



第17期OmicShare课堂

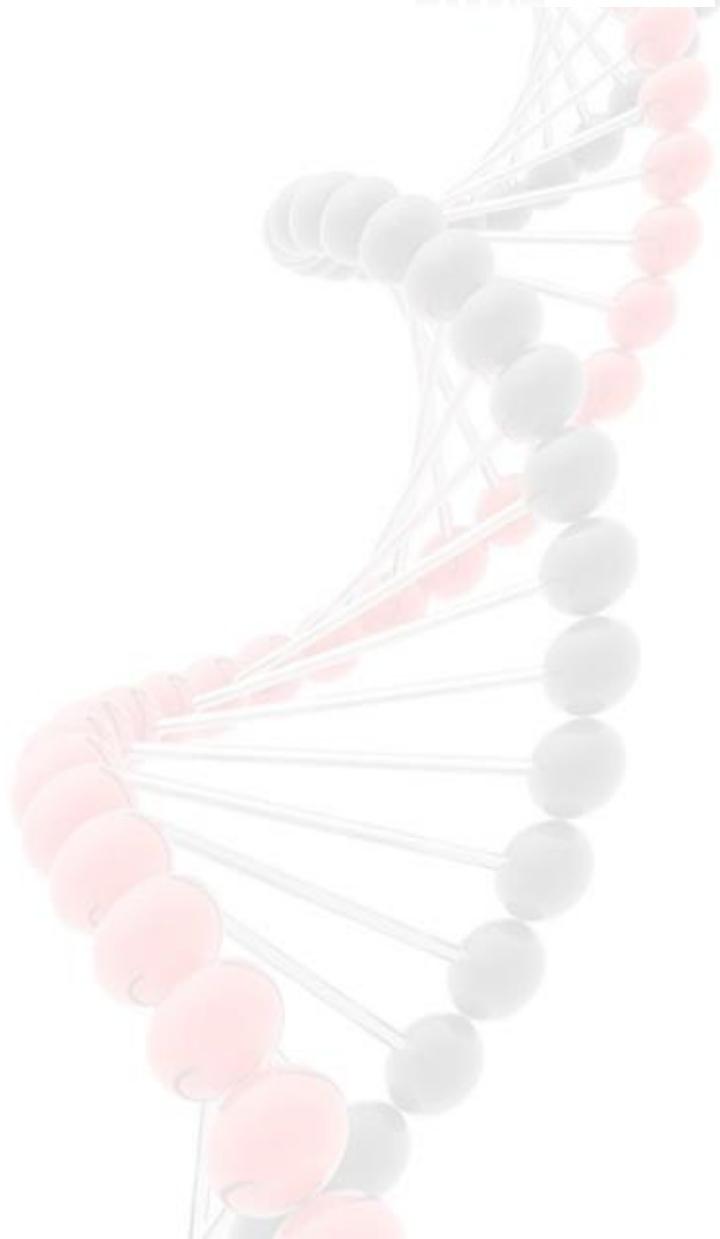


基因组浏览器 (IGV) 使用教程

提纲



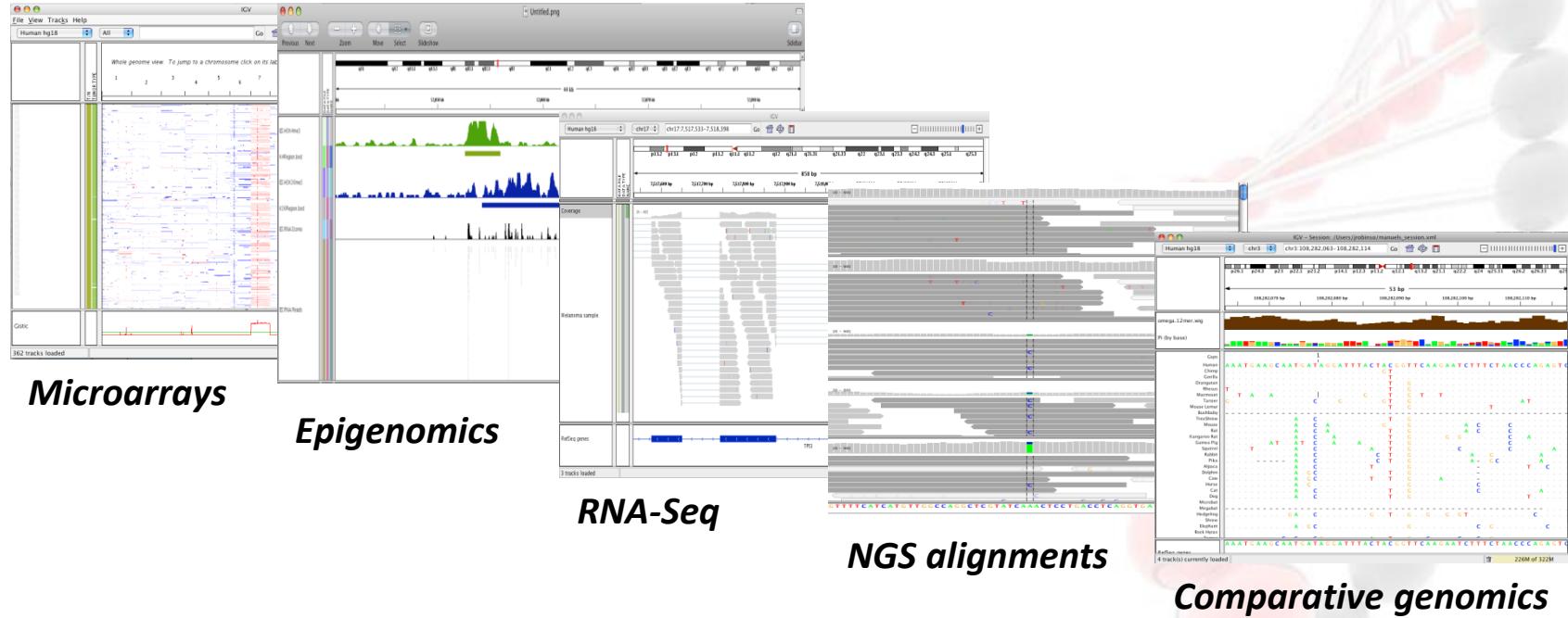
- 软件介绍与安装启动
- 数据导入与文件格式介绍
- IGVTools介绍
- 数据练习



