCTGCCAGCTCAGCCCCCTGCCCAGCCTGCCCCACAGCCTGAGCCCCAGCAGGCCAGAGAGCCCA
GTCCTGAGGTGAGCTGCTGTGGCCTGTGGCCAGGCGACCCCAGCGCTCCCAGAACTGAGGCTGGCAGCCA
GCCCCAGCCTCAGCCCCAACTGCGAGGCAGAGAGACACCAATGGGAATGCCAATGGGGAAGTCGATGCTG
GTGCTTCTCACCTTCTTGGCCTTCGCCTCGTGCTGCATTGCTGCTTACCGCCCCAGTGAGACCCTGTGCG
GCGGGGAGCTGGTGGACACCCTCCAGTTCGTCTGTGGGGACCGCGGCTTCTACTTCAGCAGGCCCGCAAG
CCGTGTGAGCCGTCGCAGCCGTGGCATCGTTGAGGAGTGCTGTTTCCGCAGCTGTGACCTGGCCCTCCTG
GAGACGTACTGTGCTACCCCCGCCAAGTCCGAGAGGGACGTGTCGACCCCTCCGACCGTGCTTCCGGACA
ACTTCCCCAGATACCCCGTGGGCAAGTTCTTCCAATATGACACCTGGAAGCAGTCCACCCAGCGCCTGCG
CAGGGGCCTGCCTGCCCTCCTGCGTGCCCGCCGGGGTCACGTGCTCGCCAAGGAGCTCGAGGCGTTCAGG
GAGGCCAAACGTCACCGTCCCCTGATTGCTCTACCCACCCAAGACCCCGCCCACGGGGGCGCCCCCCCAG
AGATGGCCAGCAATCGGAAGTGAGCAAAACTGCCGCAAGTCTGCAGCCCGGCGCCACCATCCTGCAGCCT
CCTCCTGACCACGGACGTTTCCATCAGGTTCCATCCCGAAATCTCTCGGTTCCACGTCCCCCTGGGGCTT
CTCCTGACCCAGTCCCCGTGCCCCGCCTCCCCGAAACAGGCTACTCTCCTCGGCCCCCTCCATCGGGCTG
AGGAAGCACAGCAGCATCTTCAAACATGTACAAAATCGATTGGCTTTAAACACCTTCACATACCT
```

Sequence

- GTACGACGGAGTGTTATAAGATGGGAAATCGGATACCAGATGAAATTGTGGATCAG
- MWTALPLLCAGAWLLSAGATAELTVNAIEKFHFTSWMKQHQKTYSSREYSHRLQVFAN

格式 | BED (Browser Extensible Data)

chr7	127471196	127472363	Pos1	0	+	127471196	127472363	255,0,0
chr7	127472363	127473530	Pos2	0	+	127472363	127473530	255,0,0
chr7	127473530	127474697	Pos3	0	+	127473530	127474697	255,0,0
chr7	127474697	127475864	Pos4	0	+	127474697	127475864	255,0,0
chr7	127475864	127477031	Neg1	0	-	127475864	127477031	0,0,255
chr7	127477031	127478198	Neg2	0	-	127477031	127478198	0,0,255
chr7	127478198	127479365	Neg3	0	-	127478198	127479365	0,0,255
chr7	127479365	127480532	Pos5	0	+	127479365	127480532	255,0,0
chr7	127480532	127481699	Neg4	0	-	127480532	127481699	0,0,255

格式 | BED#

BED#

BED12 包含全部 12 列

BED6 chrom, start, end, name, score, and strand

BED5 chrom, start, end, name, and score

BED4 chrom, start, end, and name

BED3 chrom, start, and end

例子

```
chr1	11873	14409	uc001aaa.3	0	+	11873	11873	0
3	354,109,1189,	0,739,1347,
```

格式 | BED#

BED#

BED12	包含全部 12 列
BED6	chrom, start, end, name, score, and strand
BED5	chrom, start, end, name, and score
BED4	chrom, start, end, and name
BED3	chrom, start, and end

例子

chr1	11873	14409	uc001aaa.3	0	+

格式 | BED#

BED#

BED12	包含全部 12 列
BED6	chrom, start, end, name, score, and strand
BED5	chrom, start, end, name, and score
BED4	chrom, start, end, and name
BED3	chrom, start, and end

例子

chr1	11873	14409	uc001aaa.3	0

格式 | BED#

BED#

BED12	包含全部 12 列
BED6	chrom, start, end, name, score, and strand
BED5	chrom, start, end, name, and score
BED4	chrom, start, end, and name
BED3	chrom, start, and end

例子

chr1	11873	14409	uc001aaa.3

格式 | BED#

BED#

BED12	包含全部 12 列
BED6	chrom, start, end, name, score, and strand
BED5	chrom, start, end, name, and score
BED4	chrom, start, end, and name
BED3	chrom, start, and end

例子

chr1 11873 14409

格式 | GFF (General Feature Format)

```
##gff-version 3
ctg123 . operon        1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA          1300  9000  .  +  .  ID=mrna0001;Parent=operon001;Name=sonichedgehog
ctg123 . exon          1300  1500  .  +  .  Parent=mrna0001
ctg123 . exon          1050  1500  .  +  .  Parent=mrna0001
ctg123 . exon          3000  3902  .  +  .  Parent=mrna0001
ctg123 . exon          5000  5500  .  +  .  Parent=mrna0001
ctg123 . exon          7000  9000  .  +  .  Parent=mrna0001
ctg123 . mRNA         10000 15000  .  +  .  ID=mrna0002;Parent=operon001;Name=subsonicsquirrel
ctg123 . exon         10000 12000  .  +  .  Parent=mrna0002
ctg123 . exon         14000 15000  .  +  .  Parent=mrna0002
```

格式 | VCF (Variant Call Format)

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS    ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1         Sample2         Sample3
2      4370   rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330   .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696 rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237 .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567 microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4        0/2:17:2        1/1:40:3
```

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

文本 | 纯文本 vs. 格式化文本

P8_Aln_Pro - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
CLUSTAL X (1.83) multiple sequence alignment


RGDV_ABC75537      MSRQAWIETSALIECISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
RGDV_AAY14576      MSRQAWIETSALIECISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
RGDV_BAA02676      MSRQAWIETSALIECISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
RGDV_AAY14579      MSRQAWIETSALIECISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
RGDV_AAY14580      MSRQAWIETSALIECISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
RGDV_AA064253      MSRQAWIETSALIERISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
RGDV_AAY14577      MSRQAWIETSALTECISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
RGDV_AAY14578      MSRQAWIETSALIECISEYGTKCSFDTFQGLTINDISTLSNLMNQISVASVGFLNDPRTP
WTV_P17380         MSRQNWVETSALVECISEYIVRSYGDTFIGLTSTDLSTLSNLLSNLSIANVGFLNDLRTP
RDVS_Q85451        MSRQMWLDTSALLEAISEYVVRCNGDTFSGLTTGDFNALSNMFTQLSVSSAGYVSDPRVP
RDVO_P17379        MSRQMWLDTSALLEAISEYVVRCNGDTFSGLTTGDFNALSNMFTQLSVSSAGYVSDPRVP
RDVA_Q85449        MSRQMWLDTSALLEAISEYVVRCNGDTFSGLTTGDFNALSNMFTQLSVSSAGYVSDPRVP
RDVF_Q85439        MSRQMWLDTSALLEAISEYVVRCNGDTFSGLTTGDFNALSNMFTQLSVSSAGYVSDPRVP
                   **** *::**** * **** .:.  *** ***  *:.:***::.::*::..*::.* *.*

