基因组功能注释分析基础（集体备课）

伊现富（Yi Xianfu）

天津医科大学（TIJMU）
生物医学工程学院

2014 年 3 月 25 日

教学提纲

1. 指导思想
2. 引言与导入
3. 基因组组装版本
4. 基因组坐标系统
5. 基因组注释常用格式
6. 文本文件与文本编辑器
7. 基因组坐标的逻辑运算
8. 操作演示
9. 总结与答疑
10. 复习思考题

教学提纲

基础知识

- 基因组组装版本与坐标系统
- 基因组注释常用格式 $\Longrightarrow$ 文本文件与文本编辑器
- 基因组坐标的逻辑运算

功能注释

- 变异位点的注释
- 基因集富集分析
- 序列标识（sequence logo）
- 盒形图（box plot）

注释平台

- Galaxy 分析平台

基础知识

- 基因组组装版本与坐标系统
- 基因组注释常用格式 $\Longrightarrow$ 文本文件与文本编辑器
- 基因组坐标的逻辑运算

功能注释

- 变异位点的注释
- 基因集富集分析
- 序列标识（sequence logo）
- 盒形图（box plot）

注释平台

- Galaxy 分析平台

基础知识

- 基因组组装版本与坐标系统
- 基因组注释常用格式 $\Longrightarrow$ 文本文件与文本编辑器
- 基因组坐标的逻辑运算

功能注释

- 变异位点的注释
- 基因集富集分析
- 序列标识（sequence logo）
- 盒形图（box plot）

注释平台

- Galaxy 分析平台

指导思想 | 备课

内容

- 引言与导入
- 基因组组装版本
- 基因组坐标系统
- 基因组注释常用格式
- 文本文件与文本编辑器
- 基因组坐标的逻辑运算
- 总结与答疑

进程（分钟）

- 5
- 5~10
- 15
- 20~25
- 10
- 35
- 5

学生

- 知识全新
- 内容抽象

教法

- 类比举例（陌生 ⇒ 熟悉）
- 详略得当（讲授 ⇒ 自学）
- 扩展发挥（知识 ⇒ 兴趣）

内容

- 引言与导入
- 基因组组装版本
- 基因组坐标系统
- 基因组注释常用格式
- 文本文件与文本编辑器
- 基因组坐标的逻辑运算
- 总结与答疑

进程（分钟）

- 5
- 5~10
- 15
- 20~25
- 10
- 35
- 5

学生

- 知识全新
- 内容抽象

教法

- 类比举例（陌生 $\Rightarrow$ 熟悉）
- 详略得当（讲授 $\Rightarrow$ 自学）
- 扩展发挥（知识 $\Rightarrow$ 兴趣）

教学提纲

基因组注释（genome annotation）

- 结构注释（structural annotation）⇐ 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（functional annotation）⇐ 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

- 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

教学提纲

- 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？

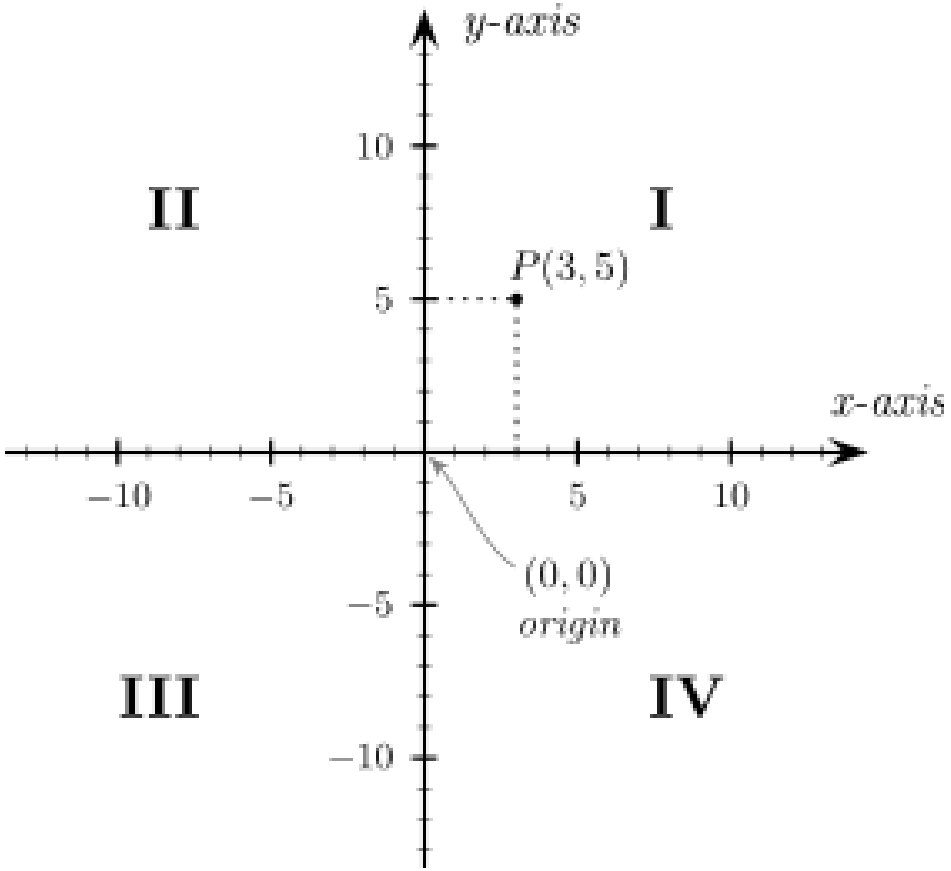
SPECIES	UCSC	DATE	NCBI
Human	hg19	Feb. 2009	Genome Reference Consortium
			GRCh37
	hg18	Mar. 2006	NCBI Build 36.1
	hg17	May 2004	NCBI Build 35
	hg16	Jul. 2003	NCBI Build 34
Mouse	mm10	Dec. 2011	Genome Reference Consortium
			GRCm38
	mm9	Jul. 2007	NCBI Build 37
	mm8	Feb. 2006	NCBI Build 36
	mm7	Aug. 2005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

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

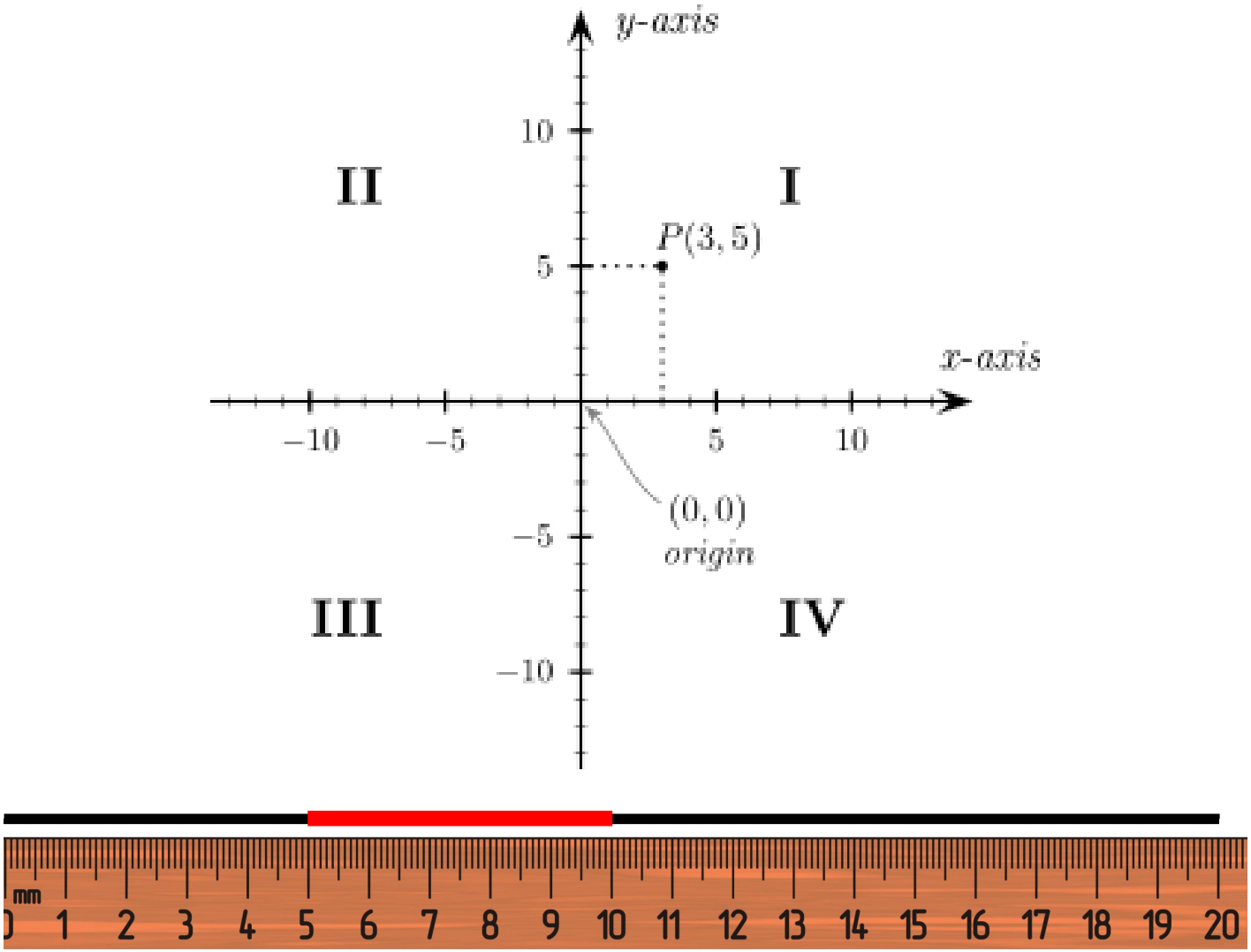
hg: human genome; GRC: Genome Reference Consortium