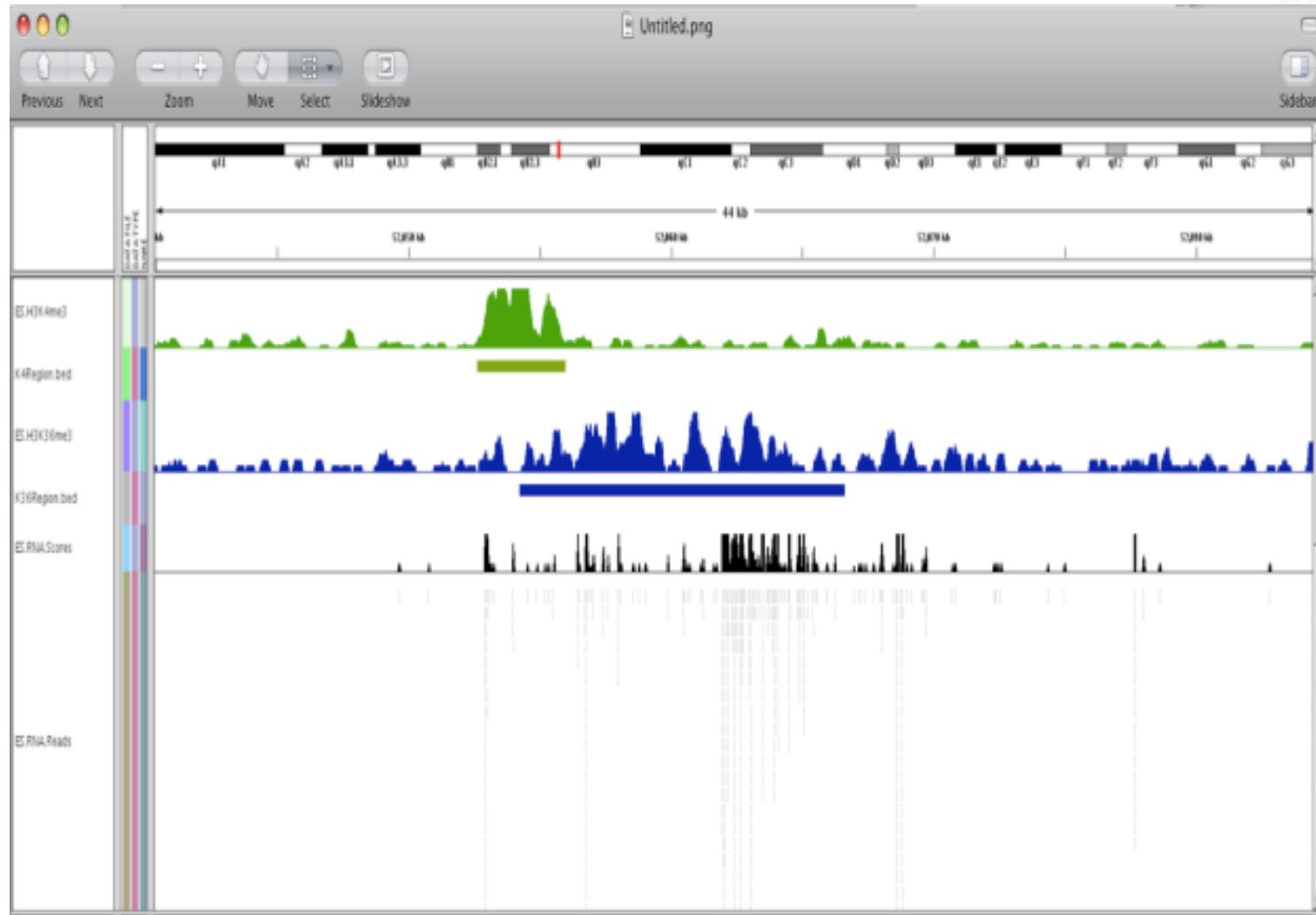
什么是 IGV



一款Java软件，可以桌面使用，也可命令行运行；
用于基因组中各类数据的整合可视化，非常适合NGS数据。

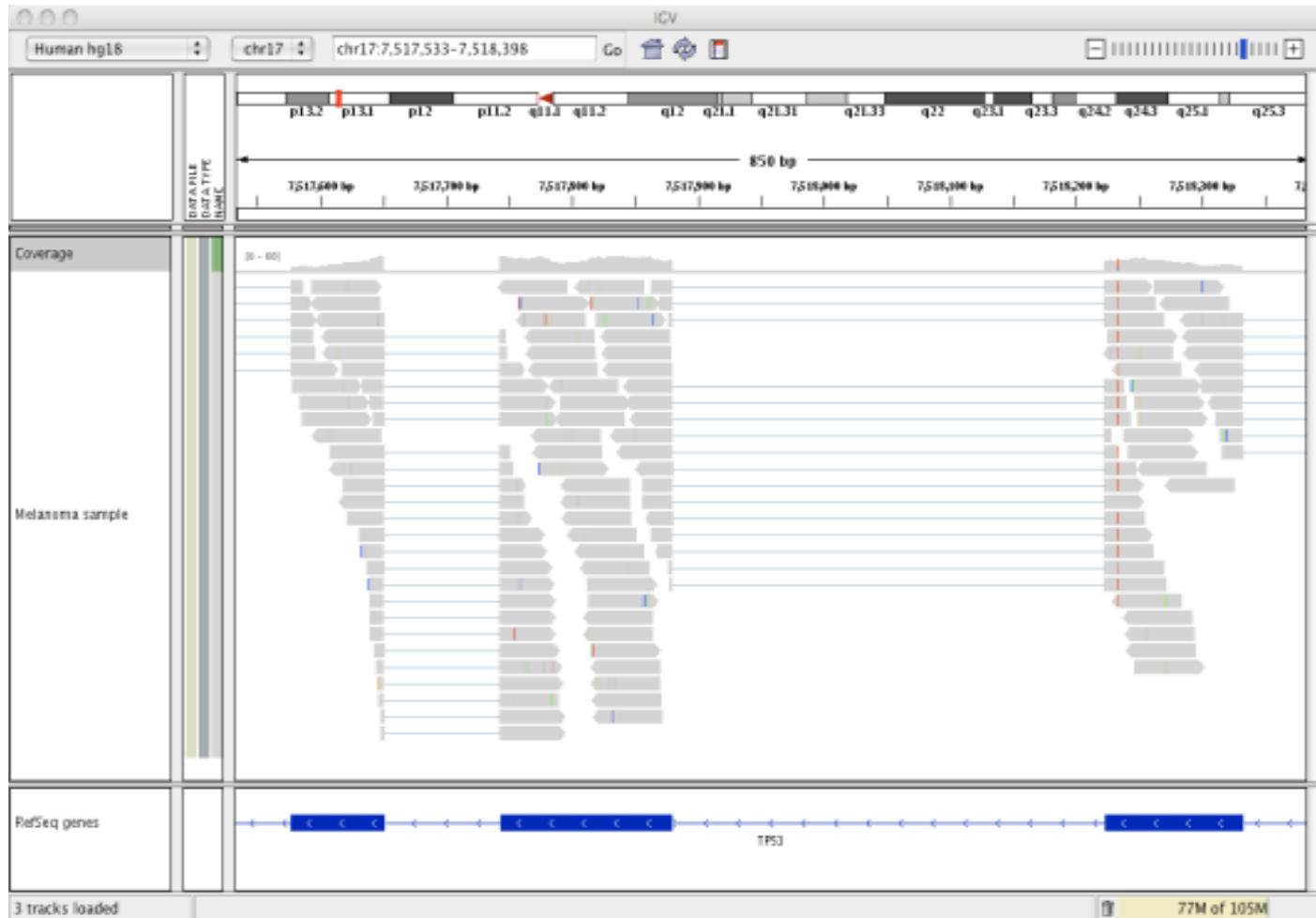


信号丰度



信号类型包括：mRNA表达量，组蛋白修饰程度，甲基化率等

比对结果



编码基因外显子剪切方式的查看 (RNA-seq比对结果)

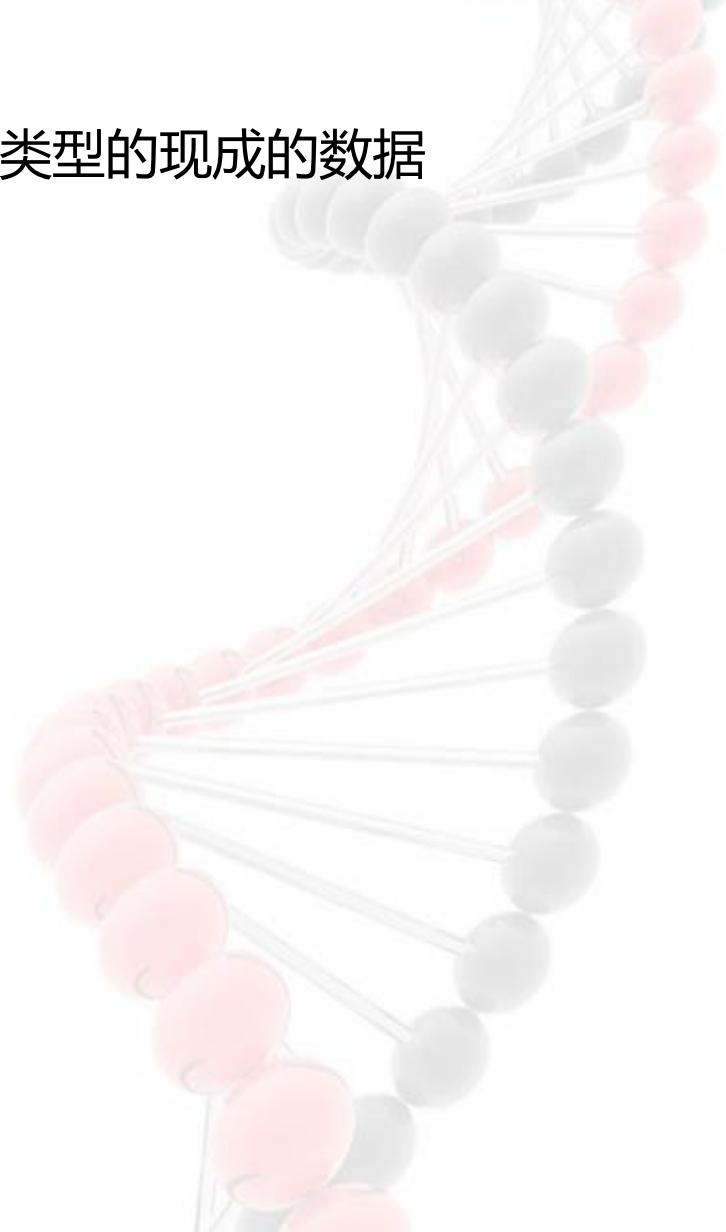
比对结果



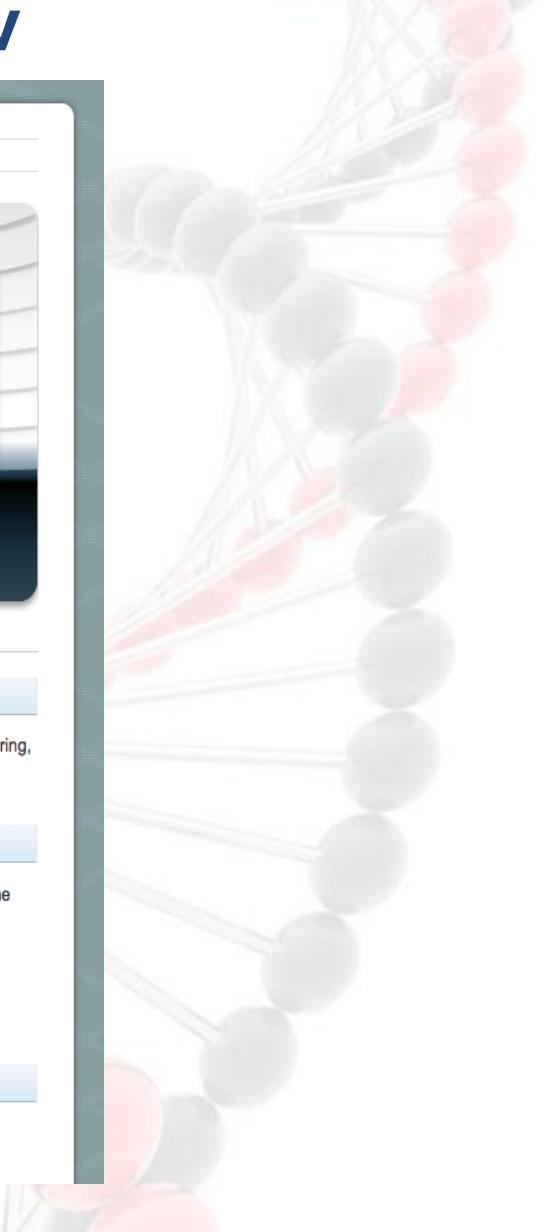
全基因组重测序，局部比对结果，例如查看局部缺失信息。以上图片是基迪奥一个分析项目——寻找拟南芥T-DNA插入位点。通过寻找跨越外源插入片段和基因组序列的paired-END序列，确定插入位点。