RGDV_ABC75537      LQAMSCEFVNFISTADRHAYMLQKNWFDSDVAPNVTTDNFIATYIKPRFSRTVSDVLRQV
RGDV_AAY14576      LQAMSCEFVNFISTADRHAYMLQKNWFDSDVAPNVTTDNFIATYIKPRFSRTVSDVLRQV
RGDV_BAA02676      LQAMSCEFVNFISTADRHAYMLQKNWFDSDVAPNVTTDNFIATYIKPRFSRTVSDVLRQV
RGDV_AAY14579      LQAMSCEFVNFISTADRHAYMLQKNWFDSDVAPNVTTDNFIATYIKPRFSRTVSDVLRQV
RGDV_AAY14580      LQAMSCEFVNFISTADRHAYMLQKNWFDSDVAPNVTTDNFIATYIKPRFSRTVSDVLRQV
RGDV_AA064253      LQAMSCEFVNFISTADRHAYMLQKNWFDSDVAPNVTTDNFIATYIKPRFSRTVSDVLRQV
RGDV_AAY14577      PQAMSCEFVNFISAADRHAYMLQKNWFDSDVAPNVTTDNFIATYIKPRFSRTVSDVLRQV
```

Ln 1, Col 1

Work on Word docs at the same time with others using SkyDrive.docx - M..

collaborate. Today, I'm thrilled to announce that we are taking a step towards improving collaboration – by bringing simultaneous editing to Word Web App in addition to Microsoft Word 2010, Microsoft Word for Mac 2011. (Word now joins Excel and OneNote on the web with simultaneous editing.)

It's no secret that we love Word around here. It's used to author all the specifications we write for our products, as well as all the blog posts we publish on this blog. These are professional documents that we produce all the time and there are countless millions of people that have been using Word to express and communicate thoughts and ideas.

Word has a long history of innovation. I remember when the red squiggly line arrived and I no longer had to manually spell check all the school papers I wrote. I also remember when Word introduced auto-correct and many common mistakes were corrected for me (still get corrected to this day). I remember when Outlook started to use Word as its default mail editor and all the power of Word arrived in the place that I write the most. In short, I have grown up with Word, first on the Mac, and now on the PC and am happy we can offer yet another feature to enable you to be more productive.

Harrison Hoffman

In fact, this post was written using Word 2010 on my ThinkPad, while my colleague Harrison has been making changes to this post such as inserting screen shots to demonstrate certain word features. And we even made a video of us doing this. It doesn't get more meta than this!

<insert video>

Here are some of the features that showcase how word communicates to you about changes collaborates are making

Notifications of other collaborators

Work on Word docs at the same time with others using SkyDrive

文本 | 换行符

三大类

- Windows：`\r\n`（CR+LF，回车 + 换行），文件尾部直接 EOF（文件结束标志）
- Unix：`\n`（LF，仅有换行），文件最后一行也会增加该字符，然后才是 EOF
- Mac：`\r`（CR，仅有回车）

识别与转换

- Windows：文本编辑器，如 Notepad++
- Unix：file 识别，fromdos/todos 或者 dos2unix/unix2dos 转换

文本 | 换行符

三大类

- Windows：`\r\n`（CR+LF，回车 + 换行），文件尾部直接 EOF（文件结束标志）
- Unix：`\n`（LF，仅有换行），文件最后一行也会增加该字符，然后才是 EOF
- Mac：`\r`（CR，仅有回车）

识别与转换

- Windows：文本编辑器，如 Notepad++
- Unix：file 识别，fromdos/todos 或者 dos2unix/unix2dos 转换

文本 | 编辑器

Classical learning curves for some common editors
Notepad
Pico
Visual Studio
vi
emacs

文本 | 编辑器 | Notepad++

文本 | 编辑器 | Vim

[No file] - GVIM
File Edit Tools Syntax Buffers Window Help
VIM - Vi IMproved
version 6.0.152
by Bram Moolenaar et al.
Vim is open source and freely distributable
Help poor children in Uganda!
type :help iccf<Enter> for information
type :q<Enter> to exit
type :help<Enter> or <F1> for on-line help
type :help version6<Enter> for version info
0,0-1 All

文本 | 编辑器 | Emacs

emacs@bandgap
File Edit Options Buffers Tools Help
This is GNU Emacs, one component of the GNU/Linux operating system.
GNU Emacs 23.1.50.1 (x86_64-pc-linux-gnu, GTK+ Version 2.16.1)
of 2009-07-31 on platinum, modified by Debian
Copyright (C) 2009 Free Software Foundation, Inc.
Authors Many people have contributed code included in GNU Emacs
Contributing How to contribute improvements to Emacs
GNU and Freedom Why we developed GNU Emacs, and the GNU operating system
Absence of Warranty GNU Emacs comes with ABSOLUTELY NO WARRANTY
Copying Conditions Conditions for redistributing and changing Emacs
Getting New Versions How to obtain the latest version of Emacs
Ordering Manuals Buying printed manuals from the FSF
Emacs Tutorial Learn basic Emacs keystroke commands
Emacs Guided Tour See an overview of the many facilities of GNU Emacs
---- *About GNU Emacs* ----
menu-bar options menu-set-font

文本 | 编辑器 | Sublime Text

Soda Light.sublime-theme — Theme - Soda

OPEN FILES
- Soda Light.sublime-theme
- Soda Dark.sublime-theme

FOLDERS
- Theme - Soda
  - Soda Dark
  - Soda Light
  - README.md
  - Soda Dark.sublime-theme
  - Soda Light.sublime-theme

Soda Light.sublime-theme | Soda Dark.sublime-theme

soda light

658 **Soda Light**.sublime-theme
Soda Light.sublime-theme

532 Widget - **Soda Light**.stTheme
Soda Light/Widget - **Soda Light**.stTheme

532 Widget - **Soda Light**.sublime-settings
Soda Light/Widget - **Soda Light**.sublime-settings

```
        "layer0.texture": "Theme - Soda/Soda Light/tabset-
            background.png",
        "layer0.inner_margin": [1, 7],
        "layer0.opacity": 1.0,
        "content_margin": [-4, 0, -4, 3],
        "tab_overlap": 5,
        "tab_width": 180,
        "tab_min_width": 45,
        "tab_height": 25,
        "mouse_wheel_switch": false
    },
    {
```

Find What: dark | Find | Replace

Replace With: light | Find All | Replace All

Line 15, Column 1 | Spaces: 4 | JSON

文本 | 编辑器 | Atom

dockra

- resources/osx
- src
- .babelrc
- .eslintrc
- .gitignore
- gulpfile.babel.js
- package.json
- README.md
- webpack.config.js

gulpfile.babel.js