教学提纲

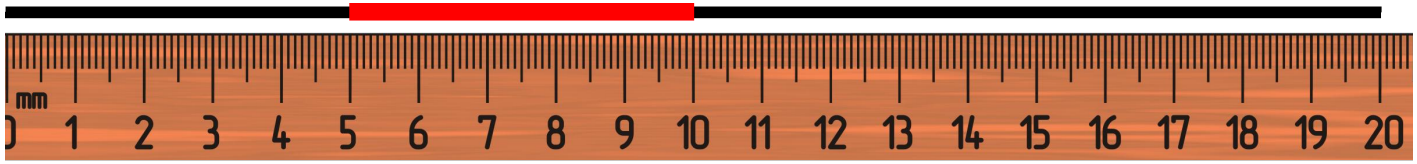
坐标系统 | 坐标轴

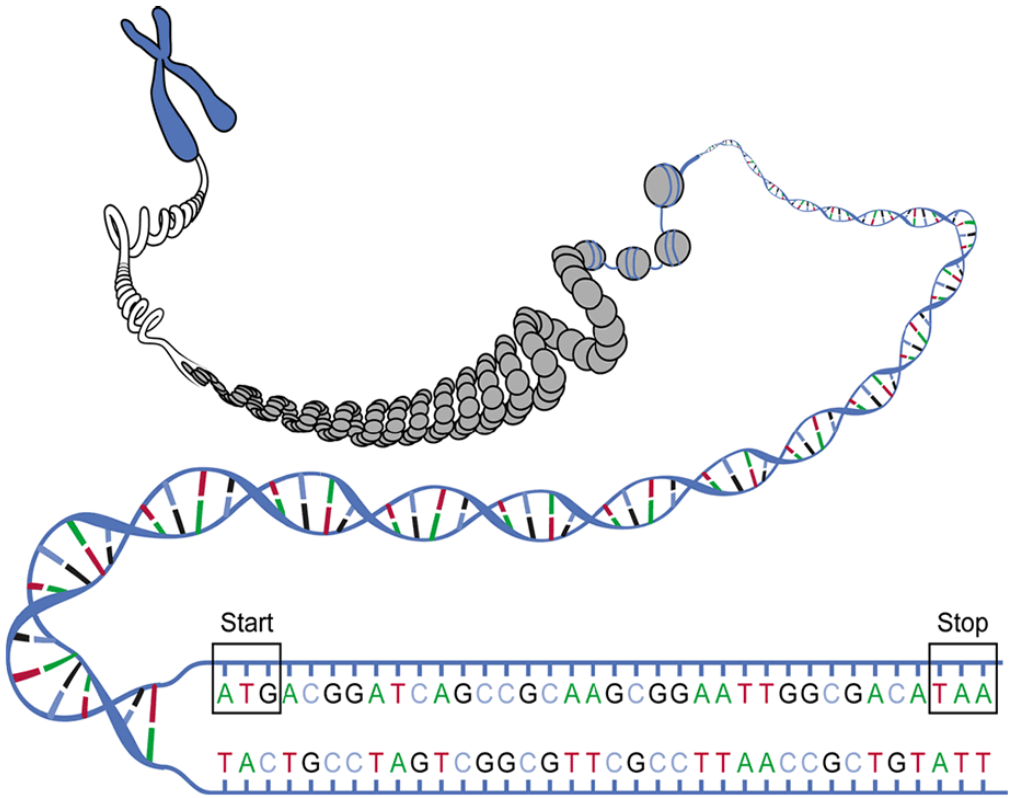


坐标系统 | 坐标轴

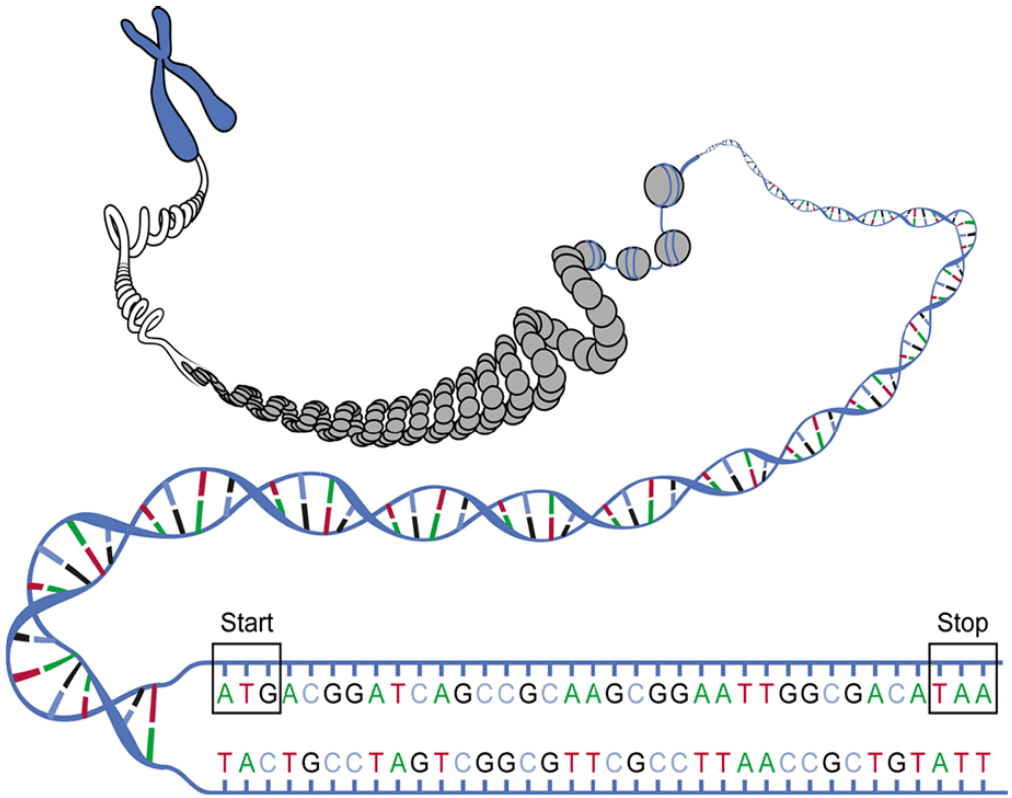

y-axis
x-axis
II
I
III
IV
P(3, 5)
(0, 0)
origin
10
5
−5
−10
−10
−5
5
10
mm
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

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"

