生物信息学

天津医科大学
生物医学工程学院

2013-2014 学年上学期

基因组功能注释分析

伊现富（Yi Xianfu）

天津医科大学（TIJMU）
生物医学工程学院

2013 年 9 月

教学提纲

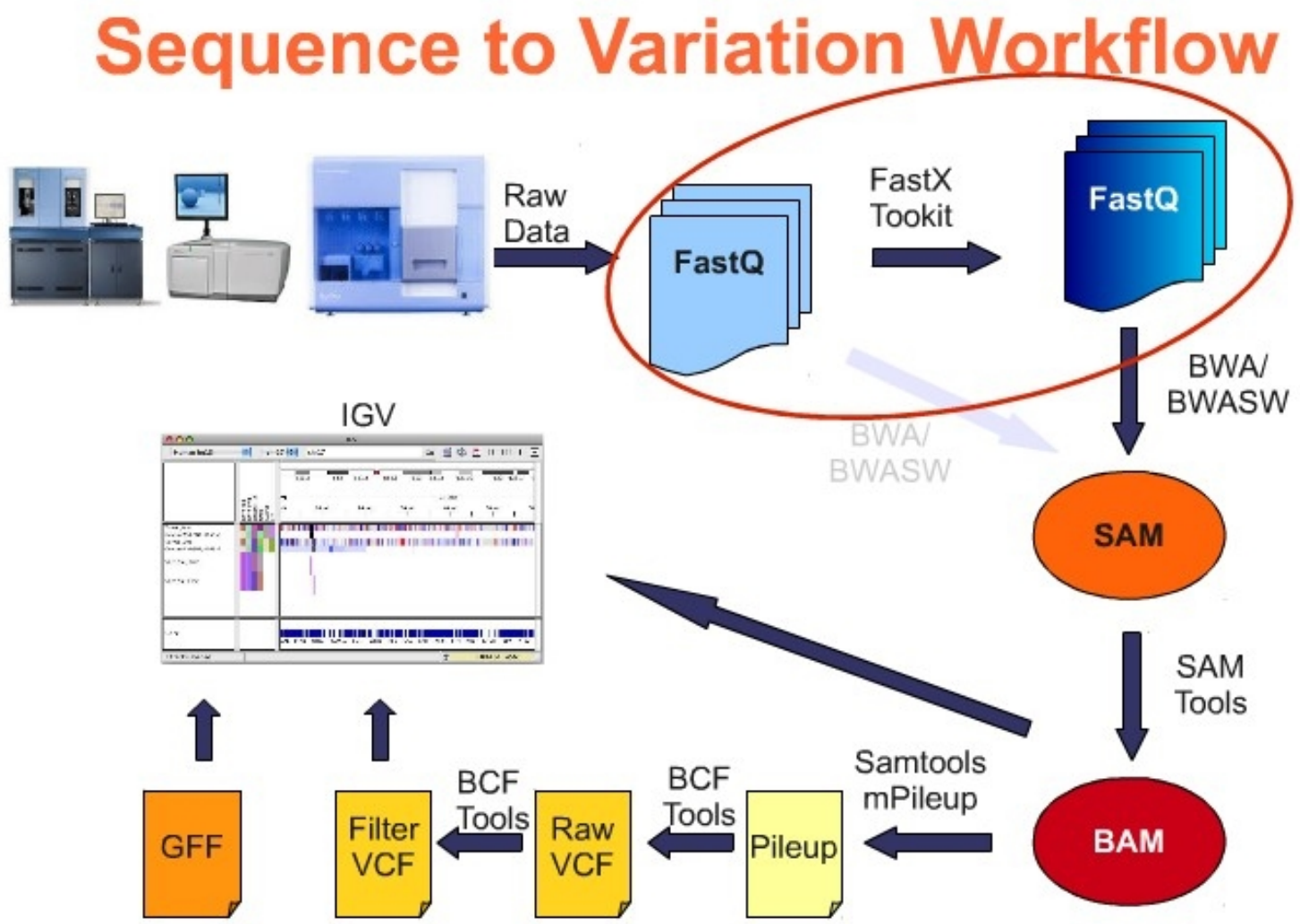
1. 引言
2. 基因组组装版本
3. 基因组坐标系统
4. 基因组注释常用格式
5. 基因组坐标的逻辑运算模式
6. 操作演示
7. 总结与答疑
8. 引言
9. 变异位点的注释
10. 基因集富集分析
11. 序列标识
12. Galaxy 分析平台
13. Galaxy 操作实例
14. 总结与答疑

教学提纲

基因组注释

基因组注释（genome annotation）

- 结构注释（structural annotation）
- 功能注释（functional annotation）


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

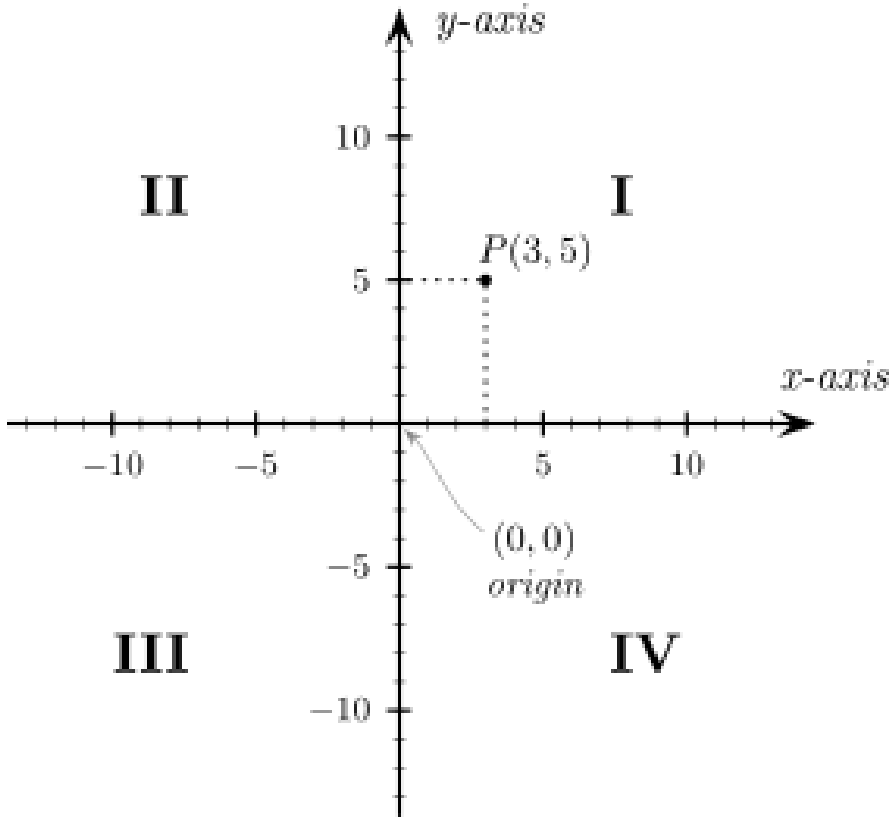
基因组组装版本 | XP VS. Win7

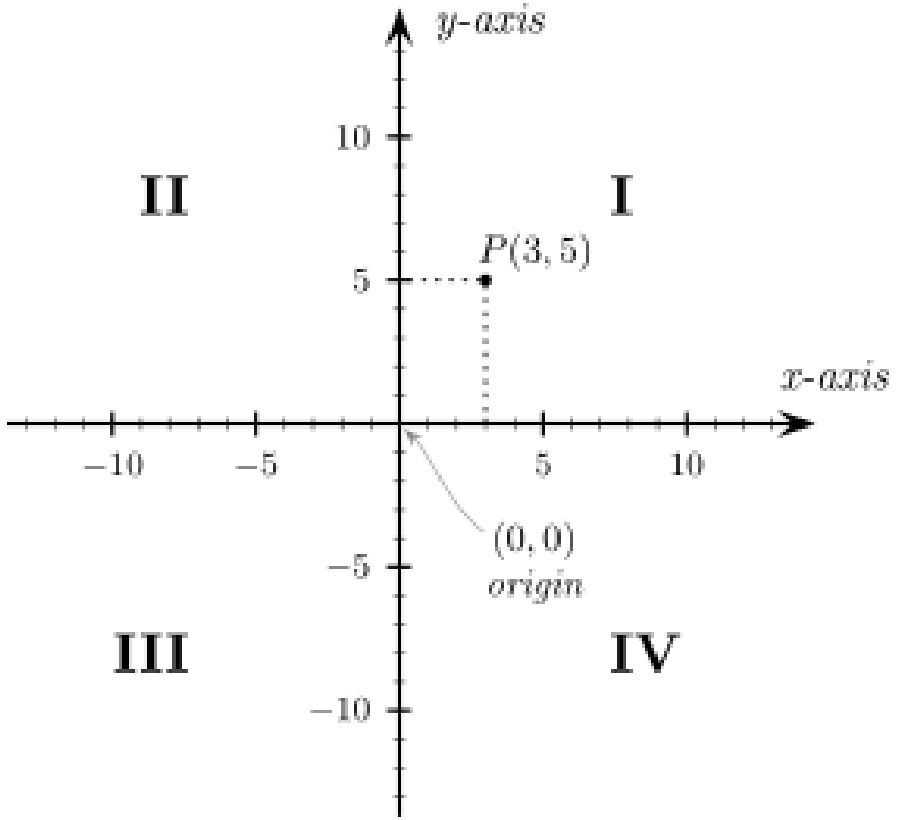
基因组组装版本 | 版本对照

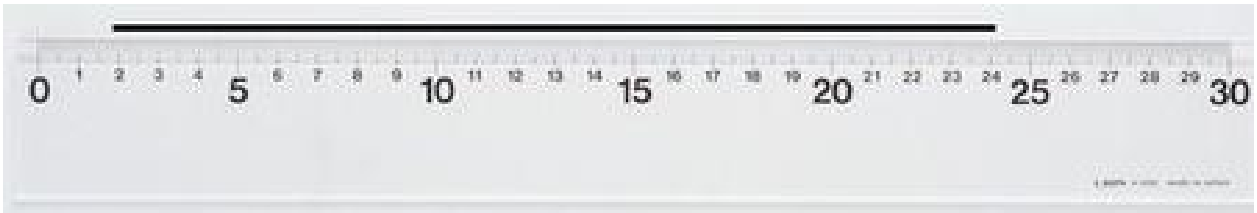
SPECIES	UCSC	DATE	NCBI
Human	hg19	Feb. 2009	Genome Reference Consortium GRCh37
	hg18	Mar. 2006	NCBI Build 36.1
	hg17	May 2004	NCBI Build 35
	hg16	Jul. 2003	NCBI Build 34
	hg15	Apr. 2003	NCBI Build 33
Mouse	mm10	Dec. 2011	Genome Reference Consortium GRCm38
	mm9	Jul. 2007	NCBI Build 37
	mm8	Feb. 2006	NCBI Build 36
	mm7	Aug. 2005	NCBI Build 35
	mm6	Mar. 2005	NCBI Build 34