```
46          'node_modules/'
47        ]
48      }).on('error', $.sass.logError))
49      .pipe($.concat('main.css'))
50      .pipe($.minifyCss())
51      .pipe($.sourcemaps.write())
52      .pipe(gulp.dest('build/dev'));
53  });
54
55  gulp.task('javascript', () => {
56    return gulp.src(javascriptFiles)
57      .pipe($.plumber())
58      .pipe($.sourcemaps.init())
59      .pipe($.babel())
60      .pipe($.uglify())
61      .pipe($.sourcemaps.write())
62      .pipe(gulp.dest('build/dev'));
63  });
64
65  gulp.task('html', () => {
66    return gulp.src(htmlFiles)
67      .pipe(gulp.dest('build/dev/'));
68  });
69
70  gulp.task('images', () => {
71    return gulp.src(imageFiles)
72      .pipe(gulp.dest('build/dev/assets/images/'));
73  });
74
75  gulp.task('fonts', () => {
76    return gulp.src(fontFiles)
```

Line 0 | File 0 | Project 0 | gulpfile.babel.js | ✔ Gulp: build | 72:34 | LF | Normal | UTF-8 | JavaScript | master

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

逻辑运算 | 常见问题

数据

gene1	chr1	10	20	+
gene2	chr1	40	60	+
gene3	chr1	50	100	+
snp1	chr1	15	+	
snp2	chr1	55	-	
exon3.1	chr1	50	60	+
exon3.2	chr1	90	100	+

问题

1. 找到 gene1 和 gene2 之间的基因间区域。
2. snp1 在 gene1 上吗？snp2 在 gene1 上吗（，在 gene2 上吗）？
3. 找到与 gene3 重叠和不重叠的基因？
4. 找到 gene3 的内含子区域。

逻辑运算 | 常见问题

数据

gene1	chr1	10	20	+
gene2	chr1	40	60	+
gene3	chr1	50	100	+
snp1	chr1	15	+	
snp2	chr1	55	-	
exon3.1	chr1	50	60	+
exon3.2	chr1	90	100	+

问题

1. 找到 gene1 和 gene2 之间的基因间区域。
2. snp1 在 gene1 上吗？snp2 在 gene1 上吗（，在 gene2 上吗）？
3. 找到与 gene3 重叠和不重叠的基因？
4. 找到 gene3 的内含子区域。

求同

交集

求异

相加

并集

相减 A-B

差集

相减 B-A

补集

逻辑运算 | 集合 ⇒ 基因组

2
1 3
4 5
3 4
5 6
x2
x1 x3
x4 x5
x3 x4
x5 x6
chr1:1-5
chr1:3-6
chr1 1 5
chr1 3 6
chr1:2
chr1:1 chr1:3
chr1:4 chr1:5
chr1:3 chr1:4
chr1:5 chr1:6
chr1 1 2 3 4 5 6

逻辑运算 | 运算模式

A Intersect
First query
Input:
Intervals to intersect with (second query)
Overlapping intervals
Output:
Overlapping pieces of intervals
B Subtract
First query
Input:
Intervals to subtract (second query)
Intervals with no overlap
Output:
Non-overlapping pieces of intervals
C Merge
Query
Input:
Merged intervals
Output:
D Concatenate
First query
Input:
Second query
Concatenate
Output:
E Complement
Query
Input:
Complement
Output:
F Cluster
Query
Input:
Find clusters
Output:
Merge clusters

逻辑运算 | intersect

逻辑运算 | intersect

```
Chromosome   ==================================================================
BED/BAM  A      ===================              =============
BED File B                 ========
Result                     ========
```

逻辑运算 | subtract

逻辑运算 | subtract

```
Chromosome   ==================================================================
BED File A                ==============              ======
BED File B          ========                  ===========
Result                          =========
```

逻辑运算 | merge

逻辑运算 | merge

```
Chromosome  ==================================================================
BED File        =============   ===============   =====================
                       ========
Result          ==============================   =====================
```

逻辑运算 | concatenate

逻辑运算 | complement

逻辑运算 | complement

```
Chromosome  ==================================================================
BED File        =============   ===============     ==================
Result      ===             ===                 =====                 =======
```

逻辑运算 | cluster

逻辑运算 | cluster

逻辑运算 | join

Input

```
Query 1:

chr1    10      100     Query1.1
chr1    500     1000    Query1.2
chr1    1100    1250    Query1.3

Query 2:

                                chr1    20      80      Query2.1
                                chr1    2000    2204    Query2.2
                                chr1    2500    3000    Query2.3
```

Output

```
(Return only records that are joined)
chr1    10      100     Query1.1    chr1    20      80      Query2.1
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

逻辑运算 | join

Input

```
Query 1:

chr1    10      100     Query1.1
chr1    500     1000    Query1.2
chr1    1100    1250    Query1.3

Query 2:

                        chr1    20      80      Query2.1
                        chr1    2000    2204    Query2.2
                        chr1    2500    3000    Query2.3
```

```
(Return all records of first query)
chr1    10      100     Query1.1    chr1    20      80      Query2.1
chr1    500     1000    Query1.2    .       .       .       .
chr1    1100    1250    Query1.3    .       .       .       .
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

逻辑运算 | join

Input

```
Query 1:

chr1    10      100     Query1.1
chr1    500     1000    Query1.2
chr1    1100    1250    Query1.3

Query 2:

                                chr1    20      80      Query2.1
                                chr1    2000    2204    Query2.2
                                chr1    2500    3000    Query2.3
```

```
(Return all records of second query)
chr1    10      100     Query1.1    chr1    20      80      Query2.1
.       .       .       .           chr1    2000    2204    Query2.2
.       .       .       .           chr1    500     3000    Query2.3
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

逻辑运算 | join

Input

```
Query 1:

chr1    10      100     Query1.1
chr1    500     1000    Query1.2
chr1    1100    1250    Query1.3

Query 2:

                        chr1    20      80      Query2.1
                        chr1    2000    2204    Query2.2
                        chr1    2500    3000    Query2.3
```

```
(Return all records of both queries)
chr1    10      100     Query1.1    chr1    20      80      Query2.1
chr1    500     1000    Query1.2    .       .       .       .
chr1    1100    1250    Query1.3    .       .       .       .
.       .       .       .           chr1    2000    2200    Query2.2
.       .       .       .           chr1    2500    3000    Query2.3
```

Return only records that are joined (INNER JOIN)
Return all records of first query (fill null with ".")
Return all records of second query (fill null with ".")
Return all records of both queries (fill nulls with ".")

逻辑运算 | 其他

coverage	Finds the number of bases each interval in the first dataset covers of the second dataset.
flank	Finds the upstream and/or downstream flanking region(s).
closest	Find the closest, potentially non-overlapping upstream and/or downstream features.
slop	Adjust the size of intervals.
window	Find overlapping intervals within a window around an interval.

逻辑运算 | 实例

Dataset 1

chr1	10	49	Feature1.1
chr1	70	119	Feature1.2
chr1	170	209	Feature1.3
chr1	180	229	Feature1.4

Dataset 2

chr1	80	109	Feature2.1
chr1	150	199	Feature2.2
chr1	250	289	Feature2.3
chr1	270	309	Feature2.4

逻辑运算 | 实例

Dataset 1

chr1	10	49	Feature1.1
chr1	70	119	Feature1.2
chr1	170	209	Feature1.3
chr1	180	229	Feature1.4

Dataset 2

chr1	80	109	Feature2.1
chr1	150	199	Feature2.2
chr1	250	289	Feature2.3
chr1	270	309	Feature2.4

intersect

chr1	80	109	Feature3.1
chr1	170	199	Feature3.2
chr1	180	199	Feature3.3

逻辑运算 | 实例

Dataset 1

chr1	10	49	Feature1.1
chr1	70	119	Feature1.2
chr1	170	209	Feature1.3
chr1	180	229	Feature1.4

Dataset 2

chr1	80	109	Feature2.1
chr1	150	199	Feature2.2
chr1	250	289	Feature2.3
chr1	270	309	Feature2.4

subtract (1-2)

chr1	10	49	Feature4.1
chr1	70	80	Feature4.2
chr1	109	119	Feature4.3
chr1	199	209	Feature4.4
chr1	199	229	Feature4.5

subtract (2-1)

chr1	150	170	Feature5.1
chr1	250	289	Feature5.2
chr1	270	309	Feature5.3

逻辑运算 | 实例

Dataset 1

chr1	10	49	Feature1.1
chr1	70	119	Feature1.2
chr1	170	209	Feature1.3
chr1	180	229	Feature1.4

Dataset 2

chr1	80	109	Feature2.1
chr1	150	199	Feature2.2
chr1	250	289	Feature2.3
chr1	270	309	Feature2.4

join

chr1	70	119	Feature1.2	chr1	80	109	Feature2.1
chr1	170	209	Feature1.3	chr1	150	199	Feature2.2
chr1	180	229	Feature1.4	chr1	150	199	Feature2.2

逻辑运算 | 应用

实际问题

1. Find genes that overlap LINEs.
2. Remove introns from gene features. Exons will (should) be reported.
3. Merge overlapping repetitive elements into a single entry.
4. Report all intervals in the human genome that are not covered by repetitive elements.