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

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

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

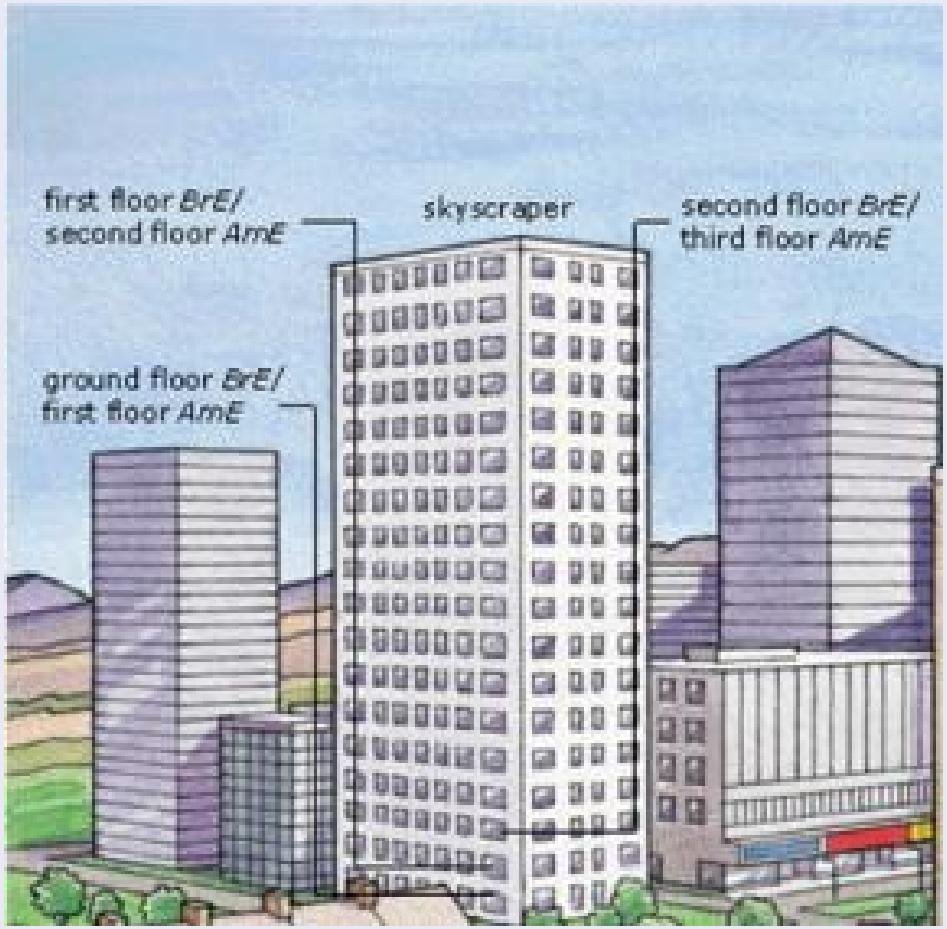
TG 的坐标

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

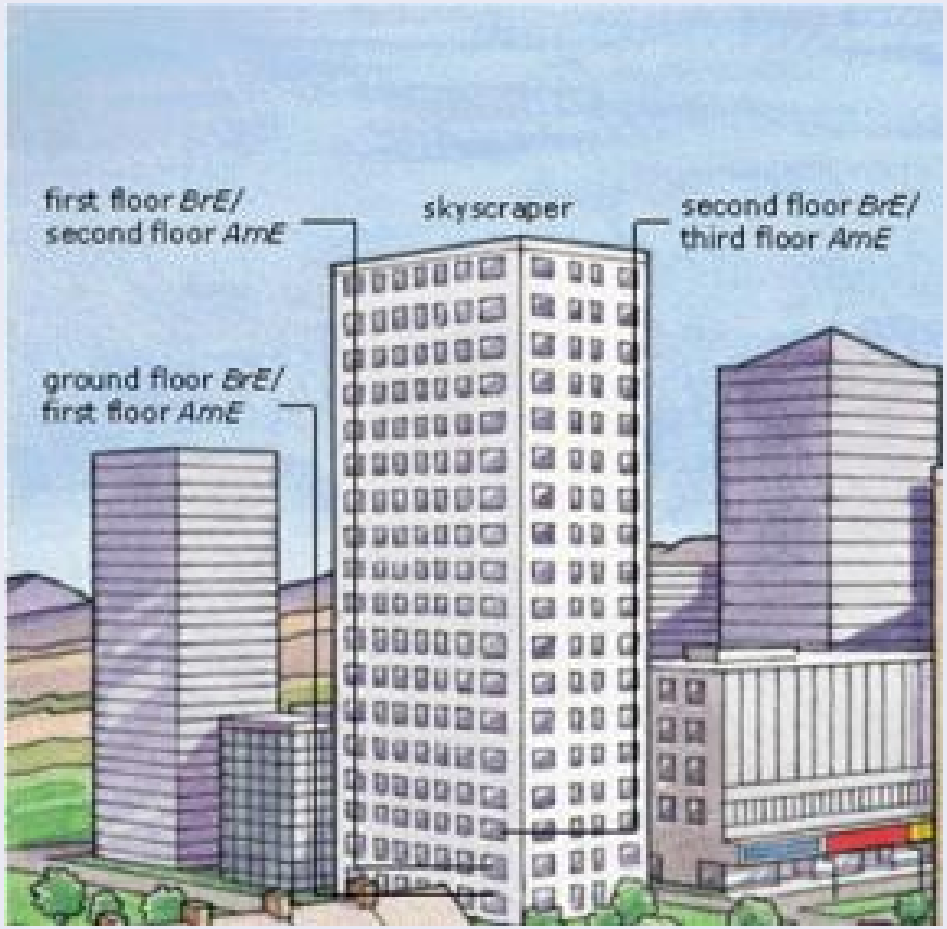
实例

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

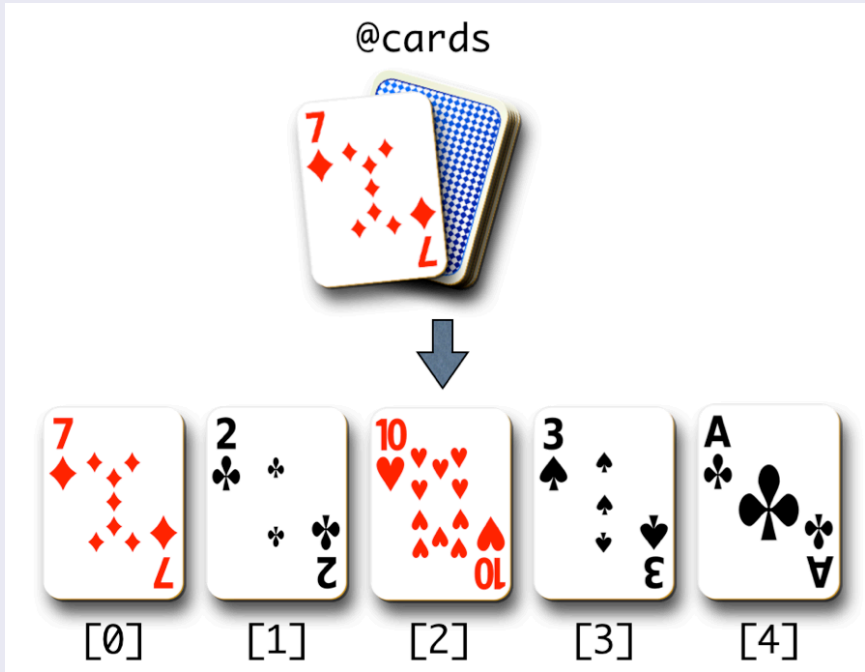
first floor


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

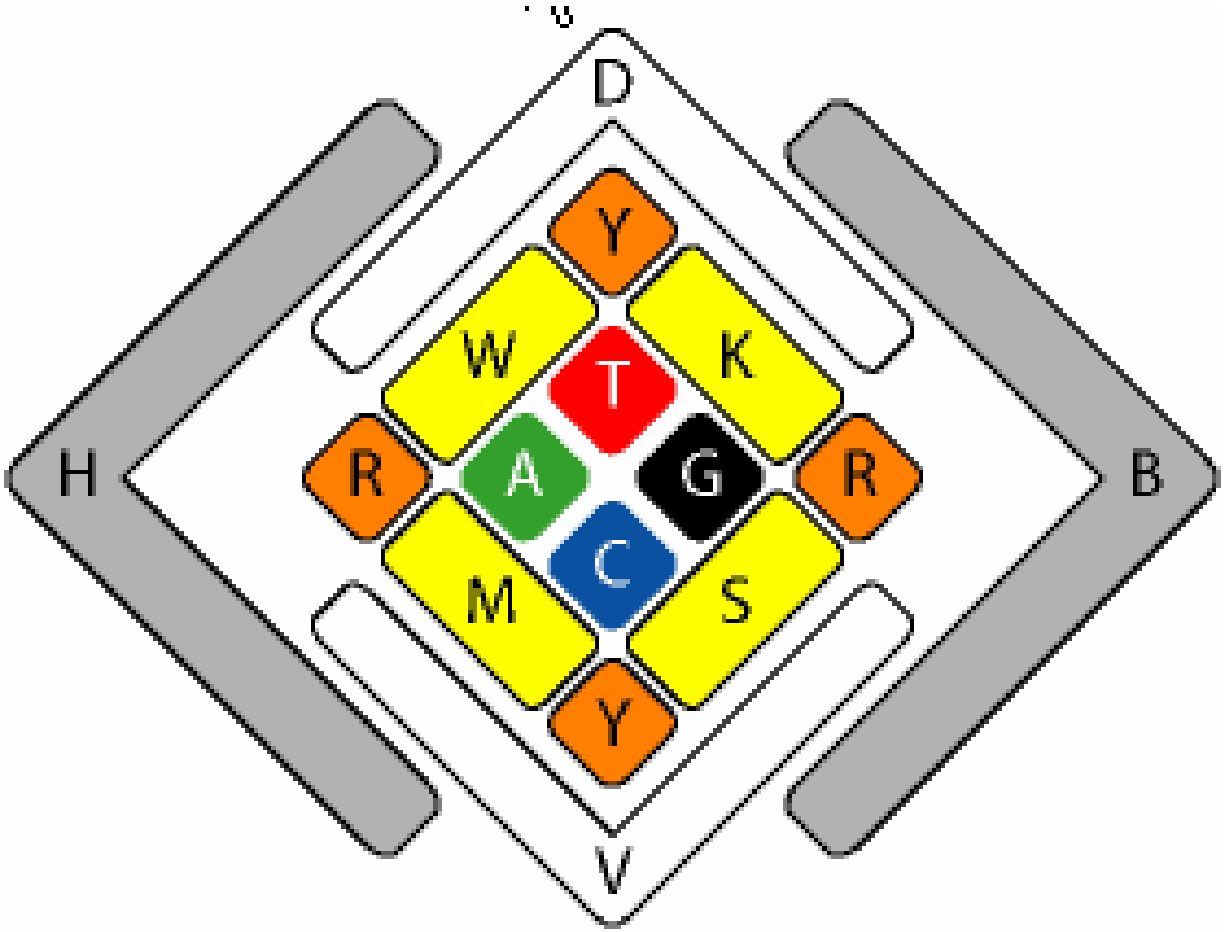
格式 | 文件格式

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
```

- 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			

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
```