教学提纲

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


0
5
10
15
20
25
30

坐标系统

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

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

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

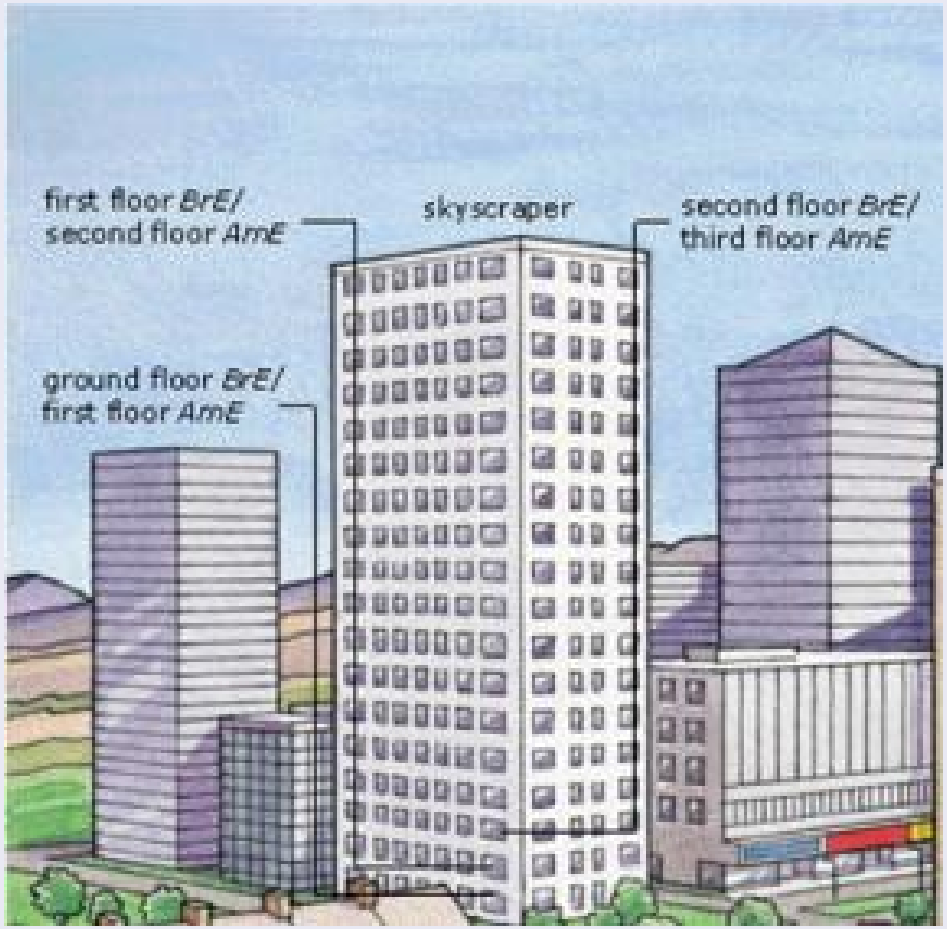
TG 的坐标

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

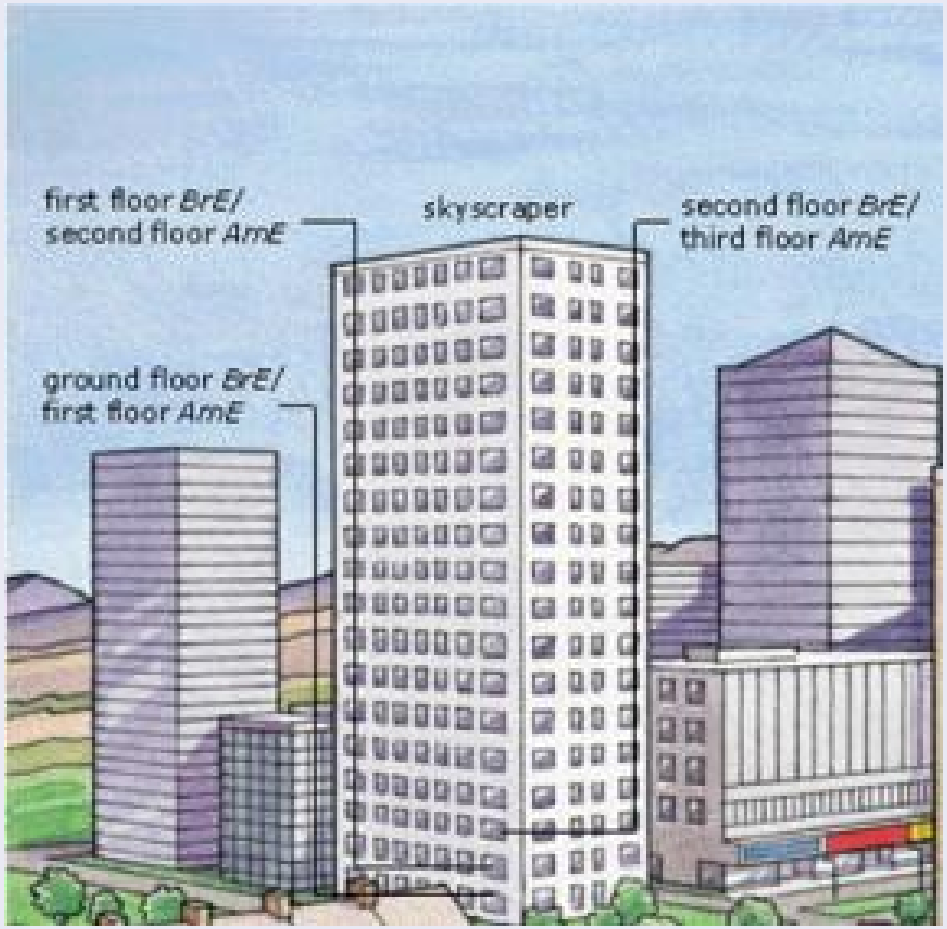
实例

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

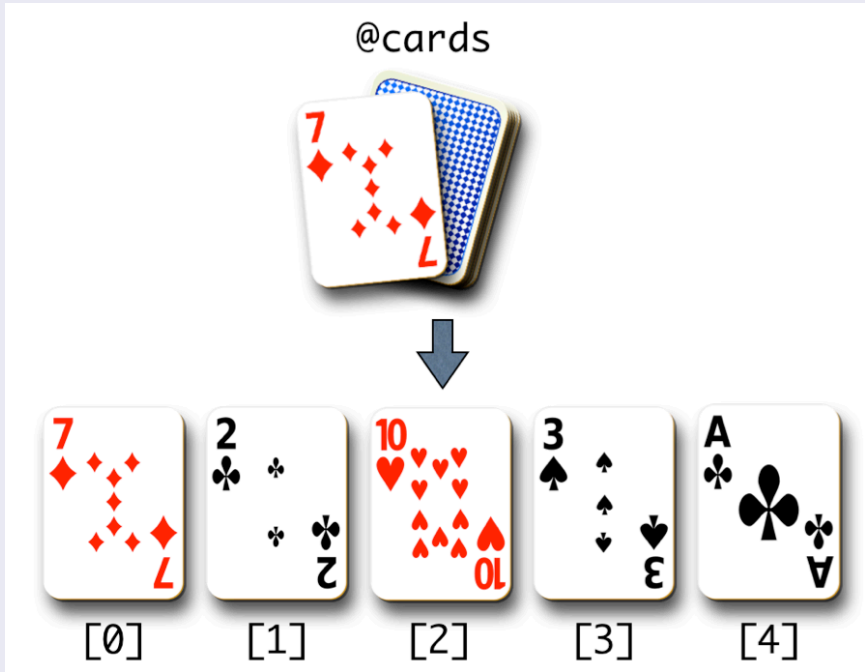
first floor


first floor BrE/
second floor AmE
skyscraper
second floor BrE/
third floor AmE
ground floor BrE/
first floor AmE