解决策略

1. intersect
2. subtract
3. merge
4. complement

逻辑运算 | 应用

实际问题

1. Find genes that overlap LINEs.
2. Remove introns from gene features. Exons will (should) be reported.
3. Merge overlapping repetitive elements into a single entry.
4. Report all intervals in the human genome that are not covered by repetitive elements.

解决策略

1. intersect
2. subtract
3. merge
4. complement

逻辑运算 | 工具

- Galaxy 中的“Operate on Genomic Intervals”工具集
- bedtools: a powerful toolset for genome arithmetic
- BEDOPS: the fast, highly scalable and easily-parallelizable genome analysis toolkit

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

总结与答疑

知识点——基因组注释基础

- 基因组组装版本——对应关系
- 两种坐标系统——0-based 和 1-based
- 四种常用格式——FASTA，BED，GFF，VCF
- 坐标逻辑运算——常见模式及其适用范围
- 坐标转换、格式转换、逻辑运算的工具

技能——纯文本与文本编辑器

- 纯文本与格式化文本
- 不同操作系统中的换行符
- 文本编辑器——Notepad++，Vim，Emacs

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

引言

前期准备工作

- 组装版本
- 坐标系统
- 常用格式
- 逻辑运算

后续功能注释

- 变异位点的注释
- 基因集富集分析
- 制作序列标识
- ...

引言

前期准备工作

- 组装版本
- 坐标系统
- 常用格式
- 逻辑运算

后续功能注释

- 变异位点的注释
- 基因集富集分析
- 制作序列标识
- ...

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

变异位点的注释 | SNP

ID	Chromosome	Position	Reference	Mutation	TotalHit
SNP00000001	15	20833654	C	A	38
SNP00000002	15	23501058	C	A	31
SNP00000003	15	45564496	C	A	20
SNP00000004	15	45564498	A	T	20
SNP00000005	15	45564501	G	T	20
SNP00000006	15	45564504	C	A	20
SNP00000007	15	45564505	C	T	20
SNP00000008	15	45564506	T	C	20
SNP00000009	15	45564508	A	T	20
SNP00000010	15	50212324	G	C	21
SNP00000011	15	50212325	A	T	21
SNP00000012	15	50212326	A	C	21
SNP00000013	15	50212328	G	A	21
SNP00000014	15	50212329	A	G	21
SNP00000015	15	50212330	G	A	21
SNP00000016	15	50212342	G	T	21
SNP00000017	15	52098626	A	G	20
SNP00000018	15	52098627	G	A	20

变异位点的注释 | SNP 注释

Pos	Alt SNP	Ref SNP	Alt SNP Codon	Ref SNP Codon	Alt SNP AA	Ref 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	--	--	--	--	Downstram

变异位点的注释 | 注释工具

- SNVs 的注释：SeattleSeq Annotation、variant tools、SnpEff、ANNOVAR
- 非同义多态性的功能注释：SIFT、PolyPhen-2、SNPs3D
- indels 的功能注释：PROVEAN

变异位点的注释 | 结果解析 | SeattleSeq Annotation

File:
/data/jboss-as-7.1.1.Final/gvsBatchOutput/SeattleSeqAnnotation137.1individual.294000040650.txt

Title:
1individual

Counts:
HapMapFreqType HapMapFreqMinor
polyPhenType polyPhenScore

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
- ☑ Tandem Repeats
- ☑ 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 36 accession lines **page 1 of 1**

inDBSNPOrNot	chromosome	position	referenceBase	sampleGenotype	sampleAlleles	allelesDBSNP	accession	functionGVS	functionDBSNP	rsID	aminoAcids	proteinPositi
dbSNP_130	10	1126383	A	R	A/G	A/G	NM_014023.3	coding-synonymous	synonymous-codon	73578536	none	121/495
dbSNP_86	10	3150973	C	Y	C/T	C/T	NM_001242339.1	coding-synonymous	synonymous-codon	1132173	none	309/777
dbSNP_86	10	3150973	C	Y	C/T	C/T	NM_002627.4	coding-synonymous	synonymous-codon	1132173	none	317/785

变异位点的注释 | 结果解析 | SeattleSeq Annotation

inDBSNPOrNot	chromosome	position	referenceBase	sampleGenotype	sampleAlleles	allelesDBSNP
dbSNP_130	10	1126383	A	R	A/G	A/G
dbSNP_86	10	3150973	C	Y	C/T	C/T
dbSNP_86	10	3150973	C	Y	C/T	C/T

accession	functionGVS	functionDBSNP	rsID	aminoAcids	proteinPositi
NM_014023.3	coding-synonymous	synonymous-codon	73578536	none	121/495
NM_001242339.1	coding-synonymous	synonymous-codon	1132173	none	309/777
NM_002627.4	coding-synonymous	synonymous-codon	1132173	none	317/785

变异位点的注释 | 结果解析 | SIFT

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

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

富集分析 | 基因集

Table 7 The minimum gene set selected in PRI dataset (gene scores rank from high to low)

Probe ID	Gene symbol	Gene name	Chromosomal regions
219868_s_at	ANKFY1	Ankyrin repeat and FYVE domain containing1	17p13.3
213613_s_at	NADK	NAD kinase	1p36.33-p36.21
208002_s_at	ACOT7	Acyl-coa thioesterase 7	1p36
222133_s_at	PHF20L1	PHD finger protein 20-like 1	8q24.22
203858_s_at	COX10	COX10 homolog, cytochrome c oxidase assembly protein, heme A: farnesyltransferase (yeast)	17p12
204051_s_at	SFRP4	Secreted frizzled-related protein 4	7p14.1
207567_at	SLC13A2	Solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 2	17p13.2
225803_at	FBXO32	F-box protein 32	8q24.13
205527_s_at	GEMIN4	Gem (nuclear organelle) associated protein 4	17p13
207017_at	RAB27B	RAB27B, member RAS oncogene family	18q21.2
206746_at	BFSP1	Beaded filament structural protein 1, filensin	20p12.1
217099_s_at	GEMIN4	Gem (nuclear organelle) associated protein 4	17p13
233638_s_at	POMGNT1	Protein O-linked mannose beta1,2-N-acetylglucosaminyltransferase	1p34.1
217381_s_at	TRGV5	T cell receptor gamma variable 5	7p14

富集分析 | 数据库与分析工具

数据库

GO Gene Ontology

KEGG Kyoto Encyclopedia of Genes and Genomes

分析工具

DAVID Database for Annotation, Visualization and Integrated Discovery

富集分析 | 数据库与分析工具

数据库

GO Gene Ontology

KEGG Kyoto Encyclopedia of Genes and Genomes

分析工具

DAVID Database for Annotation, Visualization and Integrated Discovery

富集分析 | GO

三个方面

- biological process，BP，生物学过程
- molecular function，MF，分子功能
- cellular component，CC，细胞组份

两大关系

- is_a: for simple, hierarchical connections between terms
- part_of: for describing how the components of a living system fit together