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,
```

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 +

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

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

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)

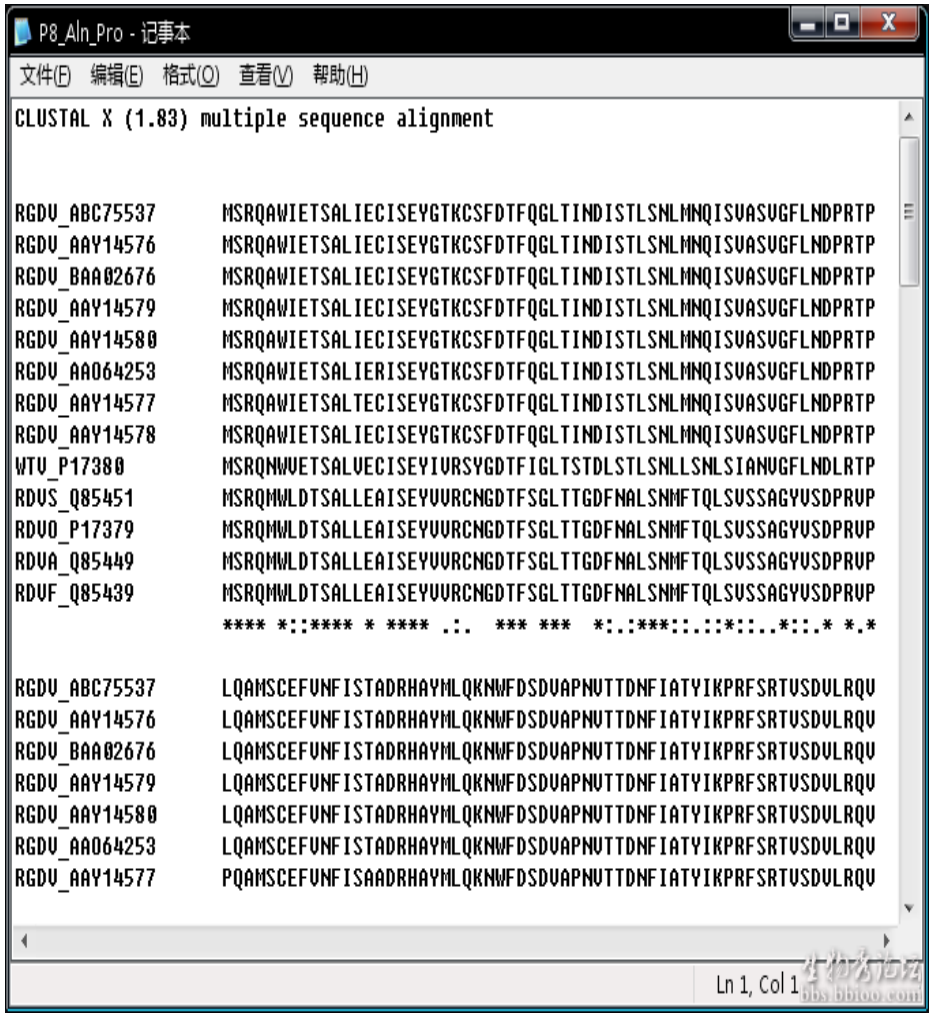
```
##gff-version 3
ctg123 . operon          1300 15000  .  +  .  ID=operon001;Name=superOperon
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Parent=operon001;Name=sonichedgehog
ctg123 . exon            1300  1500  .  +  .  Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  Parent=mrna0001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