first floor



数组

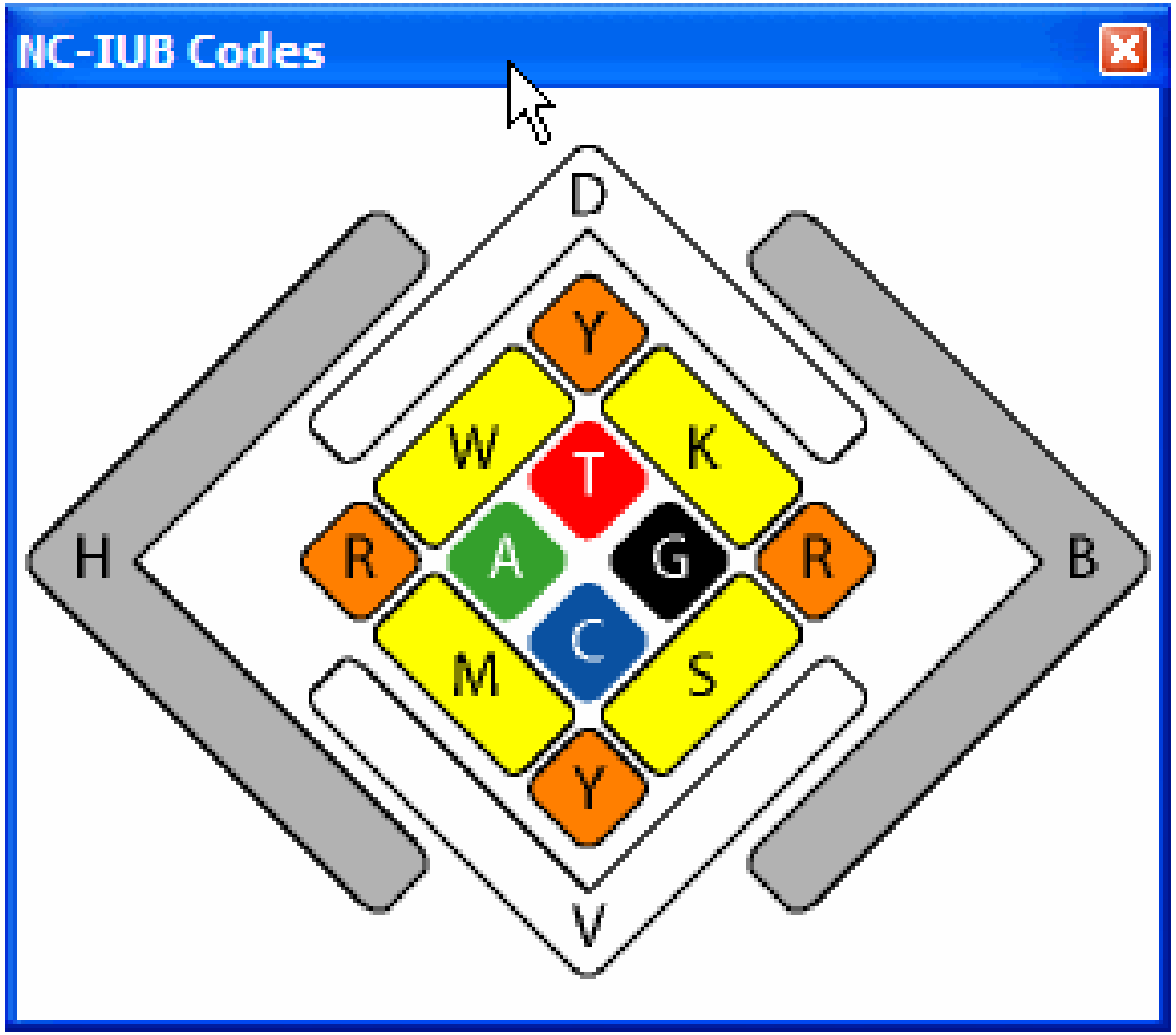


教学提纲

格式 | FASTA

```
>gi|142864|gb|M10040.1|BACDNAE B.subtilis dnaE gene encoding DNA primase, complete cds
GTACGACGGAGTGTTATAAGATGGGAAATCGGATACCAGATGAAATTGTGGATCAGGTGCAAAAGTCGGC
AGATATCGTTGAAGTCATAGGTGATTATGTTCAATTAAAGAAGCAAGGCCGAAACTACTTTGGACTCTGT
CCTTTTCATGGAGAAAGCACACCTTCGTTTTCCGTATCGCCCGACAAACAGATTTTTCATTGCTTTGGCT
GCGGAGCGGGCGGCAATGTTTTCTCTTTTTTAAGGCAGATGGAAGGCTATTCTTTTGCCGAGTCGGTTTC
TCACCTTGCTGACAAATACCAAATTGATTTTCCAGATGATATAACAGTCCATTCCGGAGCCCGGCCAGAG
TCTTCTGGAGAACAAAAAATGGCTGAGGCACATGAGCTCCTGAAGAAATTTTACCATCATTTGTTAATAA
ATACAAAAGAAGGTCAAGAGGCACTGGATTATCTGCTTTCTAGGGGCTTTACGAAAGAGCTGATTAATGA
ATTTCAGATTGGCTATGCTCTTGATTCTTGGGACTTTATCACGAAATTCCTTGTAAAGAGGGGATTTAGT
GAGGCGCAAATGGAAAAAGCGGGTCTCCTGATCAGACGCGAAGACGGAAGCGGATATTTCGACCGCTTCA
GAAACCGTGTCATGTTTCCGATCCATGATCATCACGGGGCTGTTGTTGCTTTCTCAGGCAGGGCTCTTGG
```

格式 | IUB/IUPAC 核酸



格式 | 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 | BED

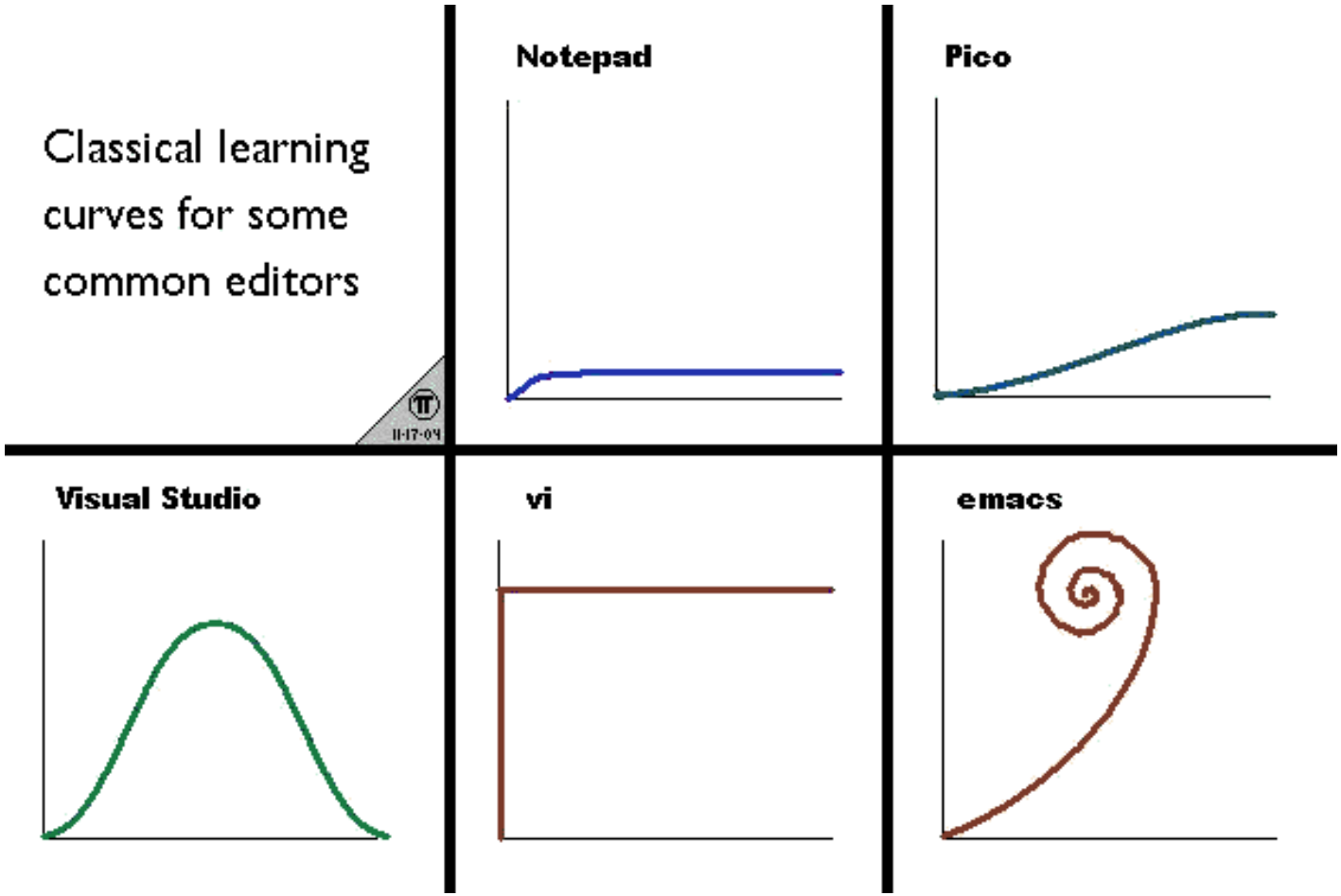
```
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  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 | GFF

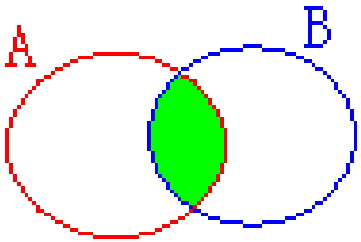
```
##gff-version 3
ctg123 . operon          1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Parent=operon001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  Parent=mrna0001
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

```
##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
```

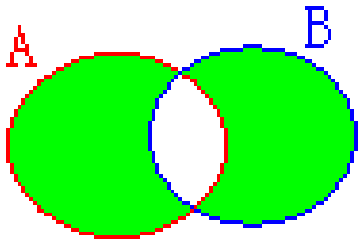

Classical learning curves for some common editors
Notepad
Pico
Visual Studio
vi
emacs