富集分析 | GO

三个方面

- biological process，BP，生物学过程
- molecular function，MF，分子功能
- cellular component，CC，细胞组份

两大关系

- is_a: for simple, hierarchical connections between terms
- part_of: for describing how the components of a living system fit together

富集分析 | GO

cellular_component
biological_process
molecular_function
pigmentation
regulation of biological process
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
negative regulation of pigmentation during development
positive regulation of pigmentation during development
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation

富集分析 | GO

Gene ontology
Biological process
Cellular process
Response to stimulus
Cell communication
Cellular physiological process
Response to stress
Signal transduction
Response to endogenous stimulus
Intracellular signalling
Cellular metabolism
Cell surface receptor linked signalling
Intracellular receptor-mediated signalling
Response to DNA-damage stimulus
Enzyme-linked receptor signalling
Cell cycle
Cell death
Regulation of cellular metabolism
Programmed cell death
Steroid hormone receptor signalling
Transmemrane RTK signalling
Cell proliferation
Nucleic acid metabolism
DNA repair
Apoptosis
DNA metabolism
MMR
Androgen receptor signalling
Regulation of cell-cycle progression
Phosphorus metabolism
Regulation of apoptosis
Transcription
DNA replication
NER
Negative regulation of cell cycle progression
Phosphate metabolism
Negative regulation of apoptosis
DNA-dependent DNA replication
Cell-cycle checkpoint
Regulation of transcription
Phosphorylation
Anti-apoptosis
Transcription-coupled NER
Transcription, DNA-dependent
Transcription from RNA pol II promoter
Regulation of transcription, DNA-dependent
Nature Reviews | Cancer

富集分析 | GO

B
biological process (5627)
cell cycle (286)
M phase (202)
prophase (4)
spindle organization (66)
metaphase (1)
anaphase (4)
telophase (0)
cul-2
elc-1
rbx-1
zyg-11
Relationships
part_of
is_a
develops_from
inferred
annotation
Nodes
ontology term
root node
cul-2
gene name

富集分析 | GO

cellular process (0)
single-organism process (0)
is a
is a
single-organism cellular process (0)
is a
is a
is a
cell cycle (318)
part of
cell division (198)
cell cycle process (0)
is a
is a
part of
cytokinetic process (0)
part of
is a
cytokinesis (1)
contractile ring contraction (0)
part of
is a
actomyosin contractile ring contraction (4)
is a
regulates
mitotic actomyosin contractile ring contraction (0)
regulation of actomyosin contractile ring contraction (0)

富集分析 | GO

receptor-mediated endocytosis
GO:0006898
is_a
endocytosis
GO:0006897
is_a
is_a
membrane invagination
GO:0010324
vesicle-mediated transport
GO:0016192
is_a
is_a
membrane organization
GO:0016044
transport
GO:0006810
is_a
is_a
is_a
cellular component organization
GO:0016043
establishment of localization
GO:0051234
is_a
part_of
cellular process
GO:0009987
is_a
localization
GO:0051179
is_a
is_a
biological_process
GO:0008150

富集分析 | KEGG

Glycan Biosynthesis and Metabolism
Lipid Metabolism
Carbohydrate Metabolism
Amino Acid Metabolism
Energy Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid
Data on KEGG graph
Rendered by Pathview

富集分析 | KEGG

CELL CYCLE
-1
0
1
Growth factor
Growth factor withdrawal
TGFβ
GSK3β
p107
E2F4,5
DP-1,2
Smad2,3
Smad4
c-Myc
Miz1
MAPK signaling pathway
SCF
Skp2
ARF
Mdm2
p300
DNA-PK
ATM/ATR
DNA damage checkpoint
Rb
p53
Apoptosis
Mrs1
Mad1
Mad2
BubR1
Bub3
Bub1
Chk1, 2
Smc1
Smc3
Stag1,2
Rad21
Cohesin
Esp1
Separin
PTTG
Securin
APC/C
Cdc20
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Kip1, 2
p21
Cip1
GADD45
14-3-3σ
PCNA
14-3-3
Cdc25A
Cdc25B,C
Ubiquitin mediated proteolysis
R-point (START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
p107,130
Rb
Abl
HDAC
Cdc6
Cdc45
ORC
MCM
Rb
Wee
Myt1
APC/C
Cdh1
Cdc14
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
Cdc7
Dbf4
Bub2
MEN
ORC (Origin Recognition Complex)
Orc1
Orc2
Orc3
Orc4
Orc5
Orc6
MCM (Mini-Chromosome Maintenance) complex
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
DNA
S-phase proteins, CycE
DNA
DNA biosynthesis
G1
S
G2
M
Data on KEGG graph
Rendered by Pathview

富集分析 | DAVID

- Gene Name Batch Viewer
- Gene ID Conversion Tool
- Gene Functional Classification Tool
- Functional Annotation Tool
  - Functional Annotation Clustering
  - Functional Annotation Chart：富集分析
  - Functional Annotation Table

富集分析 | DAVID | 结果解析

Functional Annotation Chart

Help and Manual

Current Gene List: demolist1
Current Background: Homo sapiens
155 DAVID IDs

⊞ **Options**

Rerun Using Options | Create Sublist

105 chart records

Download File

1, 2, 3, 4, 5, 6

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

富集分析 | DAVID | 工具选择

Highly recommended
Recommended

	Gene ID conversion tool	Gene name batch viewer	Gene functional classification	Functional annotation chart	Functional annotation clustering	Functional annotation table
Convert gene IDs from one type to another	■			▒		
Diagnose and fix problems of gene IDs	■	▒				▒
Explore gene names in batch		■	▒			▒
Discover enriched functionally related gene groups			■	▒	▒	
Display relationship of many-genes-to-many-terms on 2D view.			■		■	
Initial glance of major biological functions associated with gene list		▒	■	■	■	
Identify enriched (overrepresented) annotation terms				■	■	
Visualize genes on BioCarta and KEGG pathway maps				■	■	
Link gene–disease associations				■	■	
Highlight protein functional domains and motifs				■	■	
Redirect to related literatures				▒		■
List interacting proteins				■	■	▒
Cluster redundant and heterozygous annotation terms					■	
Search other functionally similar genes in genome, but not in list		■	■	1	1	
Search other annotations functionally similar to one of my interests				■	■	
Read all annotation contents associated with a gene						■

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

序列标识 | 徽标

序列标识

定义

序列标识图是显示序列保守区域的共识序列、每个位置上各个氨基酸或核苷酸出现的频率以及各个位点上的序列信息量的一种可视化方法。

含义

根据序列保守区域的多序列比对来绘制序列标识图。

在一个标识图像里，由大小不一的字符形成的一个堆栈代表序列保守区域的一个位点。每个核苷酸或氨基酸的高度和它在对应位点上出现的频率成比例。堆栈的总高度代表对应位点上的序列信息，以比特（bit）为单位。在每个堆栈里，字符按其出现的频率大小自上而下排列。所以，位于各个堆栈最上方的字符组成保守区域的共识序列。

序列标识

定义

序列标识图是显示序列保守区域的共识序列、每个位置上各个氨基酸或核苷酸出现的频率以及各个位点上的序列信息量的一种可视化方法。

含义

根据序列保守区域的多序列比对来绘制序列标识图。

在一个标识图像里，由大小不一的字符形成的一个堆栈代表序列保守区域的一个位点。每个核苷酸或氨基酸的高度和它在对应位点上出现的频率成比例。堆栈的总高度代表对应位点上的序列信息，以比特（bit）为单位。在每个堆栈里，字符按其出现的频率大小自上而下排列。所以，位于各个堆栈最上方的字符组成保守区域的共识序列。