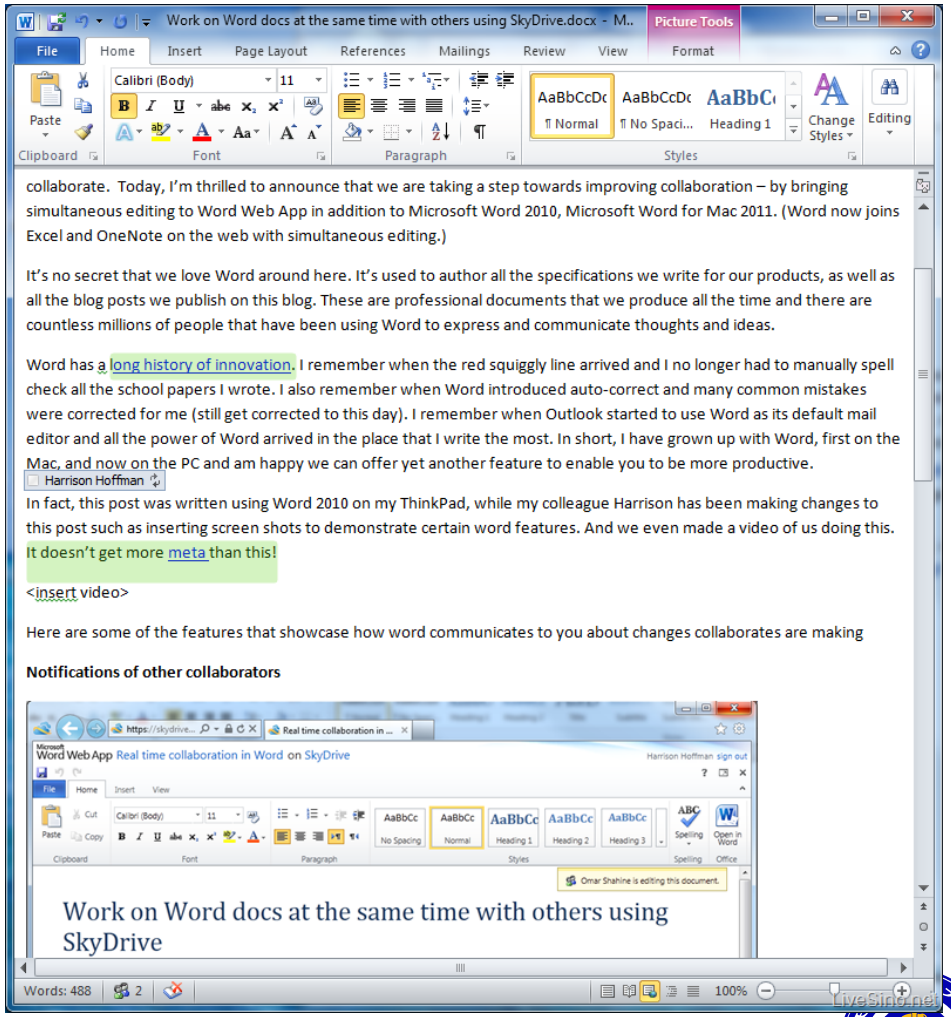
格式 | VCF (Variant Call Format)

```
##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

文本 | 纯文本 vs. 格式化文本





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

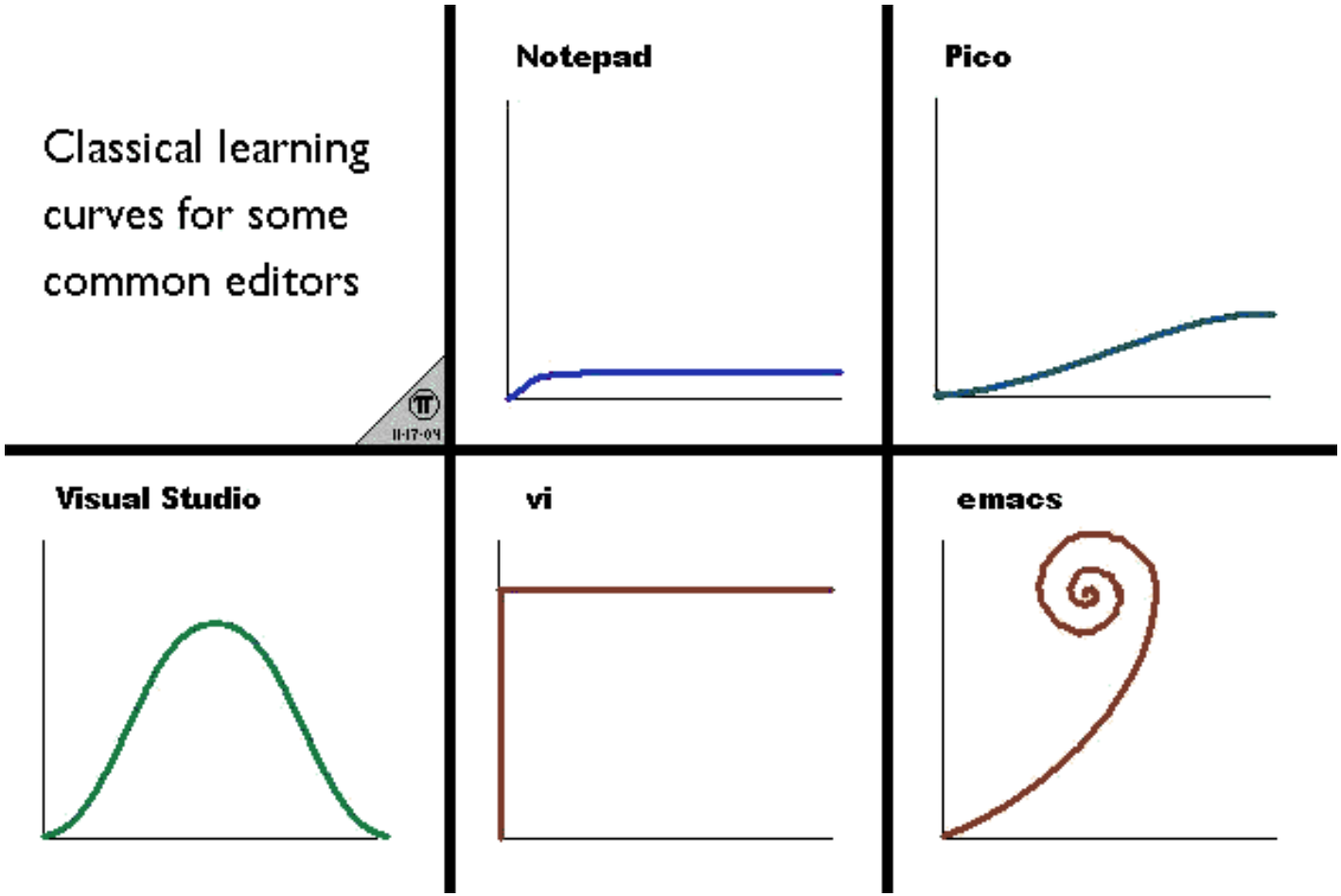
- Windows：文本编辑器，如 Notepad++
- Unix：file 识别，fromdos & todos 转换

三大类

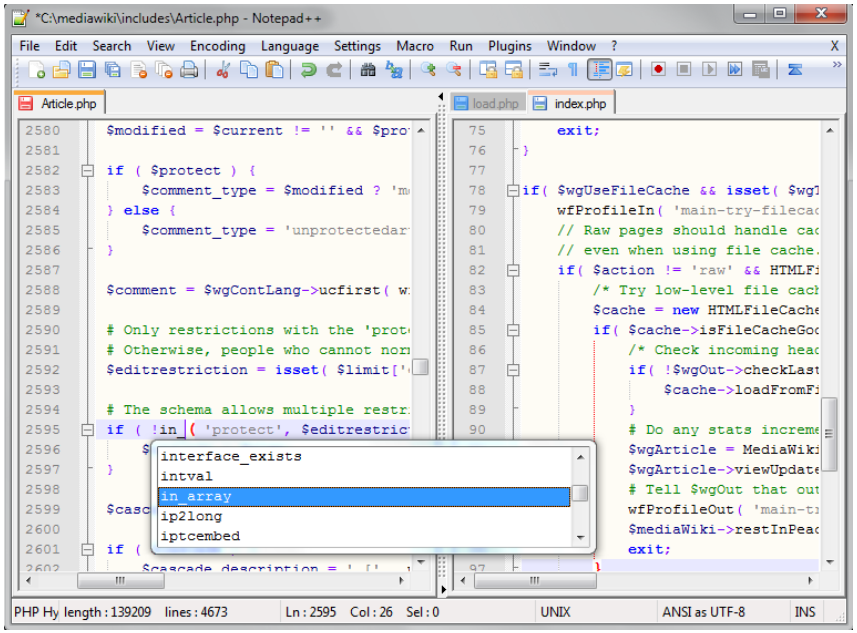
- Windows：`\r\n`（CR+LF，回车 + 换行），文件尾部直接 EOF（文件结束标志）
- Unix：`\n`（LF，仅有换行），文件最后一行也会增加该字符，然后才是 EOF
- Mac：`\r`（CR，仅有回车）

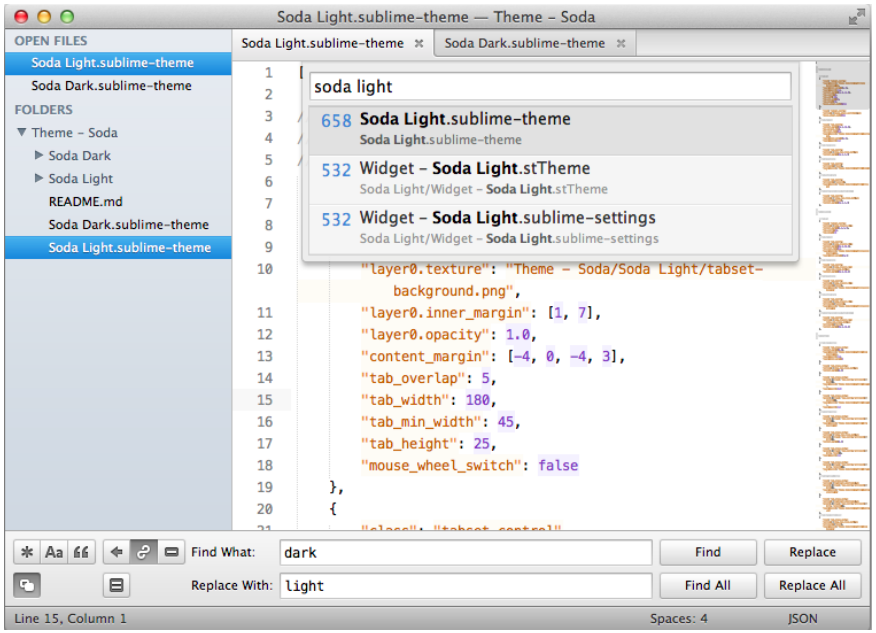
识别与转换

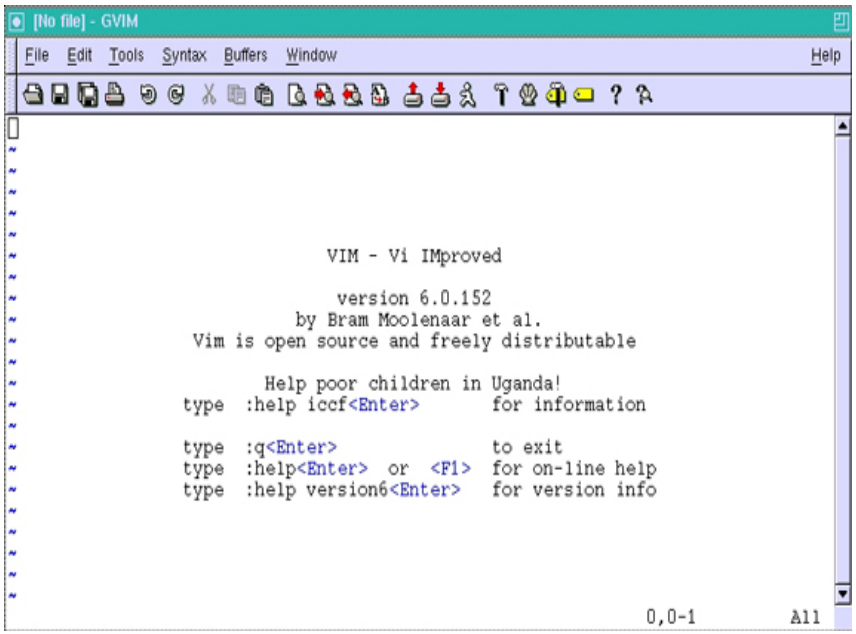
- Windows：文本编辑器，如 Notepad++
- Unix：file 识别，fromdos & todos 转换

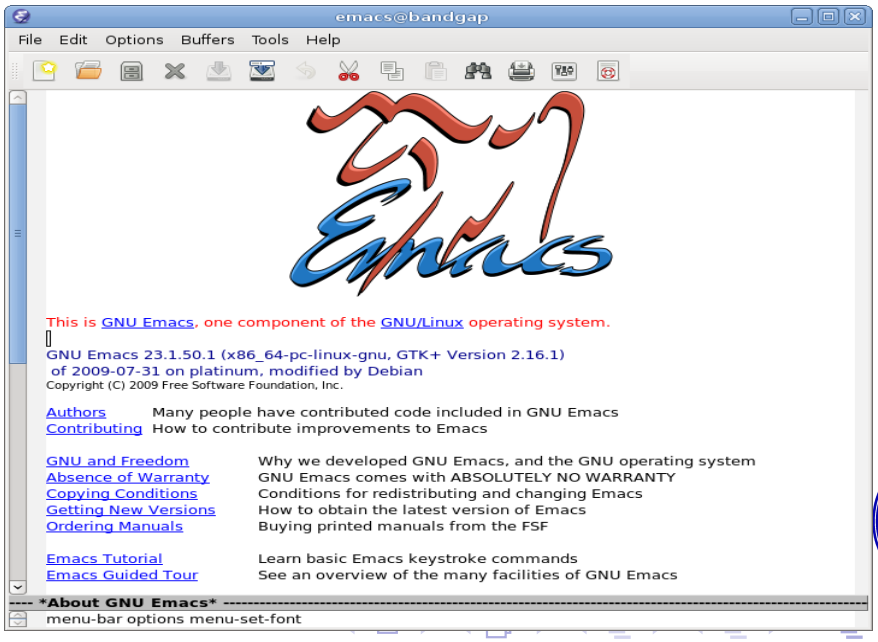

Classical learning curves for some common editors