教学提纲

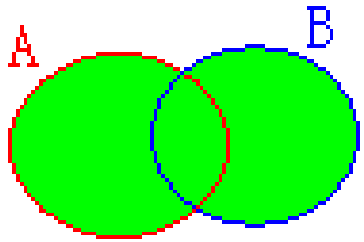


求同

交集

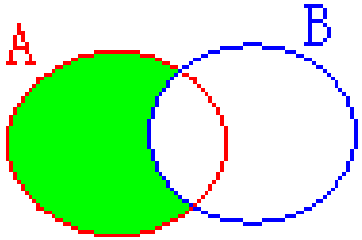


求异



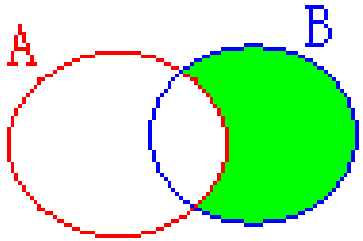
相加

并集

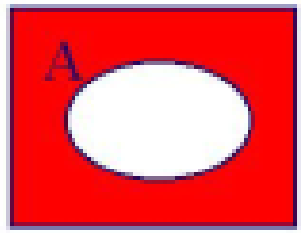


相减 A-B

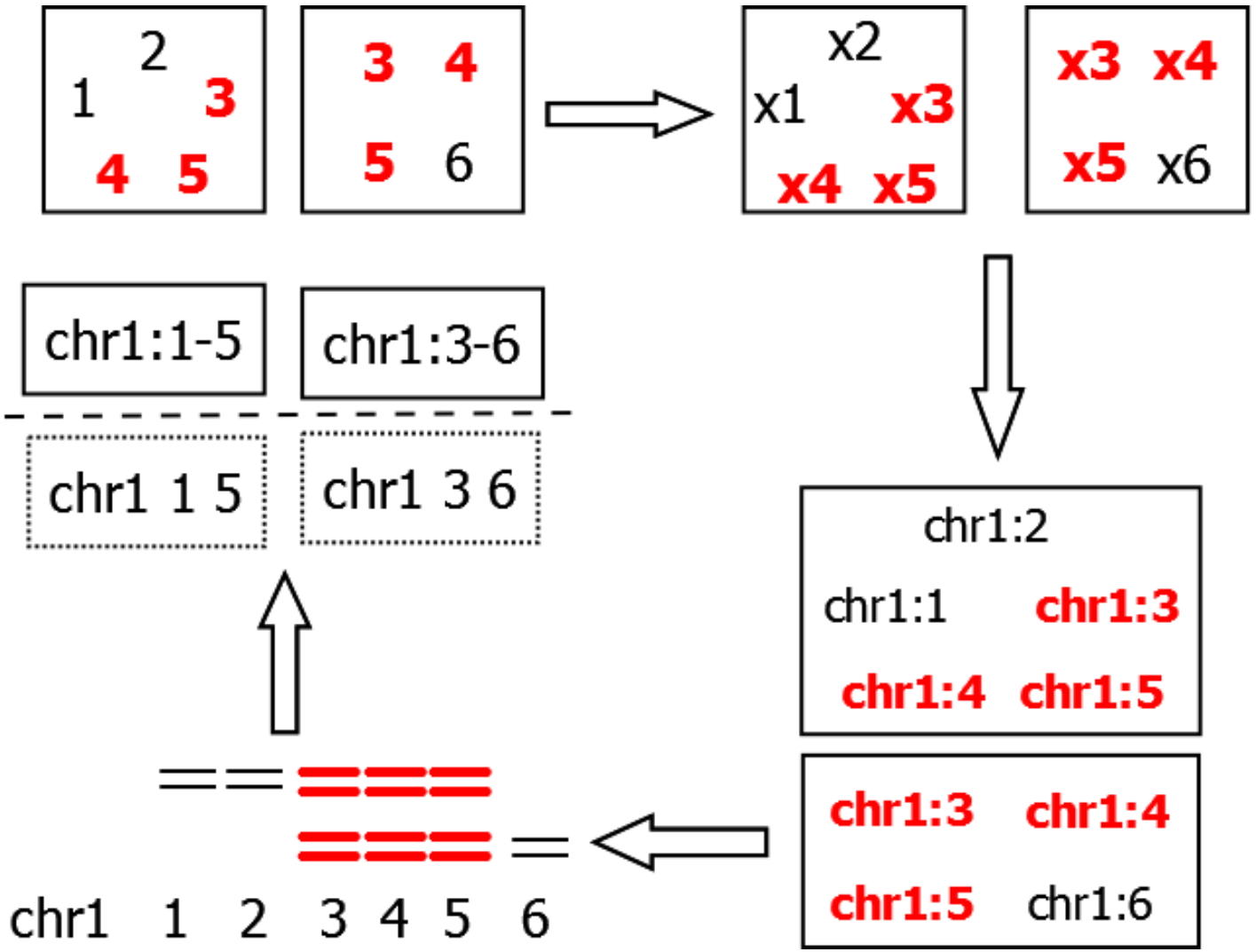
差集

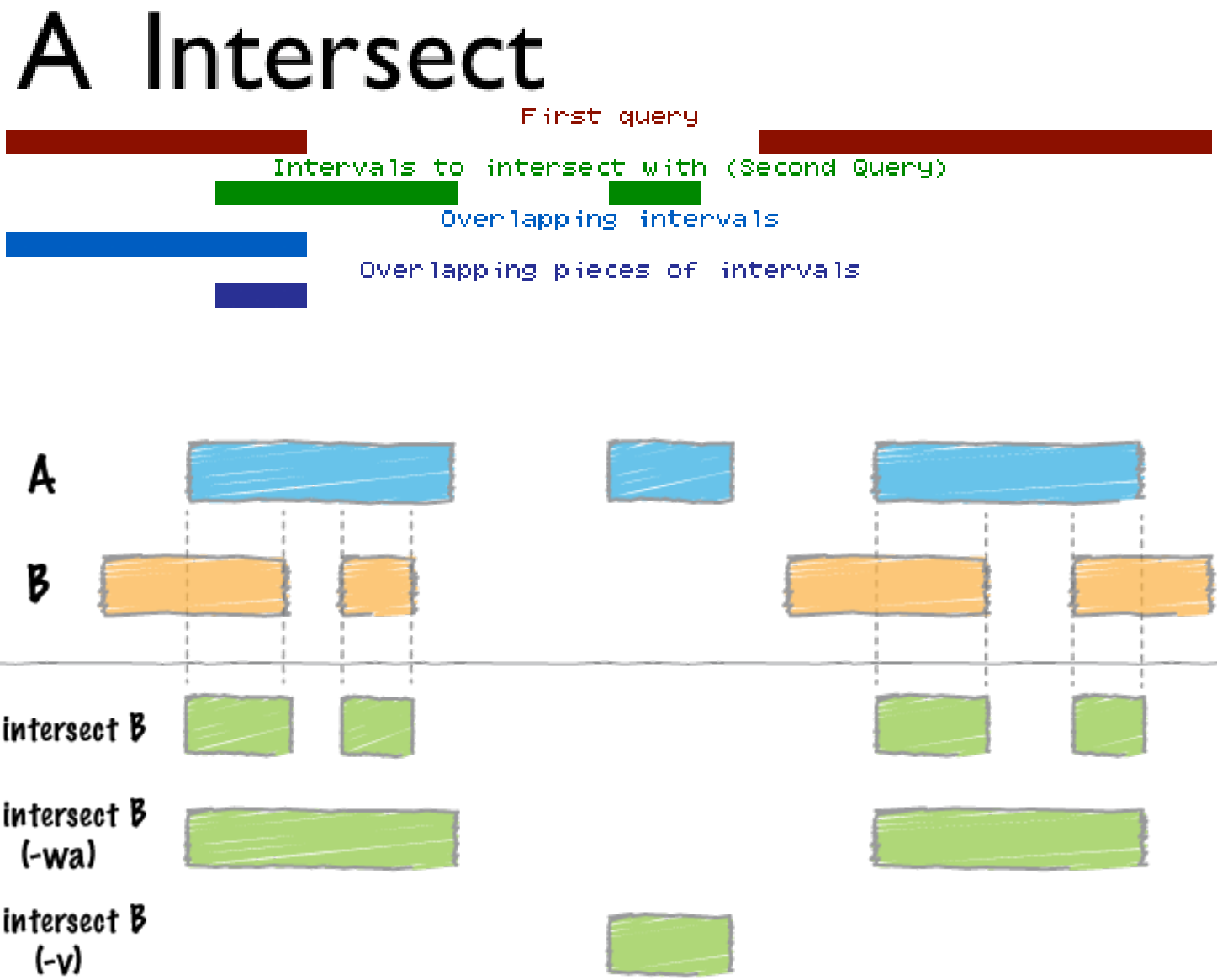