IGV 的优点

- 满足不同类型的研究的需求，便于查看各种类型的现成的数据
 - The Cancer Genome Atlas (TCGA)
 - Epigenetic & lincRNA studies
 - 1000 Genomes Project
 - 个人的研究项目
- 满足不同类型的用户 –
 生物学者和生物信息学学者
- 在单机桌面系统下也可以处理大数据
- 交互式界面，容易使用



<http://www.broadinstitute.org/igv>



igv Integrative Genomics Viewer ALMSL

- Home
- Downloads
- Documents
- FAQ
- IGV Quick Start
- IGV User Guide
- File Formats
- Release Notes
- Acknowledgments

@ Contact

Search website

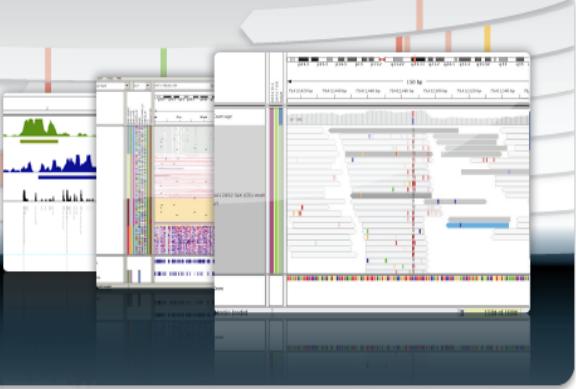
search

Broad Home
Cancer Program

 © 2009 Broad Institute

Home

Integrative Genomics Viewer



What's New

NEWS December 16, 2009. IGV version 1.4.1 has been released. See the [release notes](#) for details.

October 29, 2009. IGV version 1.4 has been released. Highlights of this release include new alignment track features and a command line utilities package, igvtools.

Overview

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated datasets. It supports a wide variety of data types and provides easy access to genomes and datasets hosted by the Broad Institute.

Downloads

Please [register](#) to download IGV. After registering, you can log in at any time using your email address.

Funding

Development of IGV is made possible by funding from the [National Cancer Institute](#) and the [National Institute of General Medical Sciences](#) of the [National Institutes of Health](#).

Citation

To cite your use of IGV, please reference
<http://www.broadinstitute.org/igv>

IGV 下载与安装



- 注册（没有限制）, <http://www.broadinstitute.org/igv>
- 点击 “Downloads”
- 选择系统类型， Mac or window, 或 linux
- 备注：如果运行，电脑需要先安装Java程序

Downloads

 Please [register](#) to [download](#) IGV. After registering, you can log in at any time using your email address. Permission to use IGV is granted under the GNU [LGPL license](#).

IGV启动

名称	修改日期	类型	大小
batik-codec_V1.7.jar	2014/6/19 22:43	Executable Jar File	173 K
goby-io-igv_V1.0.jar	2014/6/19 22:43	Executable Jar File	2,070 K
igv.bat	2014/10/10 14:08	Windows 批处理...	1 K
igv.bat.bak	2014/6/19 22:43	BAK 文件	1 K
igv.command	2014/6/19 22:43	COMMAND 文件	1 K
igv.jar	2014/6/19 22:43	Executable Jar File	29,398 K
igv.sh	2014/6/19 22:43	SH 文件	1 K
readme.txt	2014/6/19 22:43	TXT 文件	2 K

- igv.jar 是IGV软件的执行程序（ java程序 ），但在 windows下一般使用bat脚本启动jar，以便对软件参数进行重新设置。

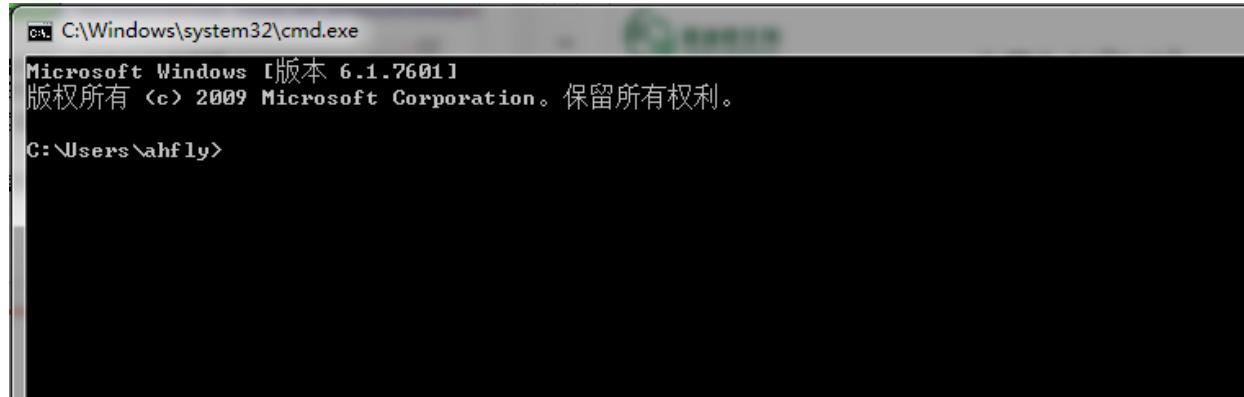
IGV启动

```
1 ::Get the current batch file's short path
2 for %%x in (%0) do set BatchPath=%~dpsx
3 for %%x in (%BatchPath%) do set BatchPath=%~dpsx
4 java -Xmx3000m -Dproduction=true -Djava.net.preferIPv4Stack=true
5
```

- 第一次启动前，使用文本编辑器（ notepad + ）修改 IGV使用的内存大小（这里我设定为3G）；
- 保存，然后双击点击igv.bat，启动IGV；
- **备注：按照你电脑的配置修改；**