Notepad
Pico
Visual Studio
vi
emacs

文本 | 编辑器 | Notepad++, Sublime Text, Vim, Emacs

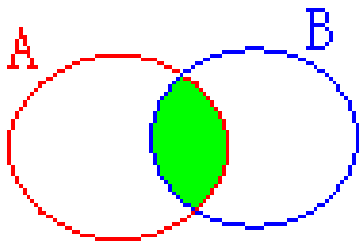






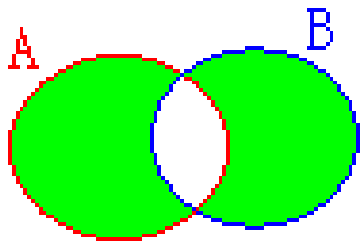


教学提纲

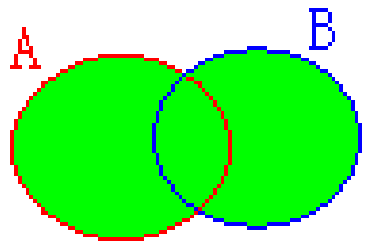


求同

交集

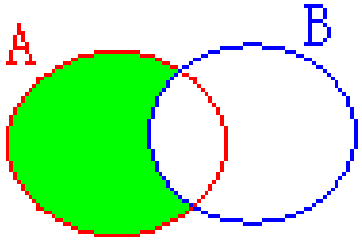


求异



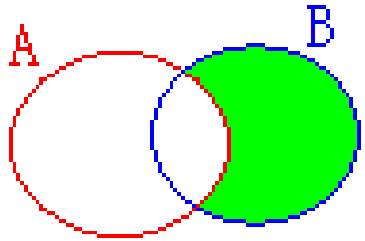
相加

并集

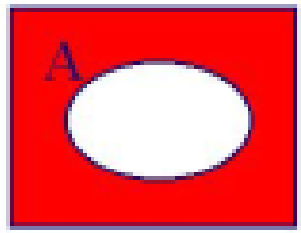


相减 A-B

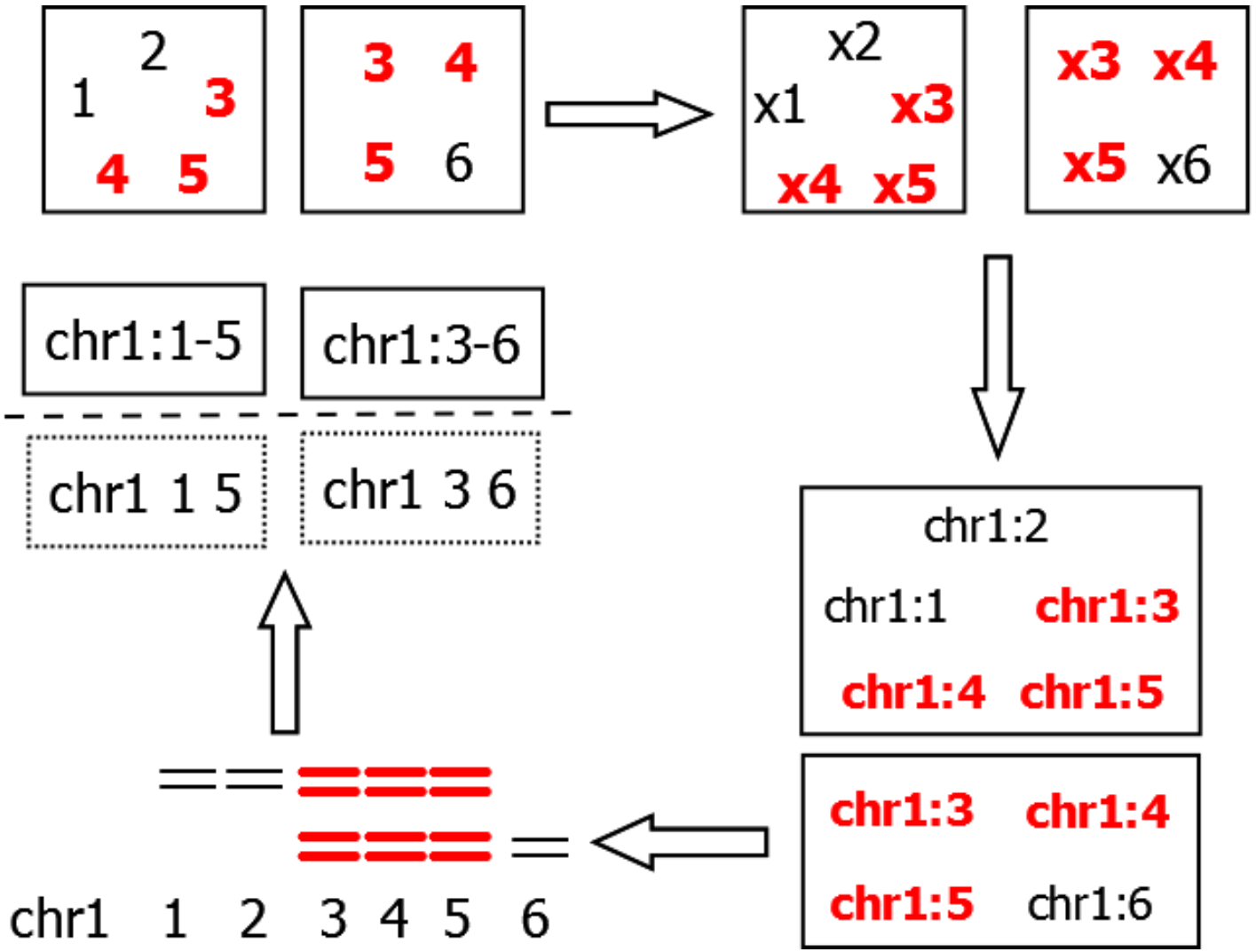
差集

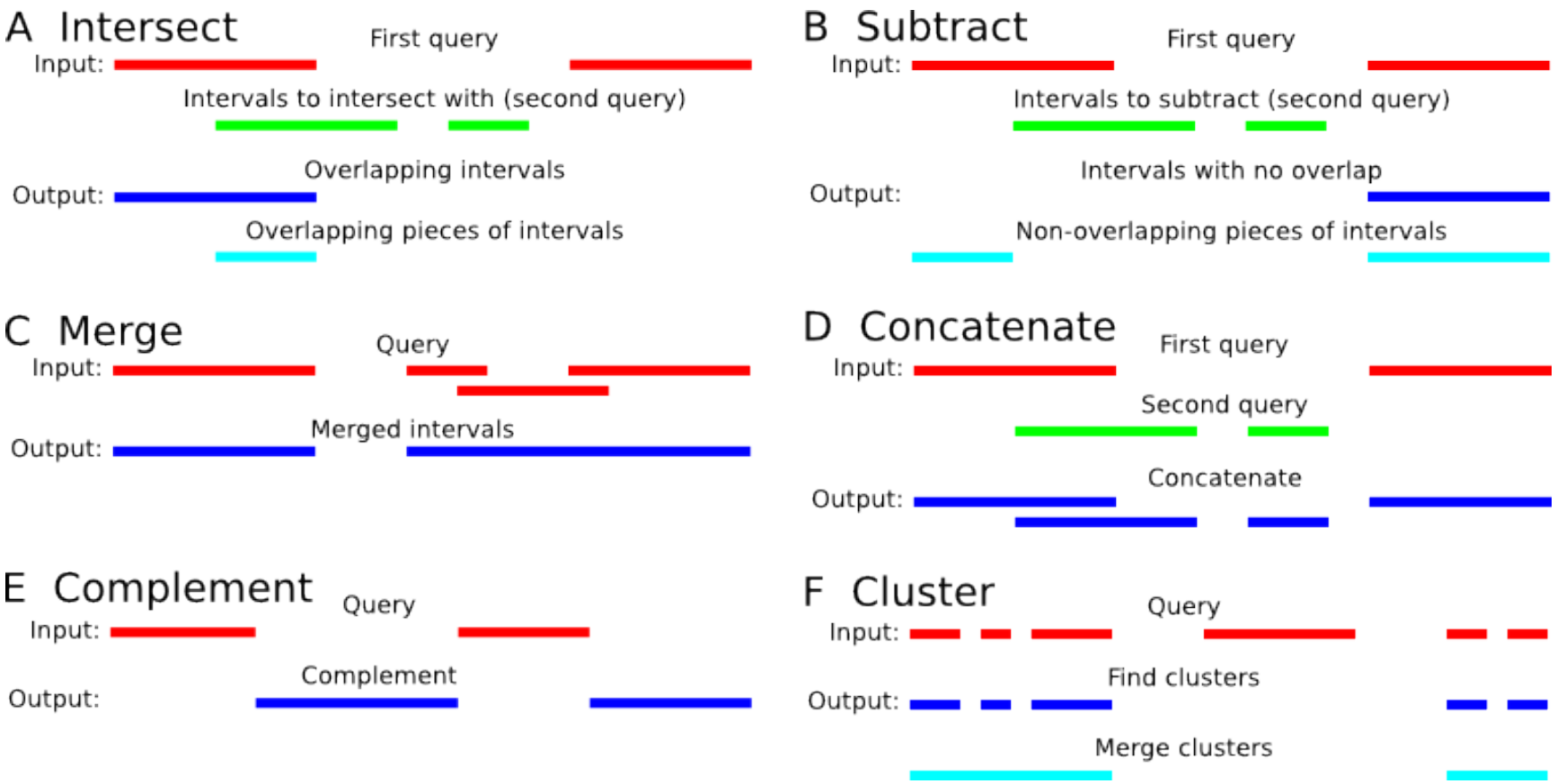


相减 B-A

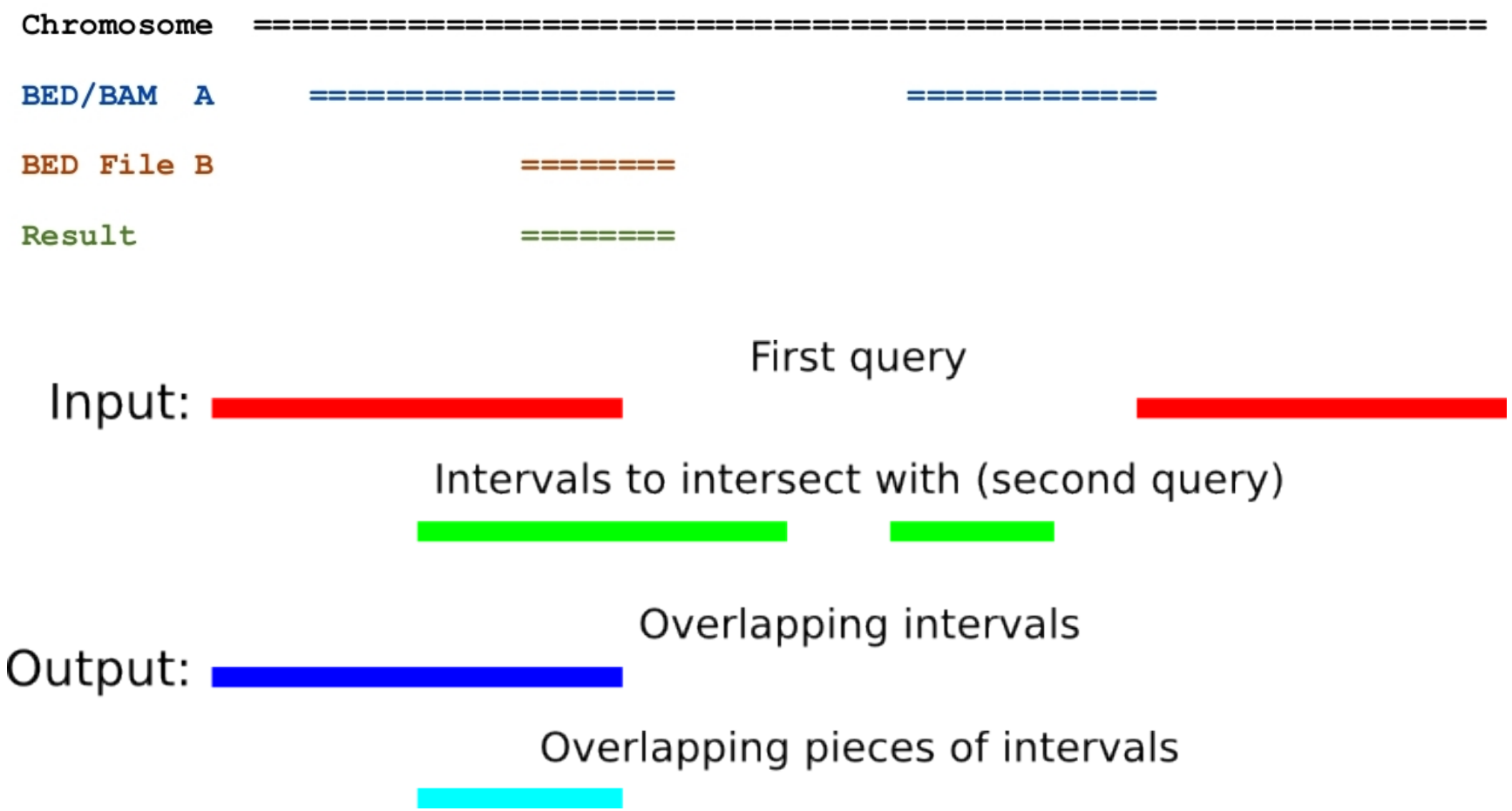


补集

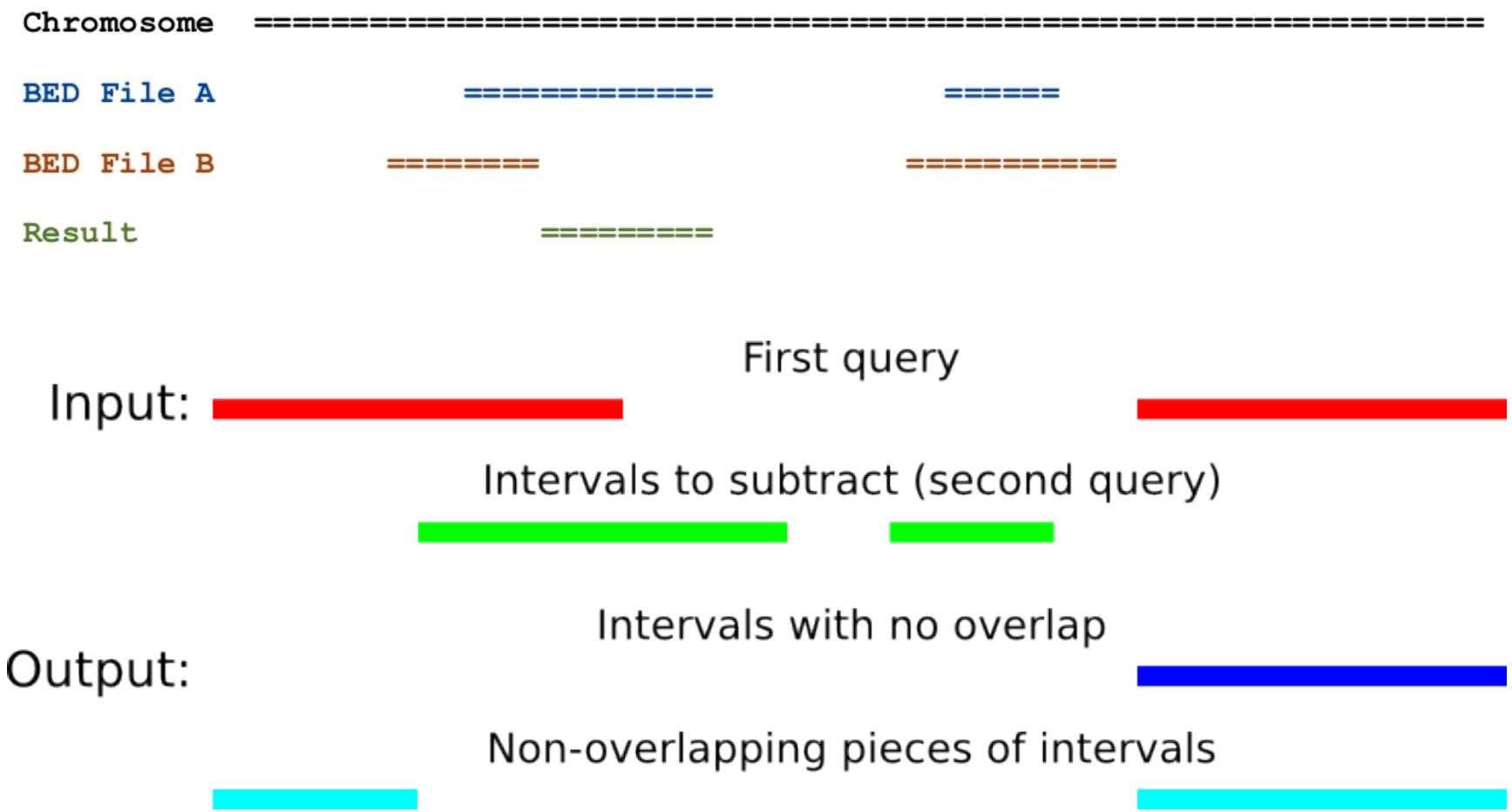

2
1 3
4 5
3 4
5 6
x2
x1 x3
x4 x5
x3 x4
x5 x6
chr1:1-5
chr1:3-6
chr1 1 5
chr1 3 6
chr1:2
chr1:1 chr1:3
chr1:4 chr1:5
chr1:3 chr1:4
chr1:5 chr1:6