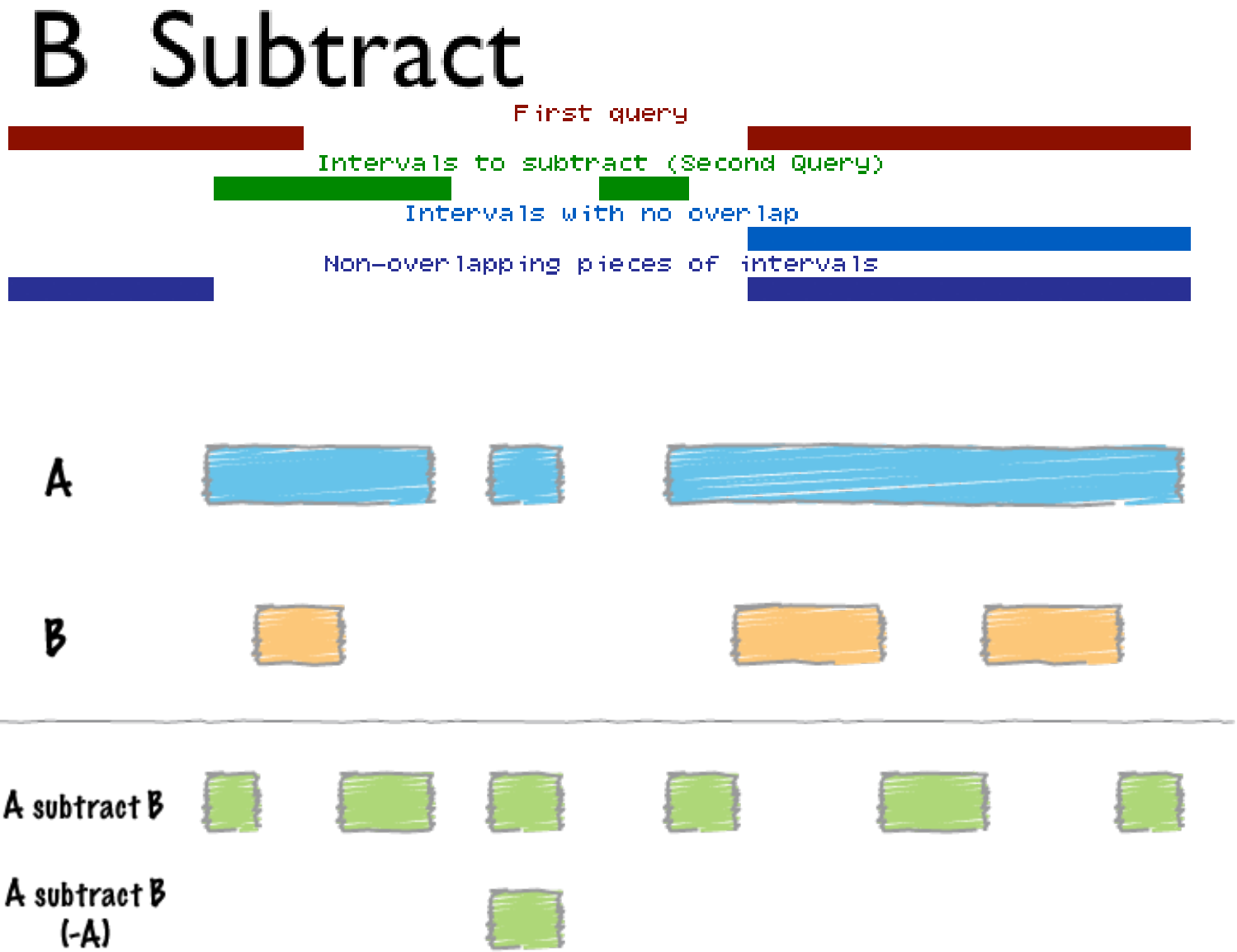
相减 B-A

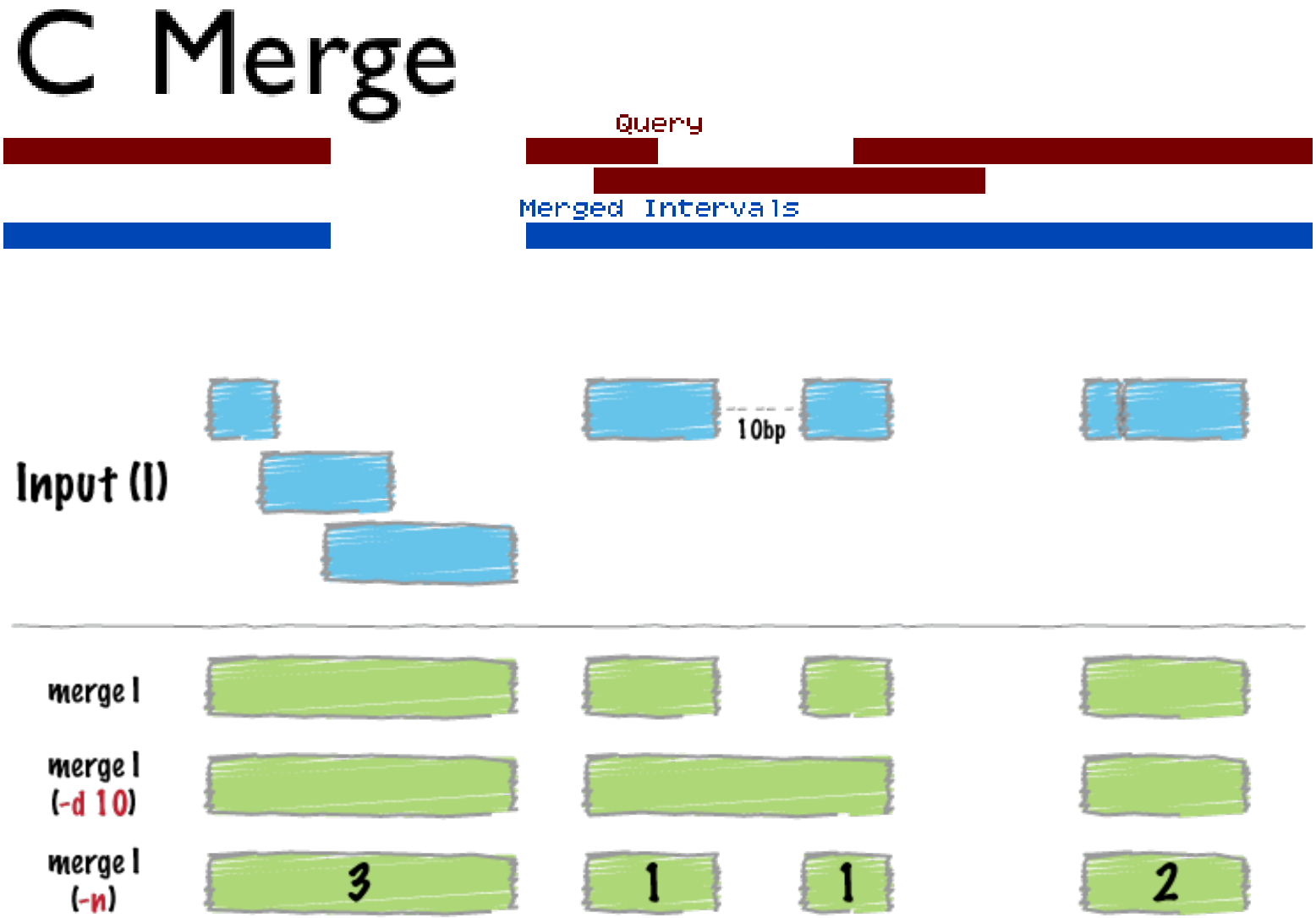


补集

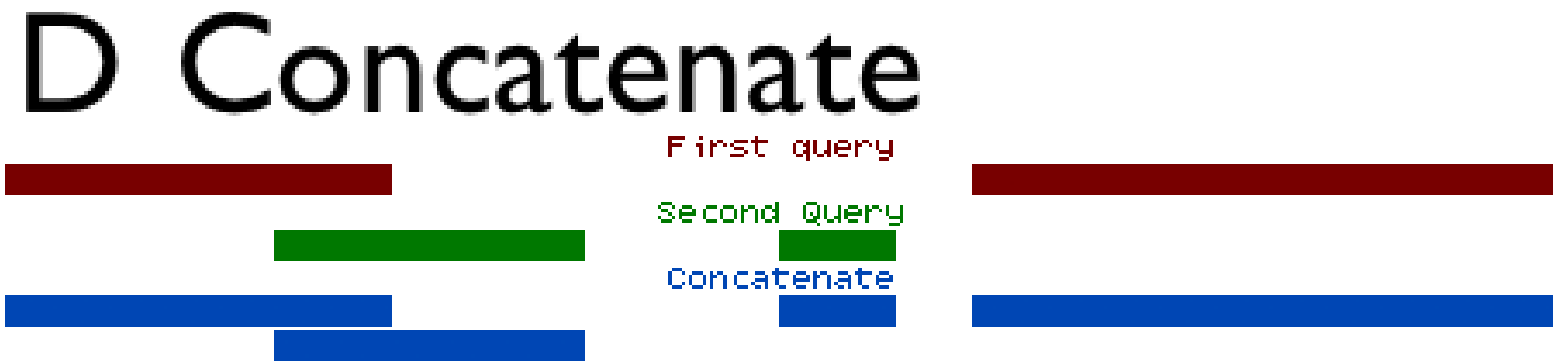

2
1 3
4 5
3 4
5 6
x2
x1 x3
x4 x5
x3 x4
x5 x6
chr1:1-5
chr1:3-6
chr1 1 5
chr1 3 6
chr1:2
chr1:1 chr1:3
chr1:4 chr1:5
chr1:3 chr1:4
chr1:5 chr1:6
chr1 1 2 3 4 5 6