IGV启动

- 机房的部分计算机双击无法启动bat文件的解决方法；
1) 进入dos (cmd) 界面

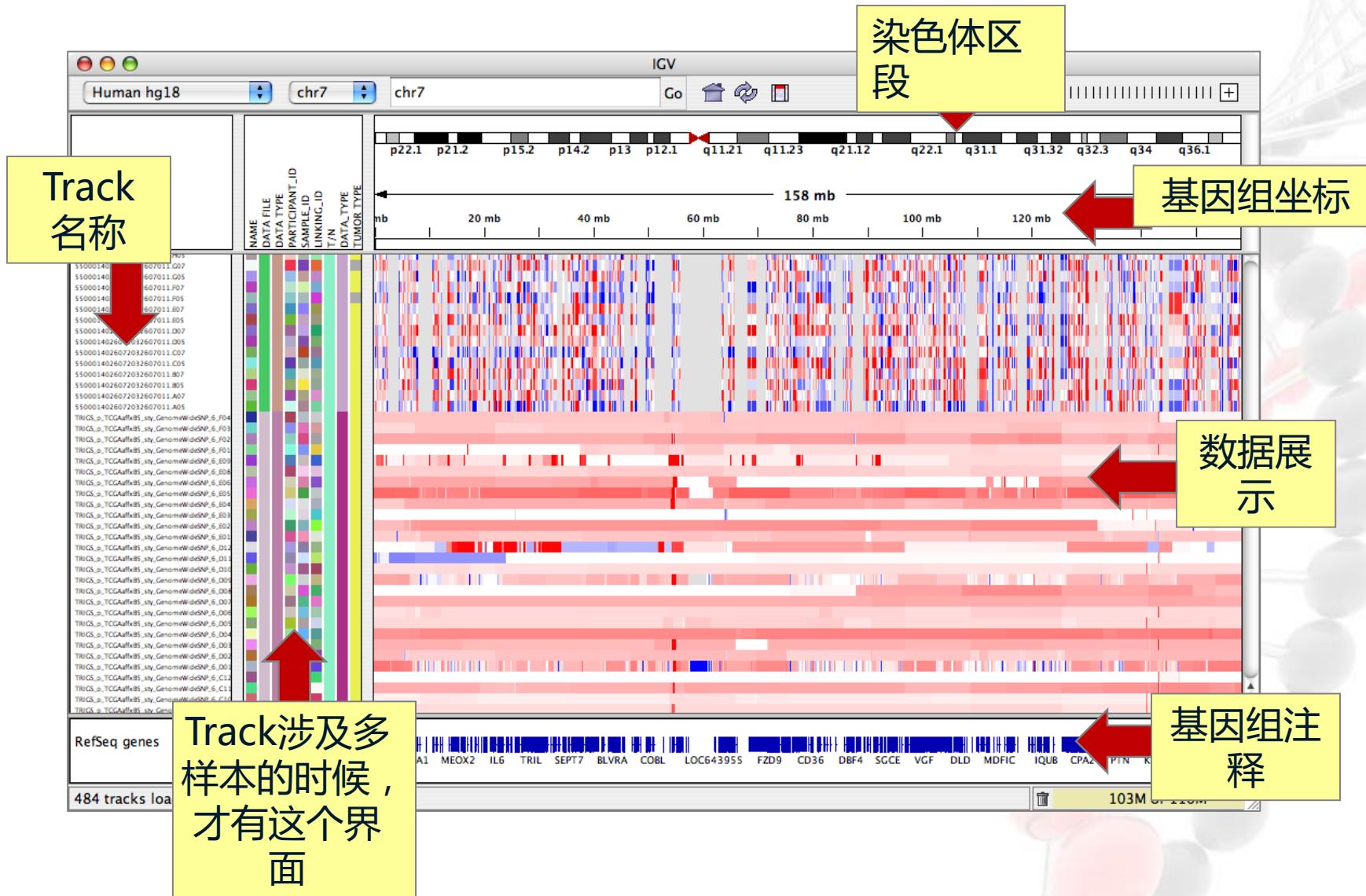


- 2) 然后使用鼠标将bat文件拖动到dos界面内，然后回车即可。

IGV 页面布局



Integrative
Genomics
Viewer



提纲



- 软件介绍与安装启动
- 数据导入与文件格式介绍
- IGV tools
- 数据练习



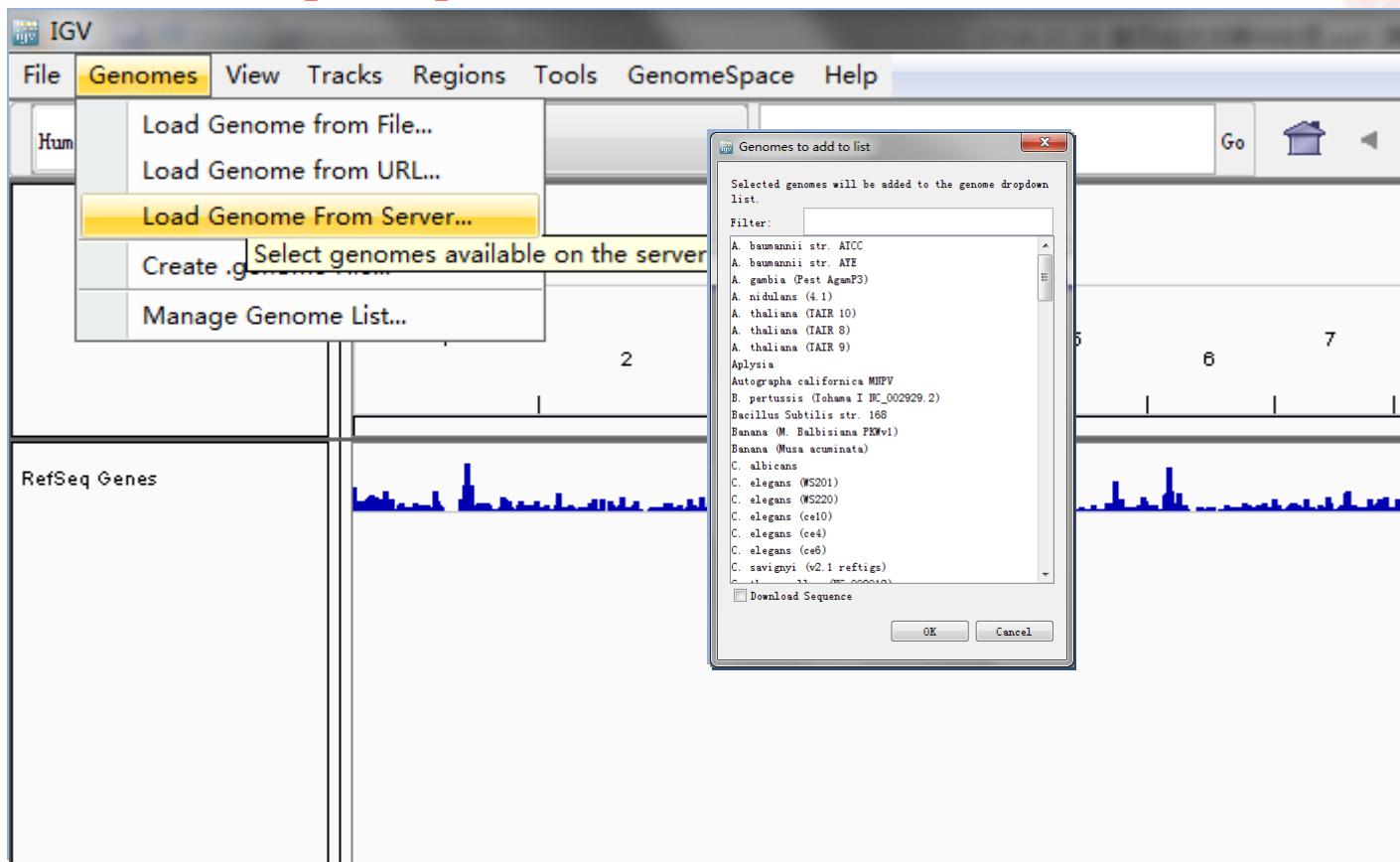
数据导入的顺序



1. 选择（并导入）参考基因组
2. 导入其他数据
3. 浏览数据，选取目标区域
4. 设置 track 的属性，优化展示结果

备注：在此类软件里，一个输入文件就被称为一个track

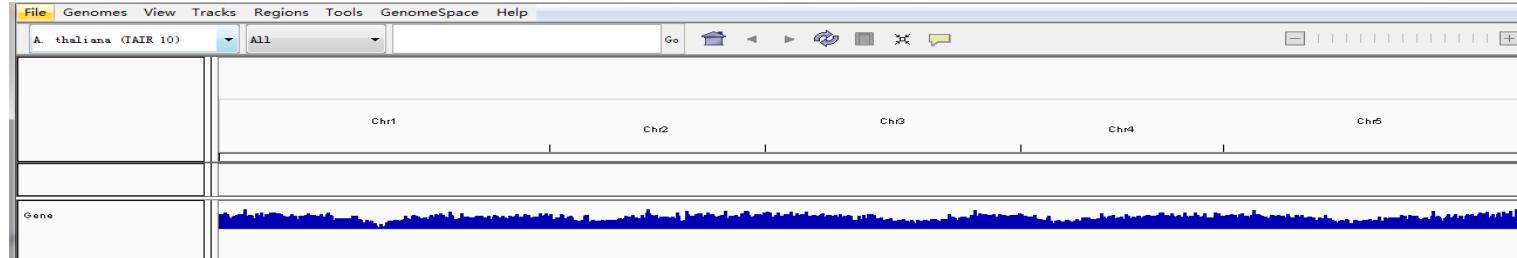
(1) 导入参考基因组



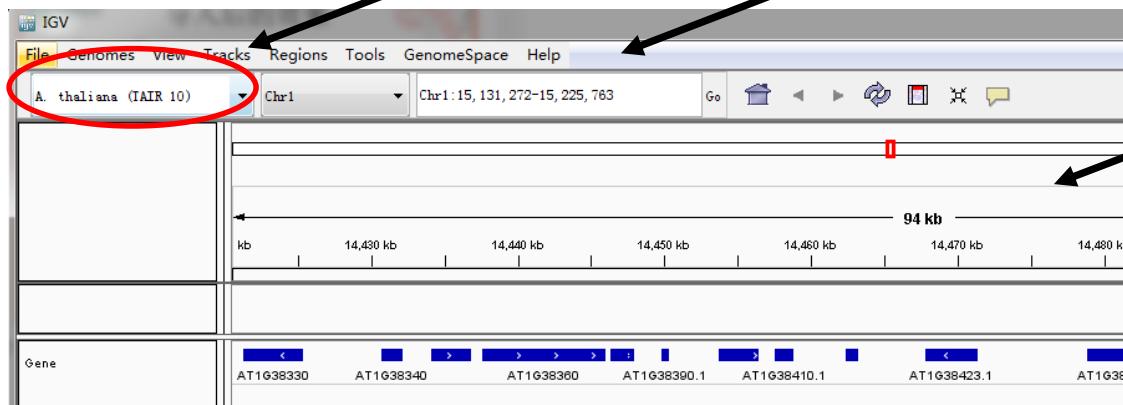
启动后，可以选择下载基因组的信息，绝大部分为动物，但也包含水稻、玉米、拟南芥、大豆这样模式植物；

导入拟南芥基因组后的效果

全局的效果



局部的效果



参考基因型名称

染色体编号和区间坐标信息

区间的标尺

基因信息

导入参考基因组——生成基因组文件

基因组文件是*.genome格式，一般存储在“我的文档里面”

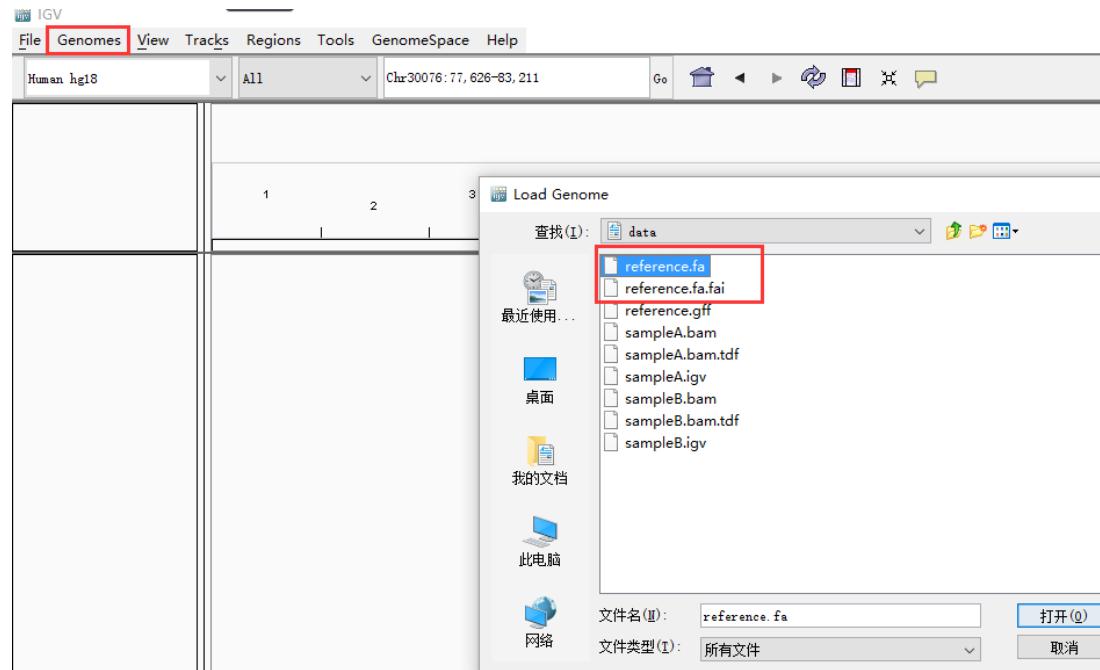
例如在机房的电脑，数据路径在c:\Documents and Settings\user\igv\genome