序列标识

序列标识（sequence logo）

- 数据：多序列比对信息
- 横轴：序列坐标位置
- 纵轴：比特，计量单位
- 总高度：信息量/保守性
- 相对高度：相对频率
- 位置自上而下：频率由大到小
- 制作工具：WebLogo, enoLOGOS, Skylign

序列标识

序列标识（sequence logo）

- 数据：多序列比对信息
- 横轴：序列坐标位置
- 纵轴：比特，计量单位
- 总高度：信息量/保守性
- 相对高度：相对频率
- 位置自上而下：频率由大到小
- 制作工具：WebLogo, enoLOGOS, Skylign

序列标识 | 着色 | WebLogo2

序列标识 | 着色 | WebLogo3

序列标识 | 剪接

5'
Exon 1
Exon 2
3'
Donor
Splice Site
Exonic Nucleotides
Intronic Nucleotides
N N N N N N N G T N N N N N
Acceptor
Splice Site
Intronic Nucleotides
Exonic Nucleotides
N N N N N A G N N N N N N N
Default splicing region
used by Variant Classification

序列标识 | 实例

Exon-Intron (Donor) Sites
bits
2.0
1.0
0.0
-10
-5
0
5
WebLogo 3.3

Intron-Exon (Acceptor) Sites
bits
2.0
1.0
0.0
-20
-15
-10
-5
0
WebLogo 3.3

序列标识 | 实例 | 真实数据

Donor Sites(%)		Acceptor Site(%)	
GT	98.797	AG	99.714
GC	0.920	AC	0.120
AT	0.143	TG	0.032
GA	0.028	AT	0.024
GG	0.025	GG	0.022
CT	0.018	AA	0.019
TT	0.016	CG	0.010
CC	0.011	CC	0.010
TG	0.007	TT	0.009
AG	0.007	CT	0.008
TA	0.006	CA	0.008
AC	0.006	GC	0.007
CA	0.006	TA	0.006
TC	0.004	TC	0.004
AA	0.004	GT	0.004
CG	0.002	GA	0.003

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

box plot | 实例

Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.

box plot | 简介

历史

- box plot, boxplot, Box-whisker Plot
- 箱线图、箱须图、盒须图、盒式图、盒状图，因形状如箱子而得名
- 1977 年由美国着名统计学家约翰 • 图基（John Tukey）发明

简介

- 显示一组数据分散情况的统计图
- 显示最大值、最小值、中位数、下四分位数和上四分位数

优缺点

- 可以粗略地看出数据是否具有有对称性、分布的离散程度
- 适合用于几个样本的比较
- 不能提供关于数据分布偏态和尾重程度的精确度量

box plot | 简介

历史

- box plot, boxplot, Box-whisker Plot
- 箱线图、箱须图、盒须图、盒式图、盒状图，因形状如箱子而得名
- 1977 年由美国着名统计学家约翰 • 图基（John Tukey）发明

简介

- 显示一组数据分散情况的统计图
- 显示最大值、最小值、中位数、下四分位数和上四分位数

优缺点

- 可以粗略地看出数据是否具有有对称性、分布的离散程度
- 适合用于几个样本的比较
- 不能提供关于数据分布偏态和尾重程度的精确度量

box plot | 简介

历史

- box plot, boxplot, Box-whisker Plot
- 箱线图、箱须图、盒须图、盒式图、盒状图，因形状如箱子而得名
- 1977 年由美国着名统计学家约翰 • 图基（John Tukey）发明

简介

- 显示一组数据分散情况的统计图
- 显示最大值、最小值、中位数、下四分位数和上四分位数

优缺点

- 可以粗略地看出数据是否具有有对称性、分布的离散程度
- 适合用于几个样本的比较
- 不能提供关于数据分布偏态和尾重程度的精确度量

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 相关概念

- 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（interquartile range），$IQR = Q3 - Q1$
- 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

box plot | 绘图步骤

- 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。（*注意：统计软件绘制的箱线图一般都没有标出内限和外限。*）
- 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*）表示。

box plot | 绘图步骤

- 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。*（注意：统计软件绘制的箱线图一般都没有标出内限和外限。）*
- 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*）表示。

box plot | 绘图步骤

- 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。*（注意：统计软件绘制的箱线图一般都没有标出内限和外限。）*
- 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*）表示。

box plot | 绘图步骤

- 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。（*注意：统计软件绘制的箱线图一般都没有标出内限和外限。*）
- 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*）表示。

box plot | 绘图步骤

- 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。*（注意：统计软件绘制的箱线图一般都没有标出内限和外限。）*
- 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 *）表示。

box plot | 绘图步骤

- 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。 *（注意：统计软件绘制的箱线图一般都没有标出内限和外限。）*
- 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。

box plot | 绘图步骤

- 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。 *（注意：统计软件绘制的箱线图一般都没有标出内限和外限。）*
- 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。

box plot | 绘图步骤

- 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。 *（注意：统计软件绘制的箱线图一般都没有标出内限和外限。）*
- 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。

box plot | 图解

```
                                  +-----+-+
  *             o      |-------|    + | |---|
                                  +-----+-+

+---+---+---+---+---+---+---+---+---+---+---+
0   1   2   3   4   5   6   7   8   9  10
```

最小值（min）=0.5；下四分位数（Q1）=7；中位数（Med）=8.5；
上四分位数（Q3）=9；最大值（max）=10；平均值 =8；
四分位数差（interquartile range，四分位间距）=Q3−Q1=2。

box plot | 图解

平均值
内限
内限
Mild outliers
Extreme outliers
最小非异常值
中位线
最大非异常值

中位线
IQR
外限
内限
内限
外限
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Extreme outliers
Mild outliers
Median
Mild outliers
Extreme outliers

box plot | 图解

Maximum
Upper Quartile
Median
Lower Quartile
Minimum

outlier
upper extreme
upper whisker
upper hinge
upper fourth
box
median
lower hinge
lower fourth
lower whisker
lower extreme

Sample, $n = 20$
$1.5 \times IQR$
IQR
$1.5 \times IQR$
Q1
m
Q3
Whiskers
Outliers

box plot | 变体

Traditional Box Plot
Notched Box Plot
Variable Width Box Plot
Variable Width Notched Box Plot
0
20
40
60
80
100
Group 1
Group 2

box plot | 变体

a
Uniform
Unimodal
Unimodal 2
Bimodal
0
20
40
60
80
100
Data values

b
s.d.
s.e.m.
Bar plot
Box plot
Violin plot
Bean plot
Data values
0
20
40
60
80
100
Uniform
Unimodal
Unimodal 2
Bimodal

box plot | 工具

- BoxPlotR
- Plotly
- ECplot
- Galaxy（“Graph/Display Data”工具集中的 Boxplot）
- R
- …

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

解析图表 | 基本策略

表格

- 行、列的含义
- 缩写的含义
- 数值的含义

图片

- 生成图片的数据
- 横、纵轴的含义
- 图片包含的元素
- 图片元素属性的含义

解析图表 | 基本策略

表格

- 行、列的含义
- 缩写的含义
- 数值的含义

图片

- 生成图片的数据
- 横、纵轴的含义
- 图片包含的元素
- 图片元素属性的含义

解析图表 | 图形的语法

- 数据（data）：我们想要可视化的对象。
- 几何对象（geom）：用以呈现数据、在图中实际看到的图形元素（点、线、条形、多边形等）。
- 图形属性（aes）：几何对象的视觉属性（位置、颜色、形状、大小和线条类型）。
- 映射（mapping）：从数据中的变量对应到图形属性。
- 统计变换（stats）：对数据进行的某种汇总（将数据分组计数以创建直方图）。
- 标度（scale）：控制着数据空间的值到图形属性空间的值的映射（用颜色、大小或形状来表示不同的取值）。
- 坐标系（coord）：描述了数据是如何映射到图形所在的平面的，它同时提供了看图所需的坐标轴和网格线。
- 引导元素（guide）：向看图者展示了如何将视觉属性映射回数据空间（坐标轴上的刻度线和标签、图例）。
- 分面（facet）：描述了如何将数据分解为各个子集，以及如何对子集作图并联合进行展示。分面也叫做条件作图或网格作图。
- 位置调整：控制着图形对象的重叠。