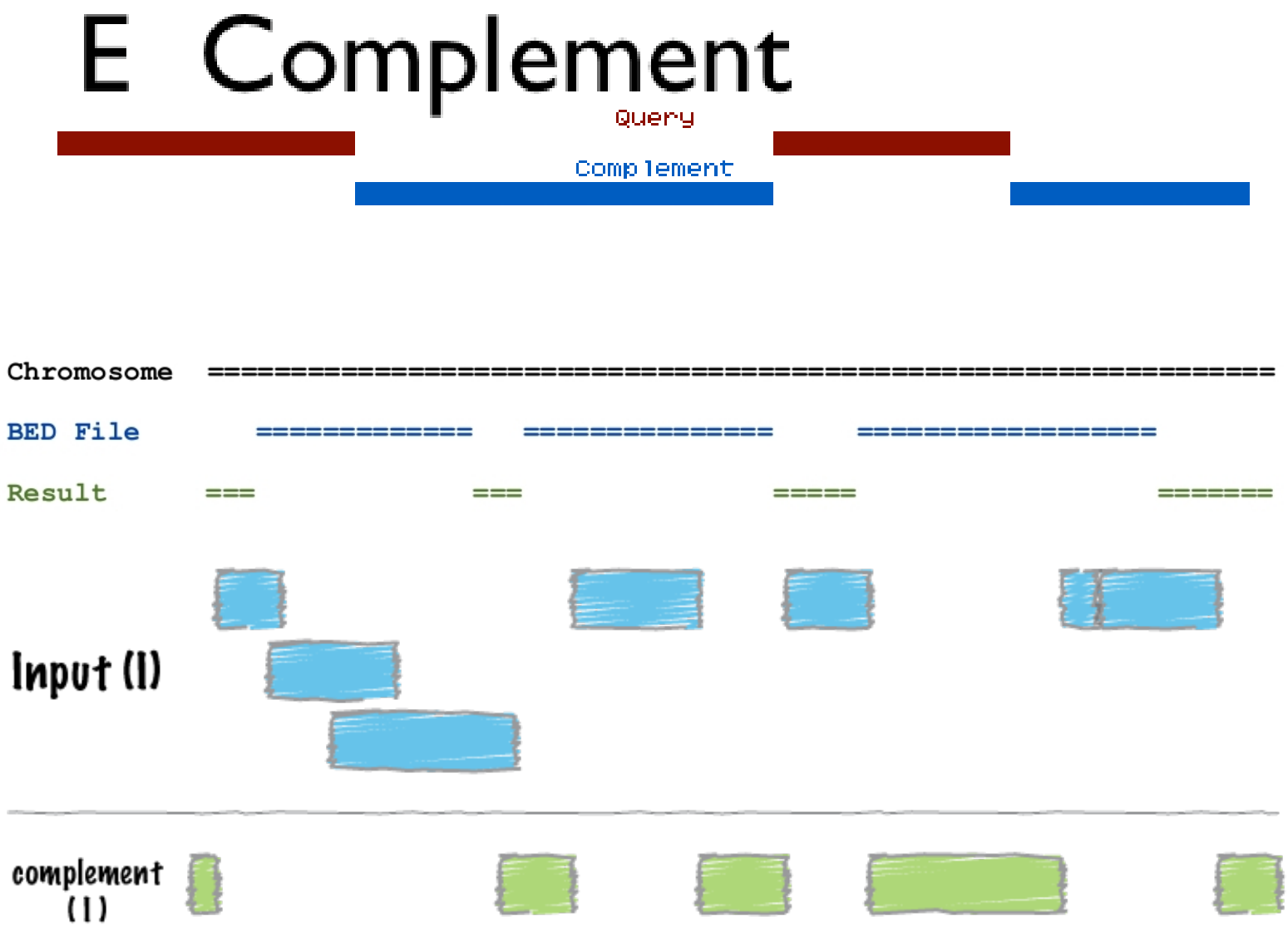

A Intersect
First query
Intervals to intersect with (Second Query)
Overlapping intervals
Overlapping pieces of intervals
A
B
A intersect B
A intersect B
(-wa)
A intersect B
(-v)

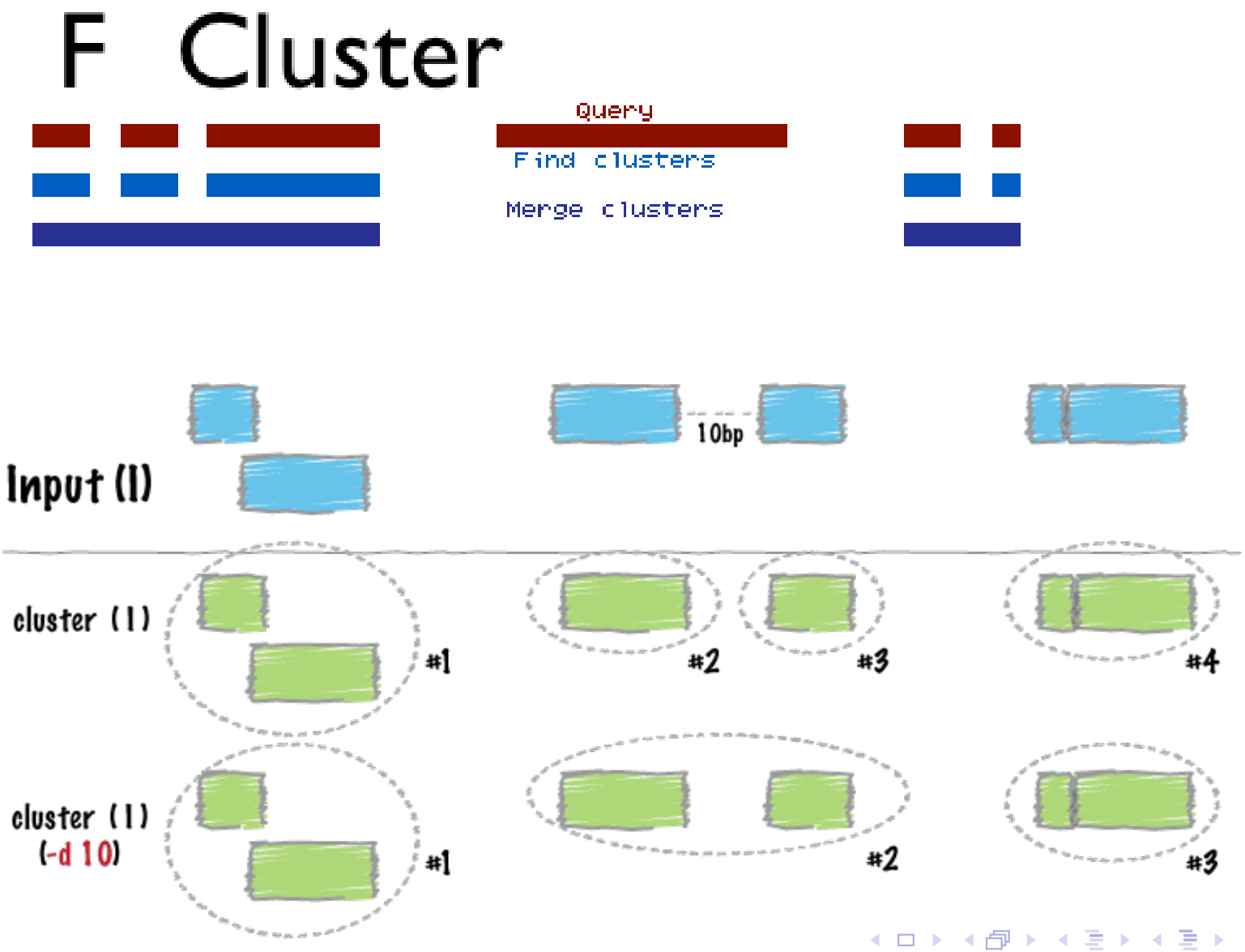

B Subtract
First query
Intervals to subtract (Second Query)
Intervals with no overlap
Non-overlapping pieces of intervals
A
B
A subtract B
A subtract B
(-A)


C Merge
Query
Merged Intervals
Input (I)
10bp
merge I
merge I
(-d 10)
merge I
(-n)
3
1
1
2

逻辑运算 | concatenate




E Complement
Query
Complement
Chromosome
BED File
Result
Input (I)
complement
(I)


F Cluster
Query
Find clusters
Merge clusters
Input (I)
10bp
cluster (I)
#1
#2
#3
#4
cluster (I)
(-d 10)
#1
#2
#3

逻辑运算 | join

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

Input

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 ".")

(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 ".")

(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 ".")

(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 ".")

教学提纲

操作演示 | liftOver

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

操作演示 | liftOver

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

操作演示 | liftOver

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

操作演示 | liftOver

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

操作演示 | liftOver

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

操作演示 | liftOver

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

操作演示 | BED ⇔ GFF

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

操作演示 | 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

操作演示 | 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

操作演示 | 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：直接比较

操作演示 | 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

操作演示 | 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

操作演示 | 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

操作演示 | subtract & join

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

操作演示 | subtract & join

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

操作演示 | subtract & join

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

操作演示 | subtract & join

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

操作演示 | subtract & join

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

操作演示 | subtract & join

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

操作演示 | subtract & join

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

操作演示 | subtract & join

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

教学提纲

总结与答疑

知识点

- 基因组组装版本的对应关系
- 两种坐标系统——0-based 和 1-based
- 四种常用格式——FASTA，BED，GFF，VCF
- 坐标的逻辑运算模式
- 坐标转换、格式转换、逻辑运算的工具