人类hg18版本的参考基因组

打开 ▾ 新建文件夹				
	名称	修改日期	类型	大
	hg18.genome	2015/3/31 21:07	GENOME 文件	

如果你研究物种的基因组文件在IGV数据库没有（版本号不对应或根本没有），怎么办？

参考基因组——直接导入Fasta格式的文件



更简单的是：Genomes → load from files → 点击选择相应的fa 文件

备注：这里为了让文件更小便于演示，序列数据来源某物种的单个scaffold。

(2) 导入其他信息



数据信息类型

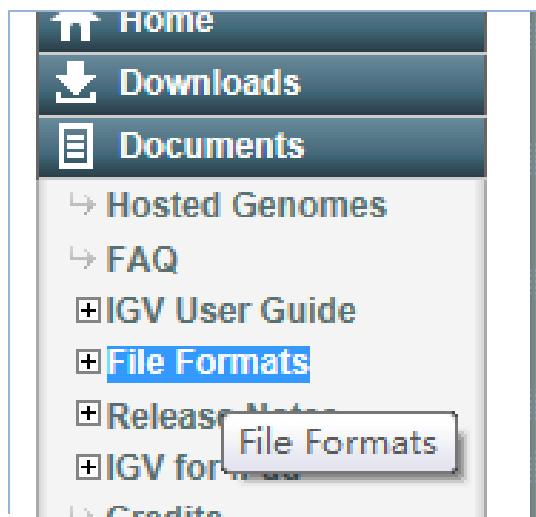
- 所有与基因组坐标有关的测序数据
- 基因组注释

文件格式

- 支持多种格式；
- 这里介绍几种常见格式，更多信息，请见官网
www.broadinstitute.org/igv/FileFormats

输入的格式

- 以IGV软件为例，其实其官网介绍了常见的输入格式：



File Formats

- File Extension
- Identifies Format
- Recommended File Formats
- BAM
- BED
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- broadPeak
- CBS
- CN
- Custom File Formats
- Cytoband
- FASTA
- GCT
- genePred
- GFF/GTF
- GISTIC
- Goby
- GWAS
- IGV
- LOH

LOH

- MAF (Multiple Alignment Format)
- MAF (Mutation Annotation Format)
- Merged BAM File
- MUT
- narrowPeak
- PSL
- RES
- SAM
- Sample Information
- chrom.sizes
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- VCF
- WIG

输入的格式

- 有几个格式，可能与我们关联很强。

BAM/Sam # 比对文件

TDF # BAM的精简版

BED # 注释文件

GFF/GTF # 注释文件

PSL # blat比对的结果

VCF # SNP、indel信息

WIG # UCSC数据库的推荐格式

IGV # IGV默认的格式

BAM

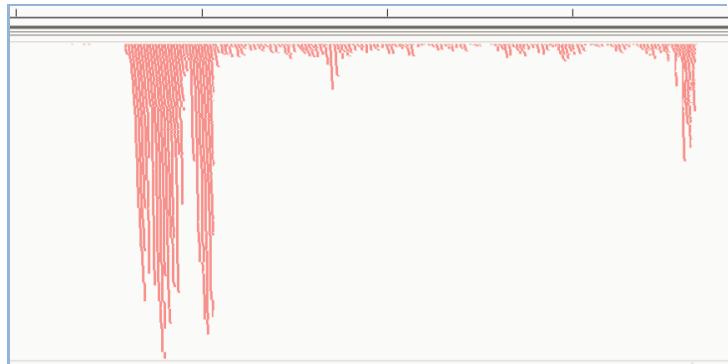
- 序列比对的直接输出结果是BAM或SAM
SAM属于文本格式，简单点说来，就是告诉我们reads比对在基因组上哪个位置，mismatch是多少；

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

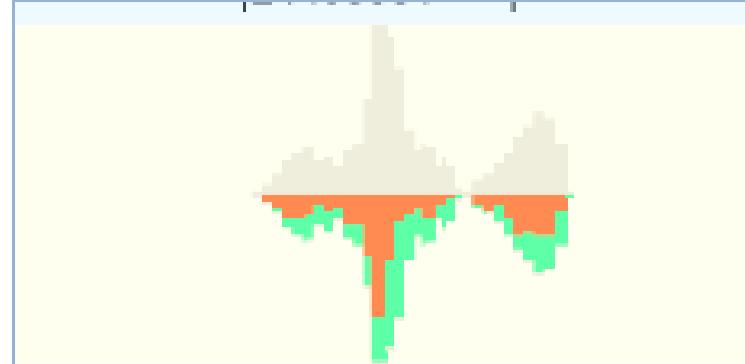
但文本文件太大，不易存储，因此一般会将SAM转化压缩为(工具：samtools) BAM文件（二进制）。

TDF

- 因为BAM文件依然保持着每一条reads的信息，信息量依然过大；
- 但绘图其实只需要保存局部的深度信息就足够；
- 所以，可见将局部深度、正负链等信息，进一步整合到简化到TDF格式的二进制文件中。



BAM格式包含每一条reads信息



TDF格式以一定区间为单位存储信息

BED

- BED常见的格式比较简单，文本文件，告诉我们一段区域的位置信息。

```
browser position chr7:127471196-127495720
browser hide all
track name="ColorByStrandDemo" description="Color by strand demonstration"
visibility=2 colorByStrand="255,0,0 0,0,255"
chr7    127471196    127472363    Pos1    0    +
chr7    127472363    127473530    Pos2    0    +
chr7    127473530    127474697    Pos3    0    +
chr7    127474697    127475864    Pos4    0    +
chr7    127475864    127477031    Neg1    0    -
chr7    127477031    127478198    Neg2    0    -
chr7    127478198    127479365    Neg3    0    -
chr7    127479365    127480532    Pos5    0    +
chr7    127480532    127481699    Neg4    0    -
```

GFF/GTF

- GFF/GTF都是基因组的注释文件，告诉你这个区域是什么（基因、重复序列、lncRNA），坐标是多少；

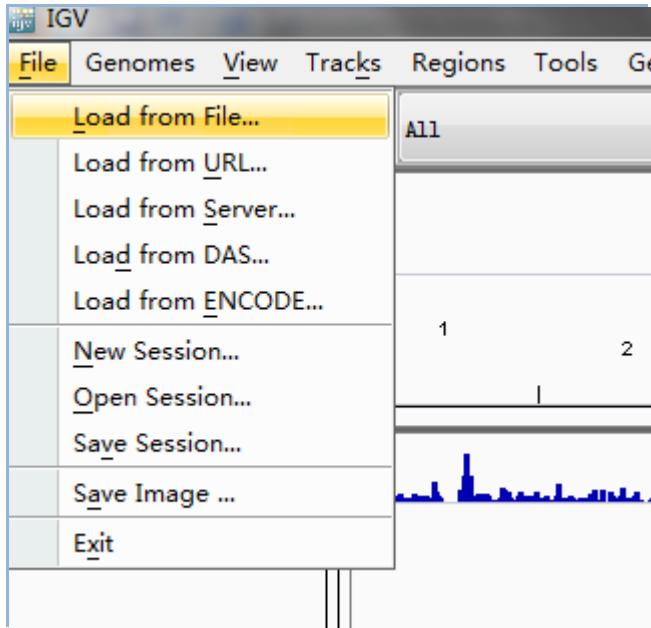
```
browser position chr22:10000000-10025000
browser hide all
track name=regulatory description="TeleGene (tm) Regulatory Regions"
visibility=2
chr22 TeleGene enhancer 10000000 10001000 500 + . touch1
chr22 TeleGene promoter 10010000 10010100 900 + . touch1
chr22 TeleGene promoter 10020000 10025000 800 - . touch2
```

其实和BED也没有本质区别，只是基因组注释文件，一般使用GFF或GTF而已；

其他

- PSL : Blat软件的输出结果
- VCF : SNP文件的常用格式
- WIG : Wiggle Track Format (WIG) , 是UCSC 基因组浏览器推荐的输入格式；

导入其他信息



导入文件

#1 : Load local file

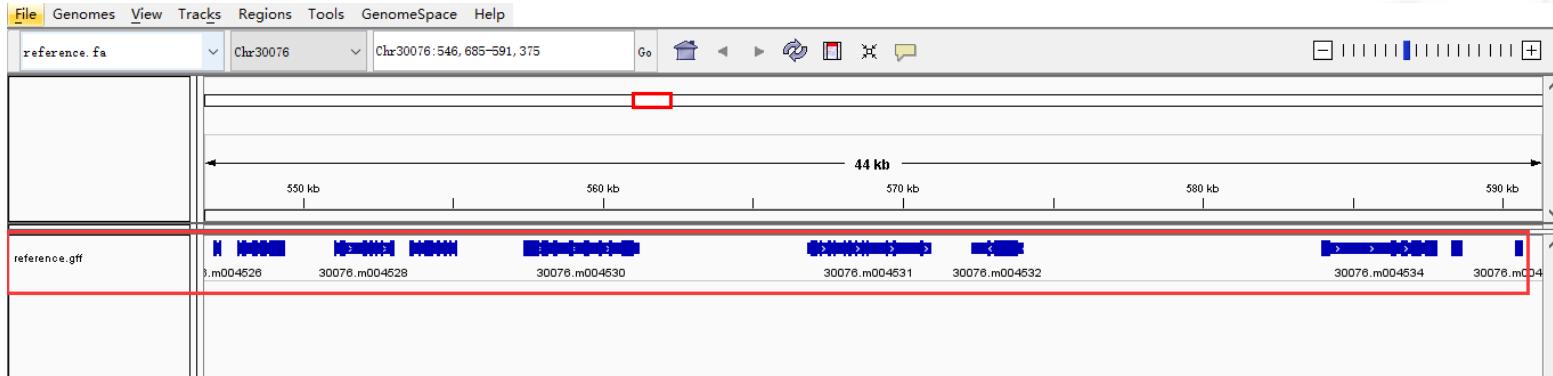
#2 : Load from URL

#3 : Load from server

(Broad IGV data server,
other data server)