解析图表 | 图形的语法 | ggplot2

解析图表 | 其他

hand holding
honeymoon
cliff of
confusion
desert of
despair
upswing of
awesome
Job Ready
Confidence
Competence
Resource
Density
Scope of
Knowledge

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

总结与答疑

知识点——基因组功能的高级注释

- 变异位点的注释——用途，注释内容，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

技能——解析图表

- 表——行列，缩写，数值
- 图——数据，横纵轴，图元素，元素属性

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

引言

基础知识

- 组装版本和坐标系统
- 常用格式
- 坐标的逻辑运算

高级注释

- 变异位点的注释
- 基因集富集分析
- 制作序列标识

分析平台

- Galaxy
- GenePattern

引言

基础知识

- 组装版本和坐标系统
- 常用格式
- 坐标的逻辑运算

高级注释

- 变异位点的注释
- 基因集富集分析
- 制作序列标识

分析平台

- Galaxy
- GenePattern

引言

基础知识

- 组装版本和坐标系统
- 常用格式
- 坐标的逻辑运算

高级注释

- 变异位点的注释
- 基因集富集分析
- 制作序列标识

分析平台

- Galaxy
- GenePattern

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

Galaxy | 工具集

- Get Data
- Text Manipulation
- Convert Formats
- Operate on Genomic Intervals
- Phenotype Association
- Statistics
- Graph/Display Data
- NGS Toolbox
- ...

Galaxy | 界面

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
工具
search tools
Get Data
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.
Galaxy 101
Start small
The very first tutorial you need
历史
Unnamed history
1.5 MB
历史已空，请单击左边窗格中'获取数据'

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

Galaxy 演示 | 坐标转换

- UCSC liftOver tool：支持 BED 和“chrN:start-end”格式的输入
- Galaxy（基于 UCSC liftOver tool）：支持 BED、GFF 和 GTF 格式的输入
- CrossMap：支持 SAM/BAM、Wiggle/BigWig、BED、GFF/GTF 和 VCF 格式的输入，输出对应格式
- NCBI Remap：支持 BED、GFF、GTF 和 VCF 等格式的输入
- Ensembl assembly converter（2015 年退休，CrossMap 继位）：支持 BED、GFF、GFT 和 PSL 格式的输入，但输出都是 GFF 格式的
- pyliftover：仅支持点坐标（point coordinates）的转换，无法对区段（ranges）坐标进行转换

hg19 ⇒ hg18

1. 获取输入
2. 数据处理

hg19 ⇒ hg18

1. 获取输入
   - 输入文件：hg19 坐标
2. 数据处理
   - 设置参数：hg19 ⇒ hg18
3. 保存输出

hg19 ⇒ hg18

1. 获取输入
   - 输入文件：hg19 坐标
2. 数据处理
   - 设置参数：hg19 ⇒ hg18
3. 保存输出

hg19 ⇒ hg18

1. 获取输入
   - 输入文件：hg19 坐标
2. 数据处理
   - 设置参数：hg19 ⇒ hg18
3. 保存输出
   - 过滤结果：MAPPED VS. UNMAPPED

hg19 ⇒ hg18

1. 获取输入
   - 输入文件：hg19 坐标
2. 数据处理
   - 设置参数：hg19 ⇒ hg18
3. 保存输出
   - 过滤结果：MAPPED VS. UNMAPPED

Galaxy 演示 | 坐标转换 | liftOver

hg19 ⇒ hg18

1. 获取输入
   - 输入文件：hg19 坐标
2. 数据处理
   - 设置参数：hg19 ⇒ hg18
3. 保存输出
   - 过滤结果：MAPPED VS. UNMAPPED

Galaxy 演示 | 坐标转换 | liftOver

hg19 ⇒ hg18

1. 获取输入
   - 输入文件：hg19 坐标
2. 数据处理
   - 设置参数：hg19 ⇒ hg18
3. 保存输出
   - 过滤结果：MAPPED VS. UNMAPPED

BED ⇔ GFF

1. 获取输入
2. 数据处理

BED ⇔ GFF

1. 获取输入
   - 输入文件：BED
2. 数据处理
   - BED ⇒ GFF
3. 保存输出

Galaxy 演示 | 格式转换 | BED ⇔ GFF

BED ⇔ GFF

1. 获取输入
   - 输入文件：BED
2. 数据处理
   1. BED ⇒ GFF
   2. GFF ⇒ BED
3. 保存输出

BED ⇔ GFF

1. 获取输入
   - 输入文件：BED
2. 数据处理
   1. BED ⇒ GFF
   2. GFF ⇒ BED
3. 保存输出

BED ⇔ GFF

1. 获取输入
   - 输入文件：BED
2. 数据处理
   1. BED ⇒ GFF
   2. GFF ⇒ BED
3. 保存输出
4. 查看结果，互相比较

Galaxy 演示 | 格式转换 | BED ⇔ GFF

BED ⇔ GFF

1. 获取输入
   - 输入文件：BED
2. 数据处理
   1. BED ⇒ GFF
   2. GFF ⇒ BED
3. 保存输出
   - 查看结果：互相比较

Galaxy 演示 | 格式转换 | BED ⇔ GFF

BED ⇔ GFF

1. 获取输入
   - 输入文件：BED
2. 数据处理
   1. BED ⇒ GFF
   2. GFF ⇒ BED
3. 保存输出
   - 查看结果：互相比较

Galaxy 演示 | 格式转换 | BED ⇔ GFF

BED ⇔ GFF

1. 获取输入
   - 输入文件：BED
2. 数据处理
   1. BED ⇒ GFF
   2. GFF ⇒ BED
3. 保存输出
   - 查看结果：互相比较

Galaxy 演示 | 逻辑运算

- Galaxy 中的“Operate on Genomic Intervals”工具集
- BEDTools: a powerful toolset for genome arithmetic
- BEDOPS: the fast, highly scalable and easily-parallelizable genome analysis toolkit

外显子 vs. SNP

1. 获取输入

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
3. 保存输出

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
3. 保存输出

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join
3. 保存输出

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join
3. 保存输出

Galaxy 演示 | 逻辑运算 | subtract & join

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join
3. 保存输出
4. 解析结果

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join
3. 保存输出
   - 解析结果

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join
3. 保存输出
   - 解析结果

外显子 vs. SNP

1. 获取输入
   - exon
   - SNP
2. 数据处理
   - subtract
   - join
3. 保存输出
   - 解析结果

问题

- 找到含有至少 N（2）个 SNP 的外显子。

输入

- 外显子数据（BED 格式）
- SNP 数据（BED 格式）

输出

- 满足要求的外显子信息（BED 格式）

问题

- 找到含有至少 N（2）个 SNP 的外显子。

输入

- 外显子数据（BED 格式）
- SNP 数据（BED 格式）

输出

- 满足要求的外显子信息（BED 格式）

问题

- 找到含有至少 N（2）个 SNP 的外显子。

输入

- 外显子数据（BED 格式）
- SNP 数据（BED 格式）

输出

- 满足要求的外显子信息（BED 格式）

外显子数据

1	chr1	10	20	exon1	0	+
2	chr1	30	40	exon2	0	+
3	chr1	50	60	exon3	0	-
4	chr1	65	75	exon4	0	+
5	chr1	85	95	exon5	0	-