技能

- “输入 -加工 -输出”三段论
- 获取输入
- 数据处理
- 解析输出

教学提纲

引言

前期准备工作

- 组装版本
- 坐标转换
- 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

分析平台

- Galaxy

引言

前期准备工作

- 组装版本
- 坐标转换
- 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

分析平台

- Galaxy

引言

前期准备工作

- 组装版本
- 坐标转换
- 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

分析平台

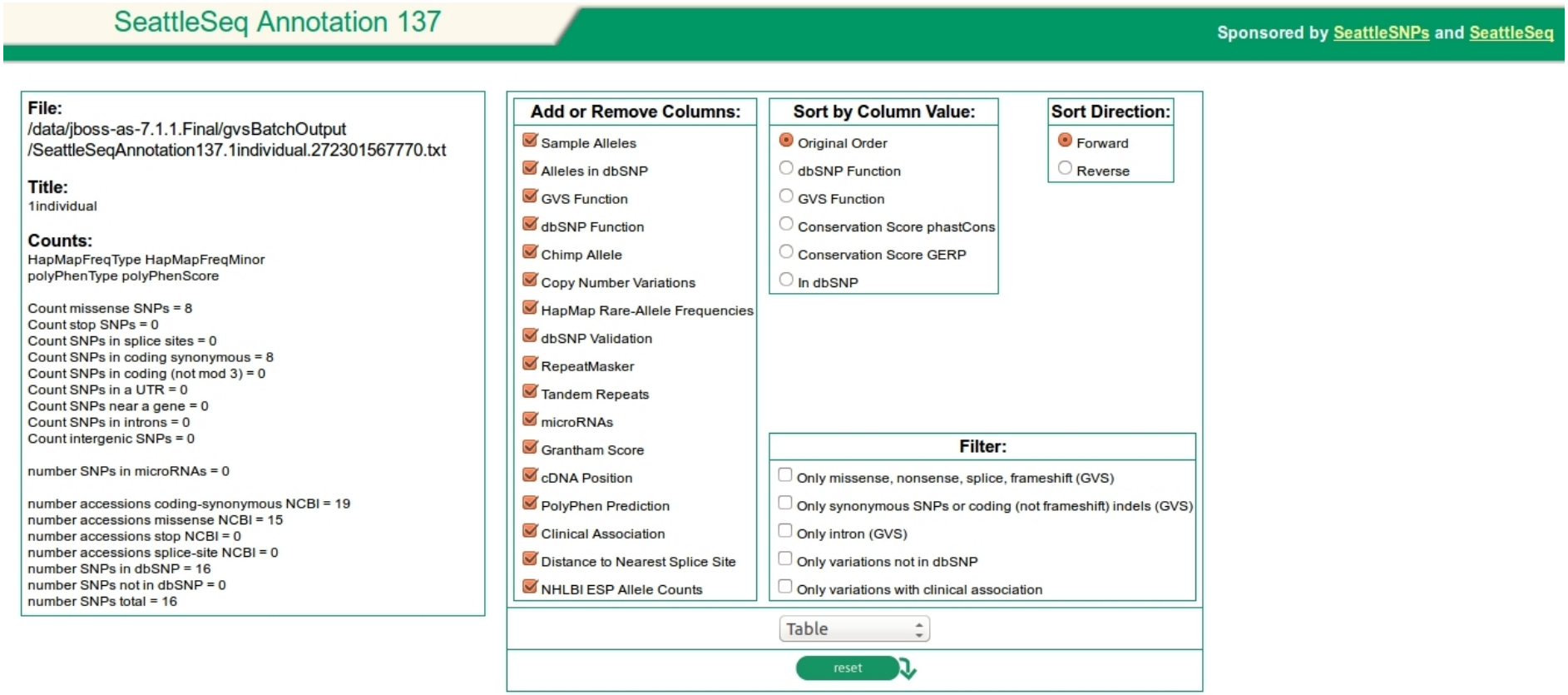
- Galaxy

教学提纲

变异位点的注释 | 注释工具

- SeattleSeq Annotation、variant tools、SnpEff
- SIFT、PolyPhen-2、SNPs3D
- PROVEAN

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



SeattleSeq Annotation 137

Sponsored by SeattleSNPs and SeattleSeq

File:
/data/jboss-as-7.1.1.Final/gvsBatchOutput
/SeattleSeqAnnotation137.1individual.272301567770.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	proteinPosition	cDNAPosition	polyPhen
dbSNP_130	10	1126383	A	R	A/G	A/G	NM_014023.3	coding-synonymous	synonymous-codon	73578536	none	121/495	363	unknown
dbSNP_86	10	3150973	C	Y	C/T	C/T	NM_001242339.1	coding-synonymous	synonymous-codon	1132173	none	309/777	927	unknown

变异位点的注释 | 结果解析 | SIFT

SIFT: PREDICTIONS

Homo sapiens GRCh37 Ensembl 63

User input	Coordinates	Codons	Transcript ID	Protein ID	Substitution	Region	dbSNP ID	SNP Type	Prediction	SIFT Score	Median Information Content	# Seqs at position
1,100382265,1,C/G	1,100382265,1,C/G	CGA-gGA	ENST00000294724	ENSP00000294724	R1487G	EXON CDS	rs12118058:G	Nonsynonymous	TOLERATED	0.46	2.45	74
1,100380997,1,A/G	1,100380997,1,A/G	GAA-GgA	ENST00000294724	ENSP00000294724	E1405G	EXON CDS	rs28730708:G	Nonsynonymous	DAMAGING	0.01	2.45	74
1,100382265,1,C/A	1,100382265,1,C/A	CGA-aGA	ENST00000294724	ENSP00000294724	R1487R	EXON CDS	rs12118058:G	Synonymous	TOLERATED	0.64	2.45	74
22,30163533,1,A/C	22,30163533,1,A/C	GAG-GcG	ENST00000330029	ENSP00000332887	E49A	EXON CDS	novel	Nonsynonymous	DAMAGING	0.02	2.57	97
20,50071099,1,G/T	20,50071099,1,G/T	ACT-AaT	ENST00000371564	ENSP00000360619	T612N	EXON CDS	rs6067785:T	Nonsynonymous	DAMAGING	0	2.81	122
2,230633386,-1,C/T	2,230633386,1,G/A	CAG-tAG	ENST00000283943	ENSP00000283943	Q1910*	EXON CDS	rs1803846:A	Nonsynonymous	N/A	N/A	N/A	N/A
2,230312220,-1,C/T	2,230312220,1,G/A	CCC-CtC	ENST00000341772	ENSP00000345229	P433L	EXON CDS	rs17853365:A	Nonsynonymous	DAMAGING	0.02	2.38	53
4,30723053,1,G/T	4,30723053,1,G/T	AGG-AGt	ENST00000333135	ENSP00000330302	R3S	EXON CDS	rs2631567:T	Nonsynonymous	TOLERATED	0.16	3.12	68
1,100624830,-1,T/A	Reference nucleotide not matched	-					N/A	N/A	Not scored	N/A	N/A	N/A
X,12905093,1,G/A	Reference nucleotide not matched	-					N/A	N/A	Not scored	N/A	N/A	N/A

教学提纲

富集分析 | 数据库与分析工具

- GO
- KEGG
- DAVID

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

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

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

Sublist	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_CC_FAT	extracellular space	RT		19	12.3	9.4E-5	6.5E-3
☐	GOTERM_MF_FAT	oxygen binding	RT		6	3.9	3.8E-5	1.4E-2
☐	GOTERM_MF_FAT	tetrapyrrole binding	RT		8	5.2	1.5E-4	1.9E-2
☐	GOTERM_MF_FAT	heme binding	RT		8	5.2	1.0E-4	1.9E-2
☐	GOTERM_MF_FAT	iron ion binding	RT		11	7.1	4.3E-4	3.9E-2
☐	GOTERM_BP_FAT	response to bacterium	RT		10	6.5	1.4E-4	9.1E-2
☐	GOTERM_BP_FAT	defense response	RT		18	11.6	1.3E-4	1.7E-1
☐	GOTERM_CC_FAT	hemoglobin complex	RT		3	1.9	5.7E-3	2.6E-1
☐	GOTERM_CC_FAT	cell fraction	RT		20	12.9	7.5E-3	2.7E-1
☐	GOTERM_BP_FAT	defense response to bacterium	RT		7	4.5	8.9E-4	3.4E-1
☐	GOTERM_MF_FAT	oxygen transporter activity	RT		3	1.9	5.8E-3	3.5E-1
☐	GOTERM_BP_FAT	response to drug	RT		9	5.8	1.5E-3	4.0E-1
☐	GOTERM_CC_FAT	cytosol	RT		22	14.2	1.5E-2	4.2E-1
☐	GOTERM_MF_FAT	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	RT		5	3.2	1.7E-2	5.1E-1
☐	GOTERM_MF_FAT	sodium ion binding	RT		5	3.2	2.1E-2	5.4E-1
☐	GOTERM_MF_FAT	ATPase activity, coupled to movement of substances	RT		5	3.2	1.7E-2	5.4E-1

富集分析 | DAVID | 工具选择

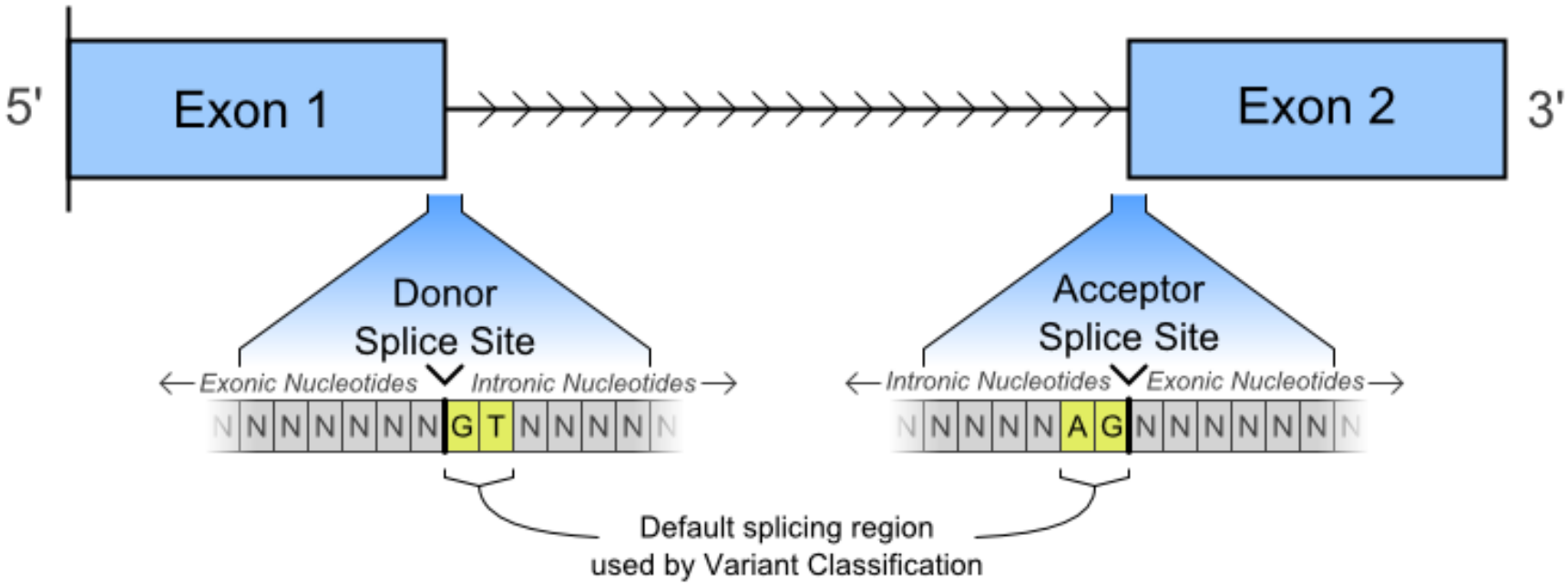
Convert gene IDs from one type to another	■			□		
Diagnose and fix problems of gene IDs	■	□				□
Explore gene names in batch		■	□			□
Discover enriched functionally related gene groups			■	□	□	
Display relationship of many-genes-to-many-terms on 2D view.			■		■	
Initial glance of major biological functions 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						■