- 预备生成好的BAM、GFF、TDF等文件都可以导入；
- 但记住BAM、PLS等文件，需要先排序再导入（虽然IGV软件也可以排序，但对于BAM这样的大文件，建议在超算上分染色体排序后，再在PC处理）

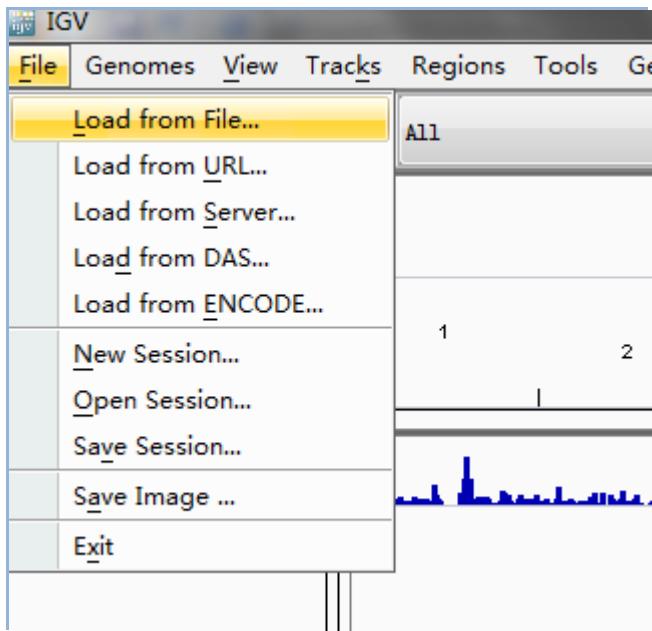
导入注释文件



- 由于导入的fa文件并没有基因组注释，这里导入gff文件进行注释。

导入步骤：file → load genome from file → 选择“reference.gff”文件

导入网络数据信息



导入文件

#1 : Load local file

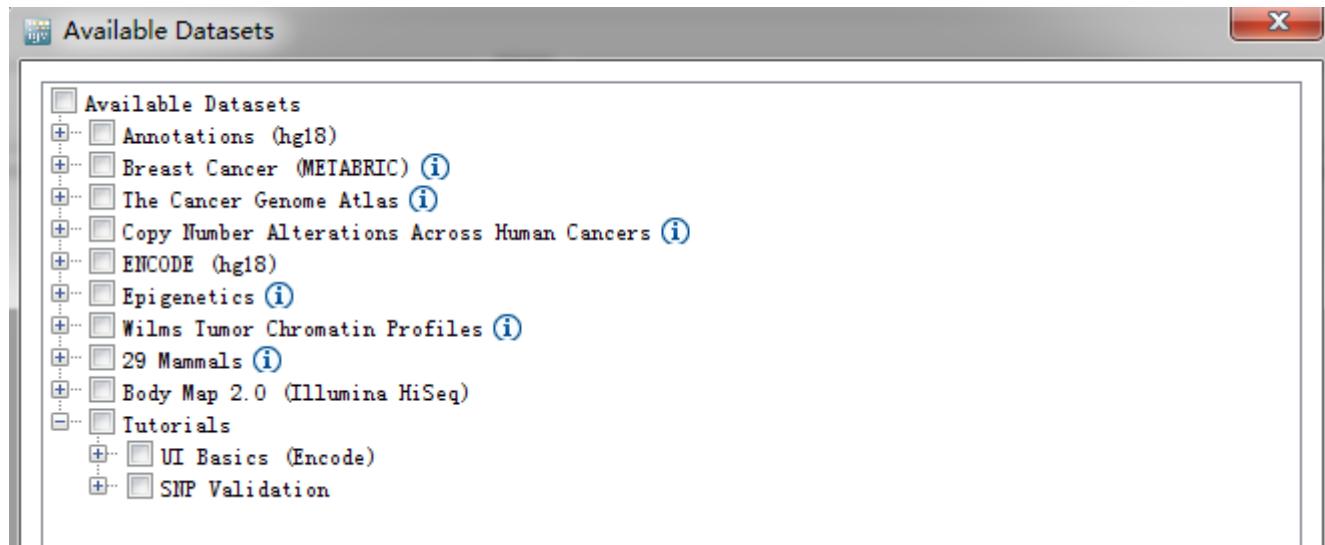
#2 : Load from URL

#3 : Load from server

(Broad IGV data server,
other data server)

- 如果是模式物种，可以从IGV的官网下载相关的信息。

“Load from server”



可以选择的数据决定于：

- (1) 你选择的服务器，默认是 Broad server
- (2) 你选择的参考基因组

“Load from server”



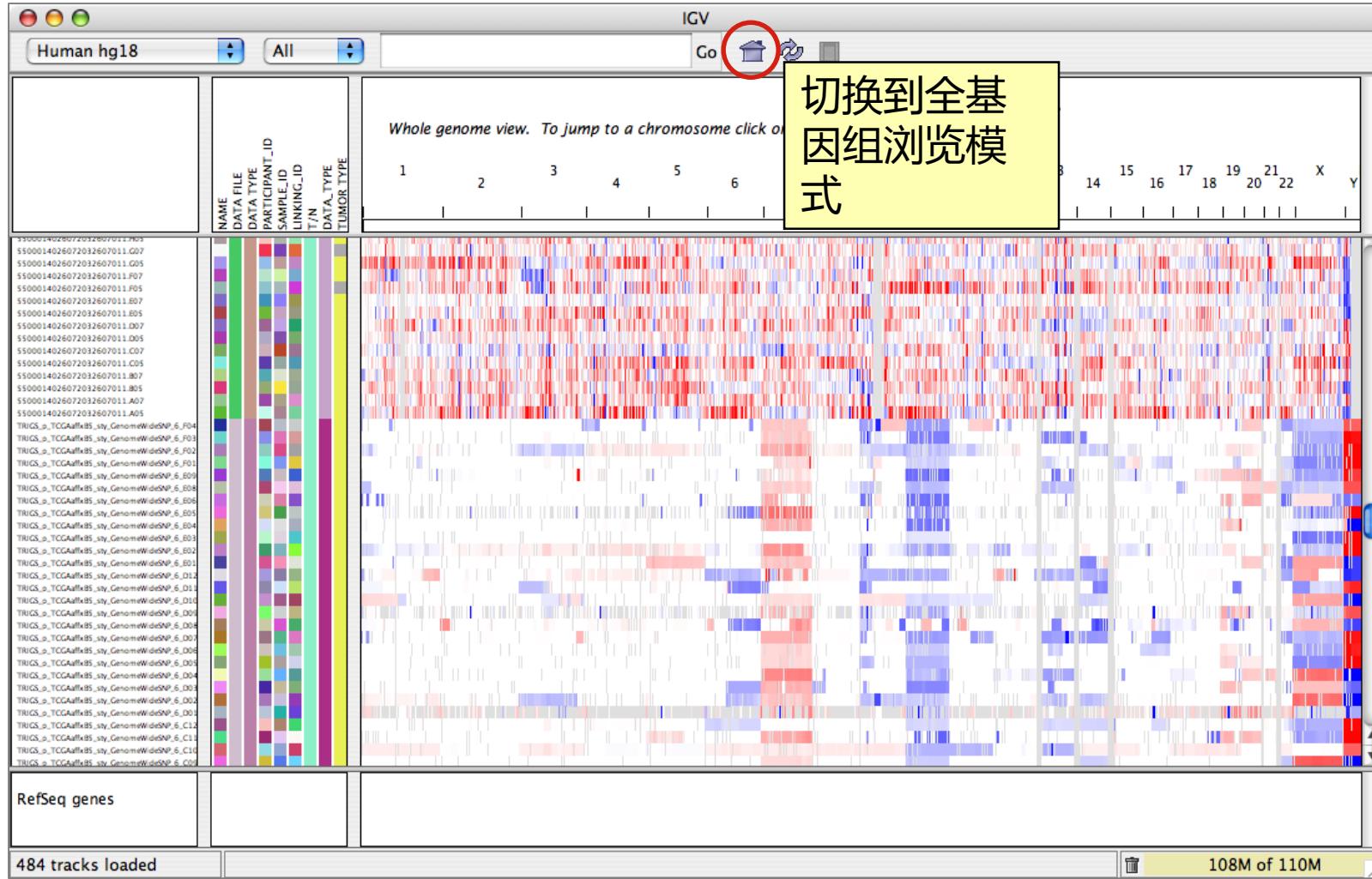
Available Datasets

- + Annotations (hg18)
- + Breast Cancer (METABRIC) (i)
- The Cancer Genome Atlas (i)
 - + Ovarian (i)
 - GBM Subtypes (Verhaak et. al.) (i)
 - Sample Information
 - Segmented Copy Number (Broad Affy 6.0)
 - Expression
 - Somatic Mutations
 - + ICGA Broad GDAC (i)
 - + GBM (Pilot Project)
- + Copy Number Alterations Across Human Cancers (i)
- + ENCODE (hg18)
- + Epigenetics (i)
- + Wilms Tumor Chromatin Profiles (i)
- + 29 Mammals (i)
- + Body Map 2.0 (Illumina HiSeq)
- + Tutorials