SNP 数据

1	chr1	11	12	snp1	0	+
2	chr1	15	16	snp2	0	+
3	chr1	17	18	snp3	0	+
4	chr1	24	25	snp4	0	+
5	chr1	33	34	snp5	0	+
6	chr1	37	38	snp6	0	+
7	chr1	44	45	snp7	0	+
8	chr1	54	55	snp8	0	+
9	chr1	57	58	snp9	0	-

Finding exons with the highest number (≥ 10) of SNPs

Join-Group-Filter-Compare-Sort

- Input: Get exons, SNPs; UCSC Table Browser
- Join[Operate on Genomic Intervals]: Join exons with SNPs
- Group: Count the number of SNPs per exon
- Filter: Filter exons that have ten or more SNPs
- Compare two Datasets: Recover exon information
- Sort: Sort the start and end coordinates
- Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Group-Filter-Compare-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Compare two Datasets: Recover exon information
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Count-Filter-Cut-Sort

- Input: Get exons, SNPs; UCSC Table Browser
- Join[Operate on Genomic Intervals]: Join exons with SNPs
- Count: Count the number of SNPs per exon
- Filter: Filter exons that have ten or more SNPs
- Cut: Cut columns to recover BED format
- Sort: Sort the start and end coordinates
- Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding exons with the highest number ($\geq$ 10) of SNPs

Join-Count-Filter-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Count: Count the number of SNPs per exon
4. Filter: Filter exons that have ten or more SNPs
5. Cut: Cut columns to recover BED format
6. Sort: Sort the start and end coordinates
7. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

- Input: Get exons, SNPs; UCSC Table Browser
- Join[Operate on Genomic Intervals]: Join exons with SNPs
- Group: Count the number of SNPs per exon
- Sort: Sort exons by SNPs count
- Select first: Select top ten
- Join[Join two Datasets]: Recover exon information
- Cut: Cut columns to recover BED format
- Sort: Sort the start and end coordinates
- Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding 10 exons with the highest number of SNPs

Join-Group-Sort-SelectFirst-Join-Cut-Sort

1. Input: Get exons, SNPs; UCSC Table Browser
2. Join[Operate on Genomic Intervals]: Join exons with SNPs
3. Group: Count the number of SNPs per exon
4. Sort: Sort exons by SNPs count
5. Select first: Select top ten
6. Join[Join two Datasets]: Recover exon information
7. Cut: Cut columns to recover BED format
8. Sort: Sort the start and end coordinates
9. Visualize: Display data in genome browser

Finding exons with the highest number (≥ 10) of SNPs

BEDTools-shell-BEDTools

考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c -s |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

不考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

Finding exons with the highest number (≥ 10) of SNPs

BEDTools-shell-BEDTools

考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c -s |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

不考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

Finding exons with the highest number (≥ 10) of SNPs

BEDTools-shell-BEDTools

考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c -s |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

不考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

Finding exons with the highest number (≥ 10) of SNPs

BEDTools-shell-BEDTools

考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c -s |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

不考虑链性

```
bedtools intersect -a exon.bed -b snp.bed -c |
awk '{if($7>=10) print;}' | cut -f1-6 |
bedtools sort -i stdin
```

Workflow：create, modify, rerun, share

1. Save: Rename the history as “Exons and SNPs”
2. Workflow: Extract workflow from history
3. Modify: Open workflow editor and modify the parameter

Workflow：create, modify, rerun, share

1. Save: Rename the history as “Exons and SNPs”
2. Workflow: Extract workflow from history
3. Modify: Open workflow editor and modify the parameter
4. Rerun: Run workflow on whole genome data
5. Share: Share or publish workflow
6. Create: Create workflows from scratch (e.g. Find the 50 longest exons)

Workflow：create, modify, rerun, share

1. Save: Rename the history as “Exons and SNPs”
2. Workflow: Extract workflow from history
3. Modify: Open workflow editor and modify the parameter
4. Rerun: Run workflow on whole genome data
5. Share: Share or publish workflow
6. Create: Create workflows from scratch (e.g. Find the 50 longest exons)

Workflow：create, modify, rerun, share

1. Save: Rename the history as “Exons and SNPs”
2. Workflow: Extract workflow from history
3. Modify: Open workflow editor and modify the parameter
4. Rerun: Run workflow on whole genome data
5. Share: Share or publish workflow
6. Create: Create workflows from scratch (e.g. Find the 50 longest exons)

Workflow：create, modify, rerun, share

1. Save: Rename the history as “Exons and SNPs”
2. Workflow: Extract workflow from history
3. Modify: Open workflow editor and modify the parameter
4. Rerun: Run workflow on whole genome data
5. Share: Share or publish workflow
6. Create: Create workflows from scratch (e.g. Find the 50 longest exons)

Workflow：create, modify, rerun, share

1. Save: Rename the history as “Exons and SNPs”
2. Workflow: Extract workflow from history
3. Modify: Open workflow editor and modify the parameter
4. Rerun: Run workflow on whole genome data
5. Share: Share or publish workflow
6. Create: Create workflows from scratch (e.g. Find the 50 longest exons)

Workflow：create, modify, rerun, share

1. Save: Rename the history as “Exons and SNPs”
2. Workflow: Extract workflow from history
3. Modify: Open workflow editor and modify the parameter
4. Rerun: Run workflow on whole genome data
5. Share: Share or publish workflow
6. Create: Create workflows from scratch (e.g. Find the 50 longest exons)

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

三段论：“输入-加工-输出”

Input device(s)
Processor (CPU)
Output device(s)
Receive input
Process Information
Produce Output

PUTPUT
INPUT
OUTPUT

三段论：“输入-加工-输出”

获取输入

数据来源
文件格式
过滤数据
...

加工处理

选择工具
调整参数
记录步骤
...

解析输出

文件格式
解析图表
保存结果
...

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

总结与答疑

知识点——Galaxy 分析平台

- Galaxy——界面、学习、使用

技能——“输入-加工-输出”三段论

- 获取输入——格式、来源、过滤
- 数据处理——工具、版本、参数
- 解析输出——格式、注释、解析

教学提纲

1 引言
2 基因组组装版本
3 基因组坐标系统
4 基因组注释常用格式
5 文本文件与文本编辑器
6 基因组坐标的逻辑运算
7 总结与答疑
8 引言
9 变异位点的注释
10 基因集富集分析
11 序列标识
12 box plot
13 解析图表
14 总结与答疑
15 引言
16 Galaxy 分析平台
17 Galaxy 使用演示
18 数据处理三段论
19 总结与答疑
20 复习思考题

复习思考题

知识点

1. hg19 和 mm10 分别代表什么含义？hg19 是和 GRCh37 相对应，还是和 GRCm38 相对应？
2. 常见的基因组坐标系统是哪两种，举例进行说明。
3. 简述 BED 格式前 6 列的含义，能解释实际的 BED 记录。
4. 基于基因组坐标的常见逻辑运算模式有哪些，画图进行解释。
5. 简述序列标识的含义，能解释实际的序列标识图。

技能

1. 不同操作系统的换行符有何区别？
2. 以 SNP 的注释结果为例，论述如何解析一张表。
3. 以 box plot 为例，论述如何解析一张图。
4. 以坐标转换为例，论述“输入-加工-输出”的工作流程。

Powered by

TEX LATEX XETEX Beamer