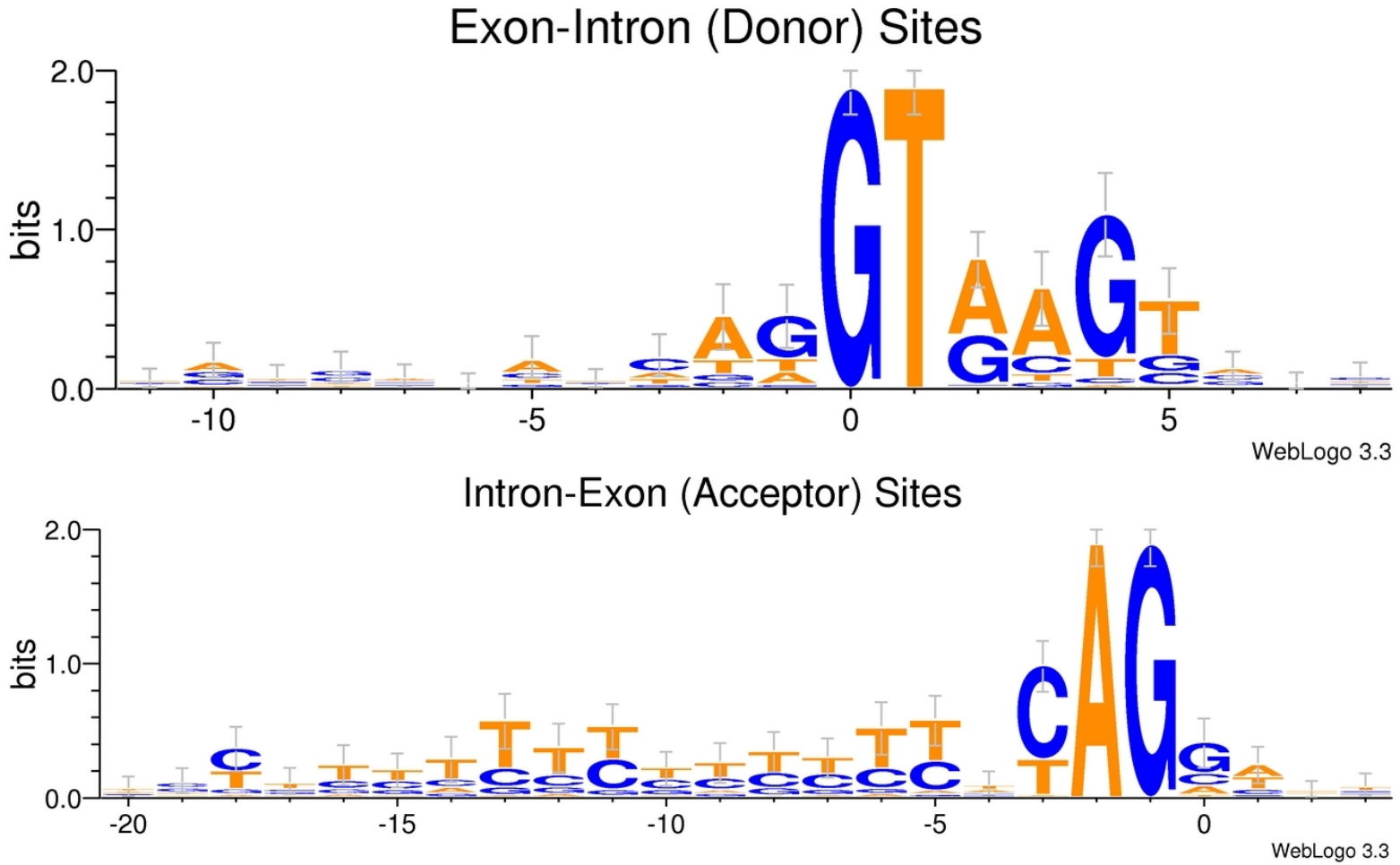
教学提纲

序列标识

序列标识（sequence logo）

- 横轴：位置
- 纵轴：比特
- 总高度：保守性
- 相对高度：相对频率
- 制作工具：WebLogo, enoLOGOS


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


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

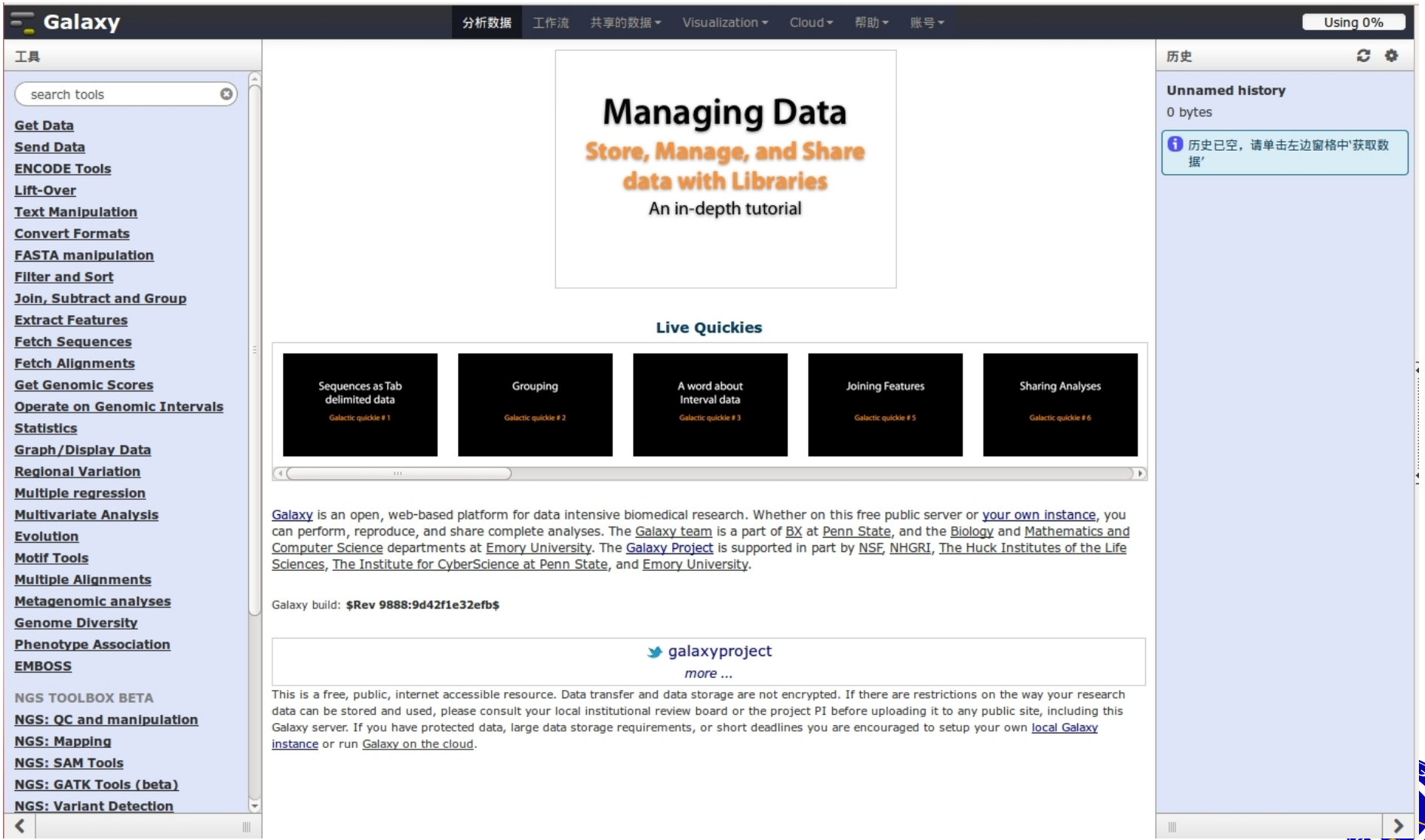
Donor Sites(%)		Acceptor Site(%)	
GT	98.797	AG	99.714
GC	0.920	AC	0.120
AT	0.143	TG	0.032
GA	0.028	AT	0.024
GG	0.025	GG	0.022
CT	0.018	AA	0.019
TT	0.016	CG	0.010
CC	0.011	CC	0.010
TG	0.007	TT	0.009
AG	0.007	CT	0.008
TA	0.006	CA	0.008
AC	0.006	GC	0.007
CA	0.006	TA	0.006
TC	0.004	TC	0.004
AA	0.004	GT	0.004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

教学提纲

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
分析数据
工作流
共享的数据
Visualization
Cloud
帮助
账号
Using 0%
工具
search tools
Get Data
Send Data
ENCODE Tools
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
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity
Phenotype Association
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Variant Detection
Managing Data
Store, Manage, and Share data with Libraries
An in-depth tutorial
Live Quickies
Sequences as Tab delimited data
Galactic quickie # 1
Grouping
Galactic quickie # 2
A word about Interval data
Galactic quickie # 3
Joining Features
Galactic quickie # 5
Sharing Analyses
Galactic quickie # 6
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
Galaxy build: $Rev 9888:9d42f1e32efb$
galaxyproject
more ...
This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.
历史
Unnamed history
0 bytes
历史已空，请单击左边窗格中'获取数据'

教学提纲

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

Finding exons with the highest number of SNPs

1. Input: exons, snps; UCSC Table Browser
2. Join[Genomic Operations Join]: identify those exons that contain SNPs
3. Group: obtain the number of SNPs within each exon
4. Sort: sort exon by SNP count
5. Filter: filter exons that have ten or more SNPs
6. Join[Join two Queries]: restore genomic location for exons containing ten or more SNPs
7. Visualize: visualize dataset in UCSC Genome Browser

教学提纲

总结与答疑

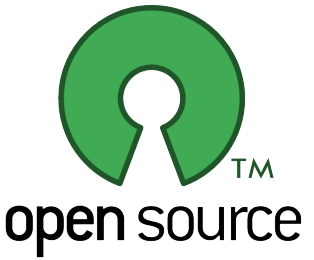
知识点

- 变异位点注释的用途及注释工具
- 基因集富集分析的功能及分析工具
- 序列标识的含义与制作工具
- Galaxy 分析平台的使用方法

技能

- 查找工具使用的 protocol
- 学习新工具的方法与步骤
- 数据处理流程的保存与共享

Powered by



TEX LATEX XETEX Beamer