(3) 数据浏览, 确定目标区域



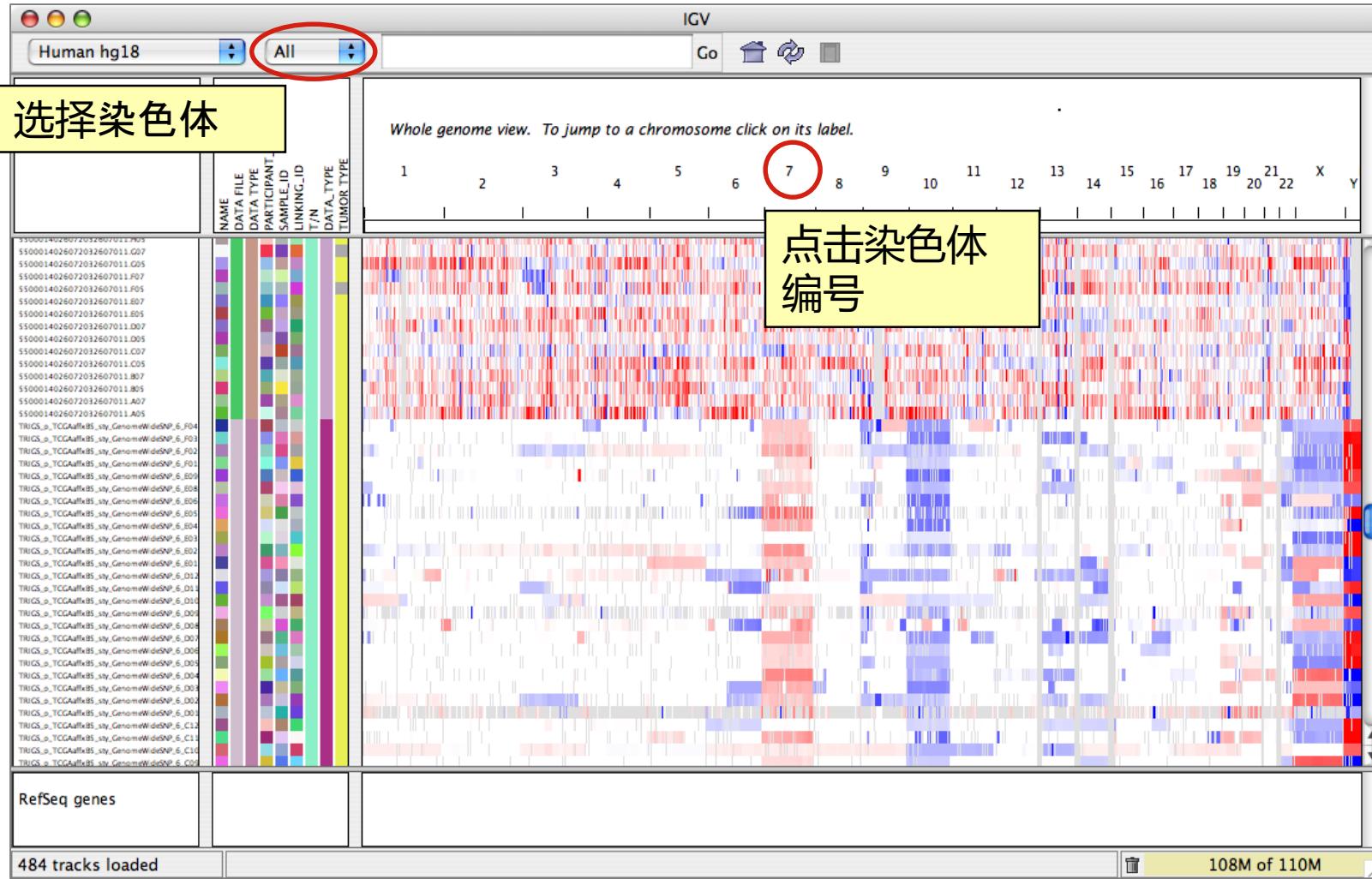
Whole genome view



数据浏览



缩放到染色体水平



数据浏览



Integrative
Genomics
Viewer

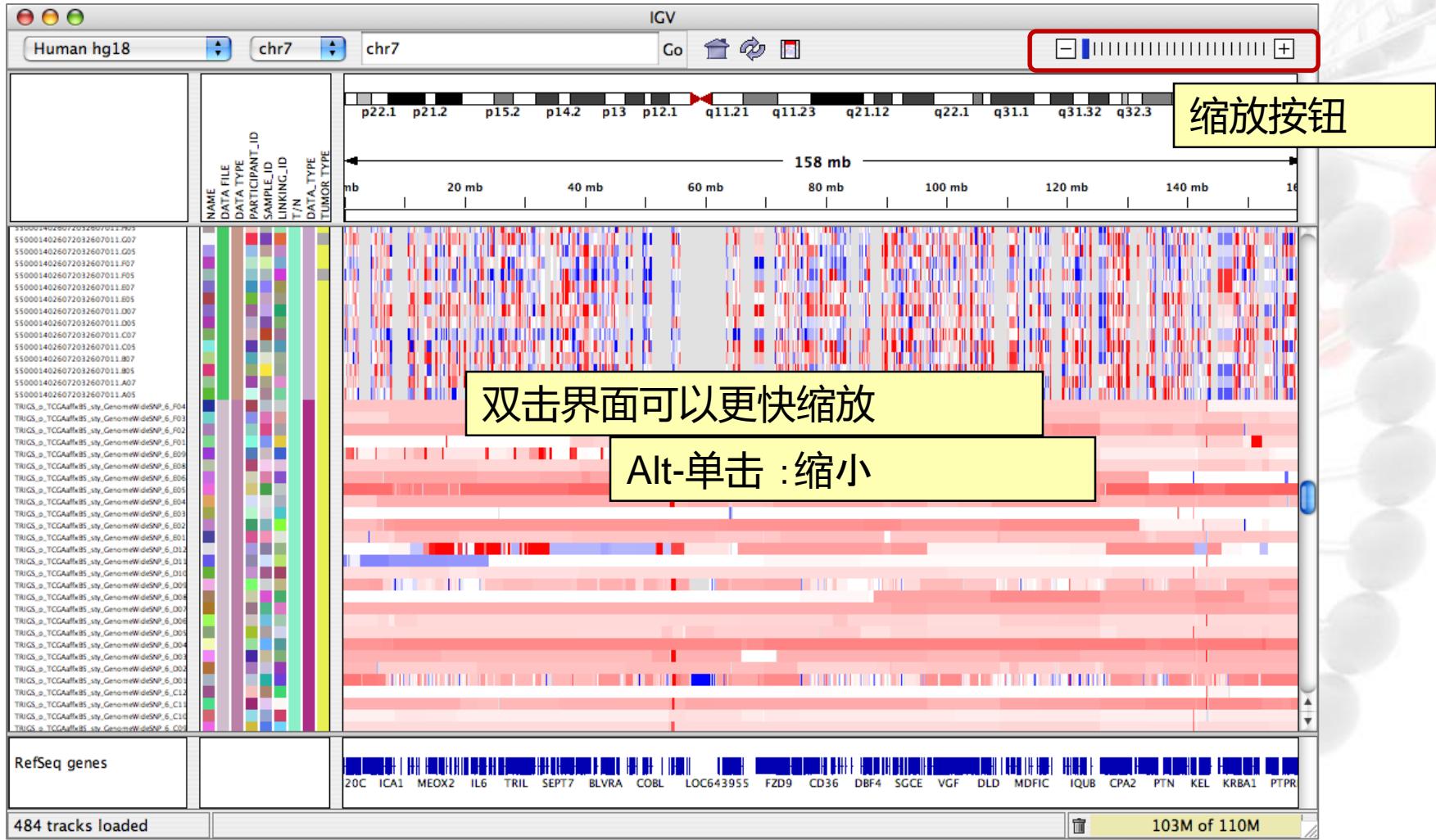
缩放到染色体水平



数据浏览



进一步放大