chr1 1 2 3 4 5 6


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

逻辑运算 | intersect

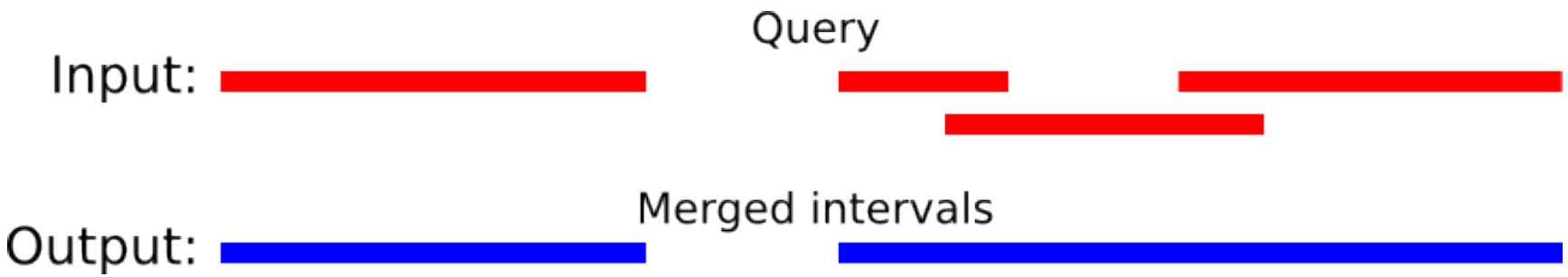


逻辑运算 | subtract

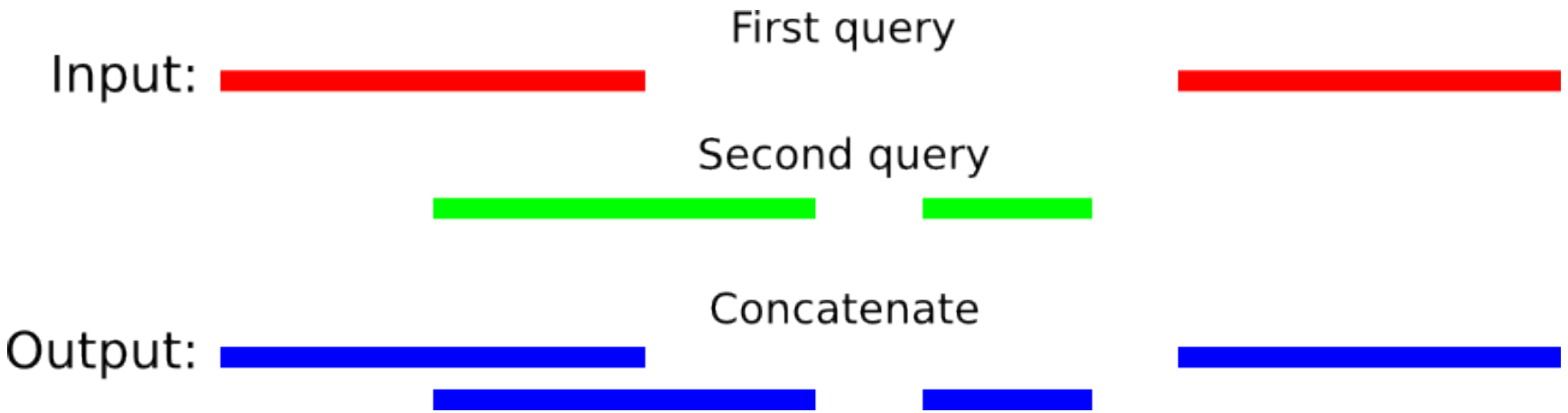


逻辑运算 | merge

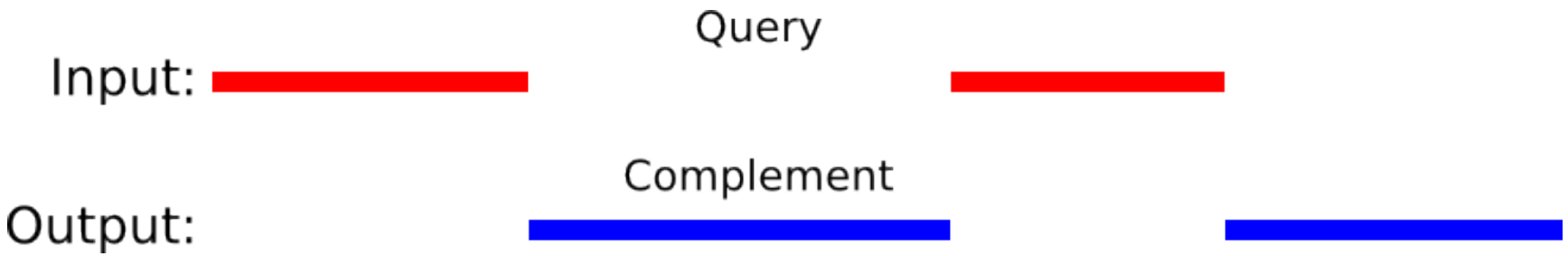

Chromosome
BED File
Result


Query
Input:
Merged intervals
Output:

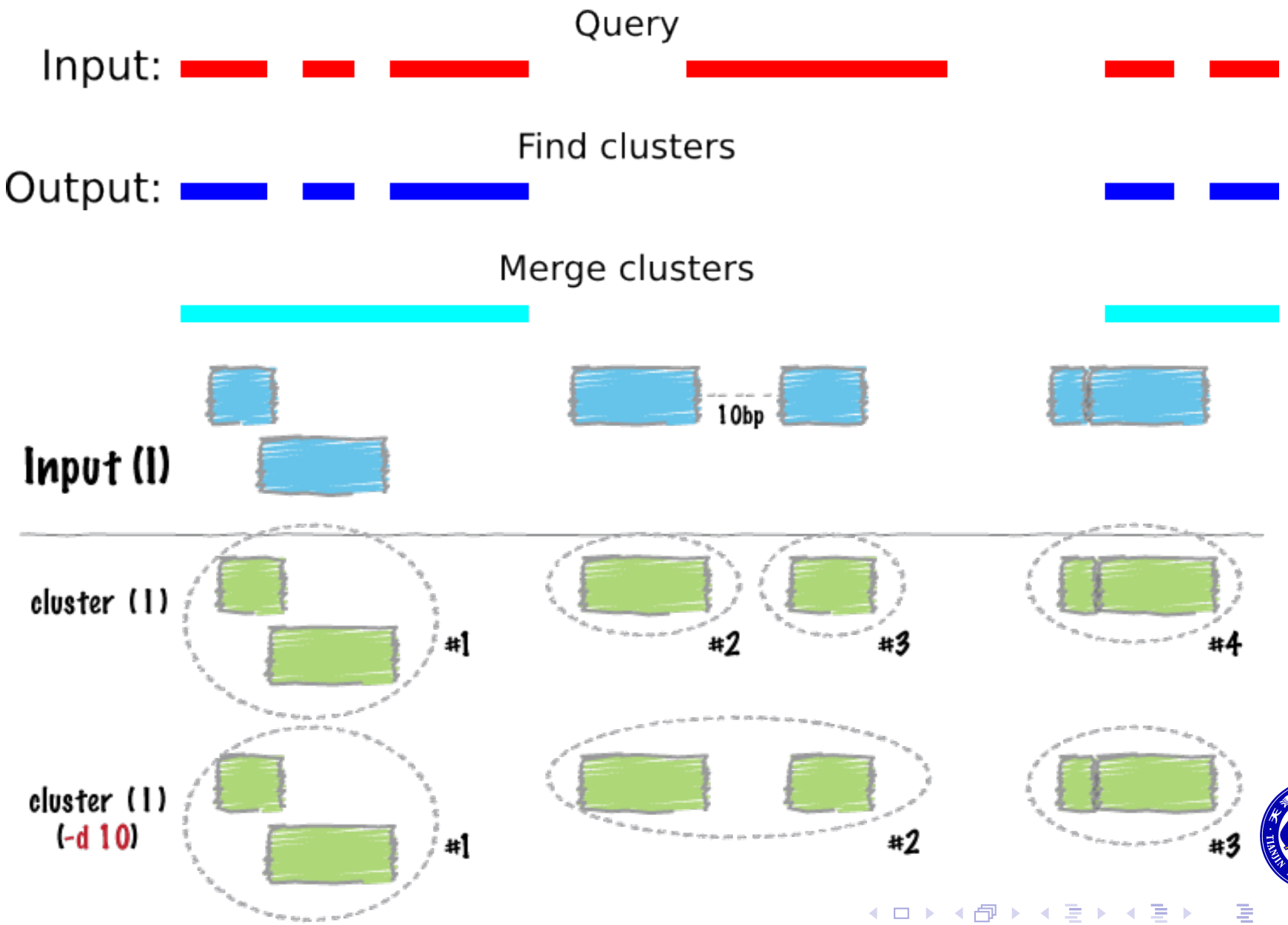
逻辑运算 | concatenate



逻辑运算 | complement



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

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

(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

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

(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 ".")

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.

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

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

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

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

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

- 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

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
   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
   - 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

复习思考题

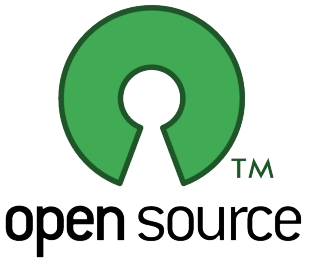
知识点

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有哪些，画图进行解释。

技能

1. 不同操作系统的换行符有何区别？如何进行查看和转换？

Powered by



TEX LATEX XETEX Beamer