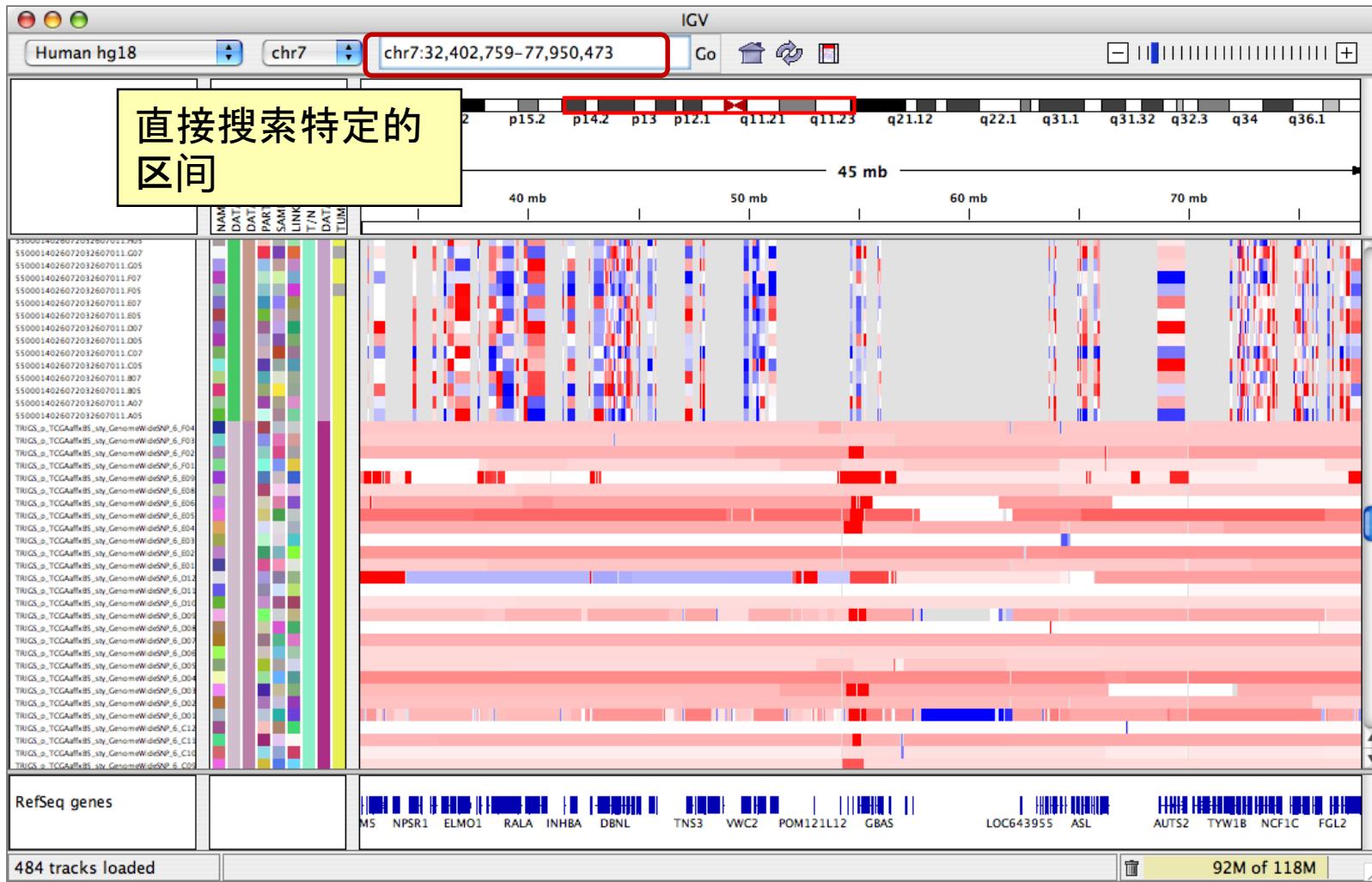


数据浏览



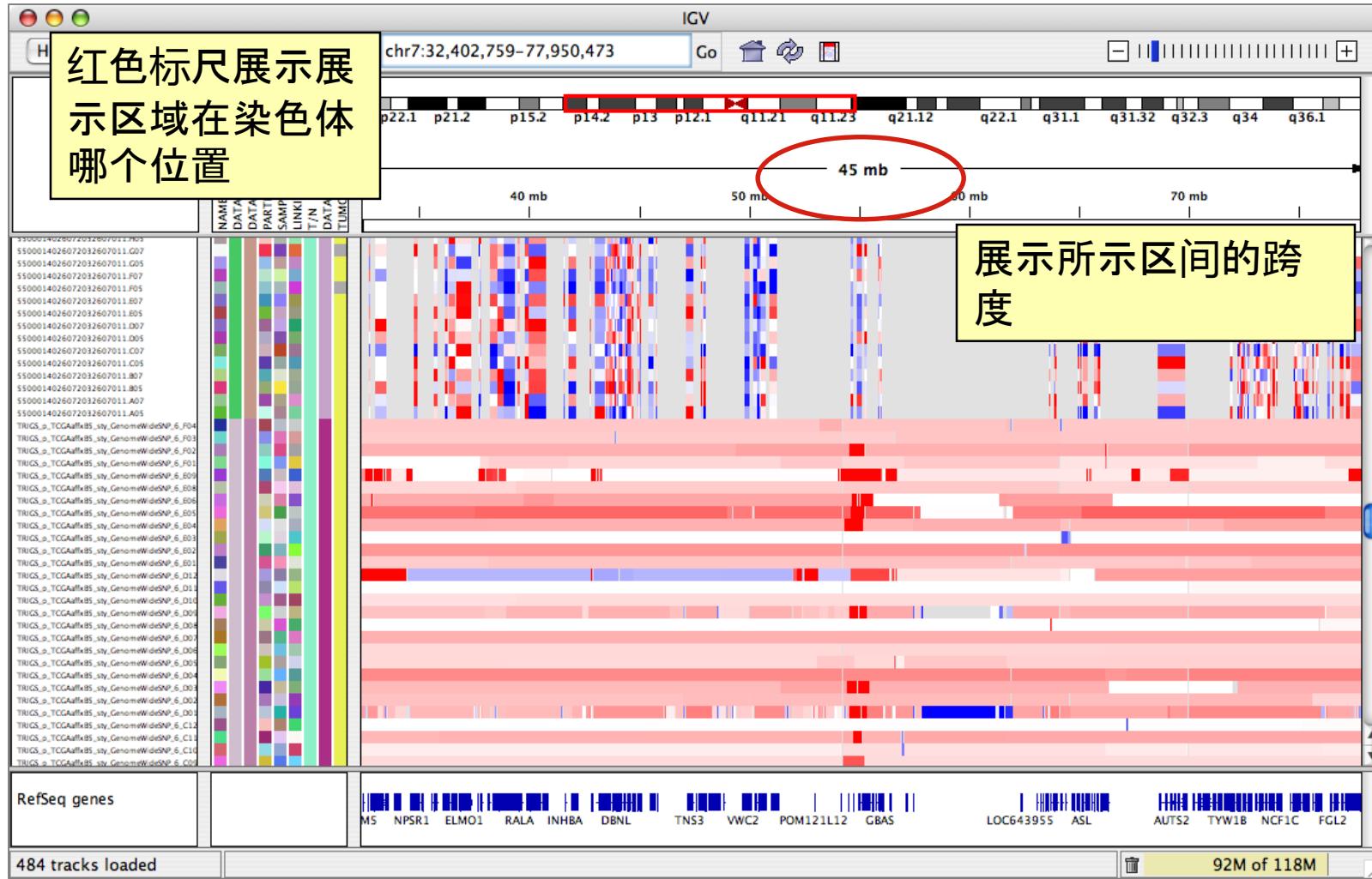
Integrative
Genomics
Viewer

进一步放大



数据浏览

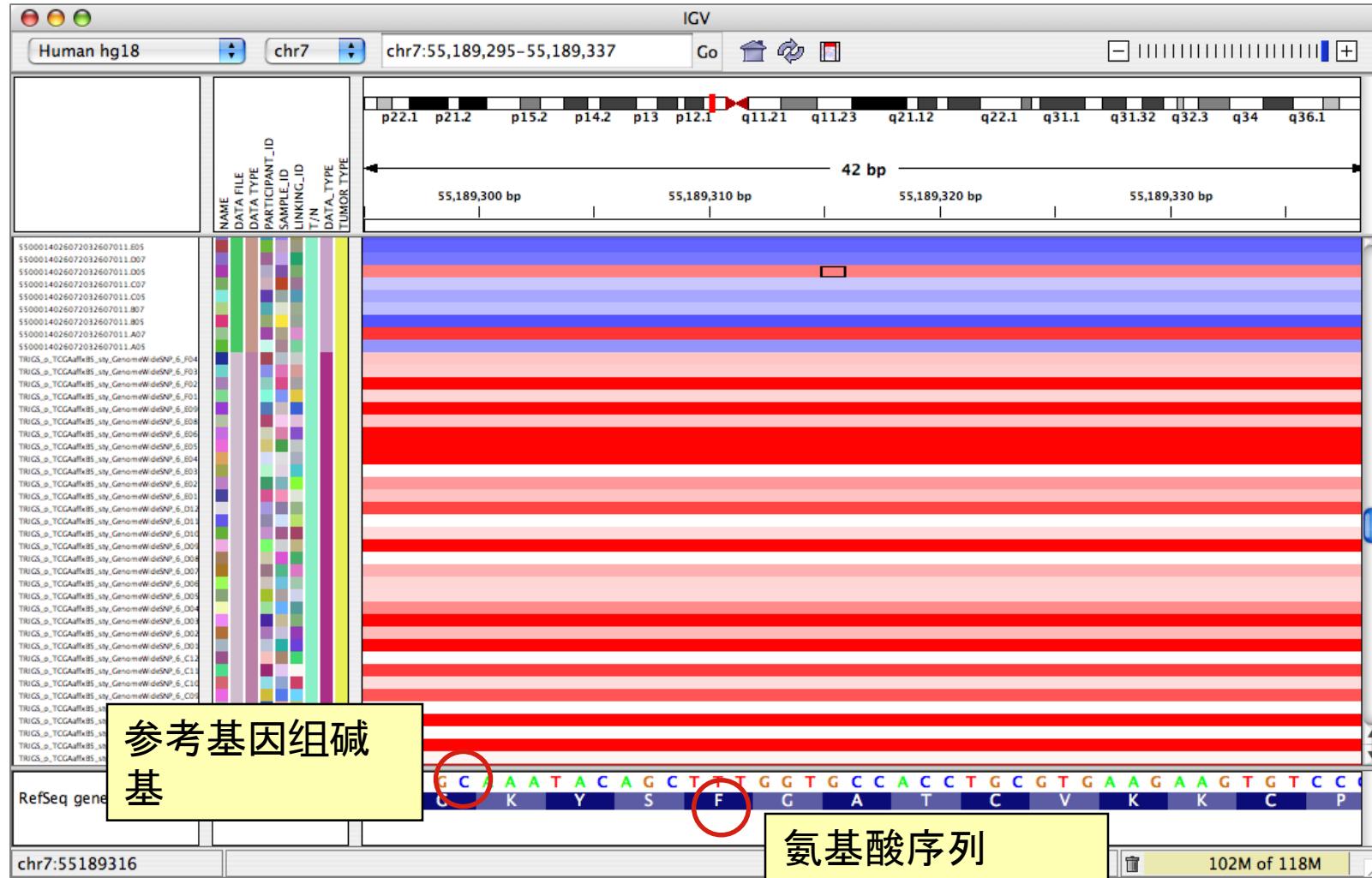
进一步放大



数据浏览



放大到单个碱基水平

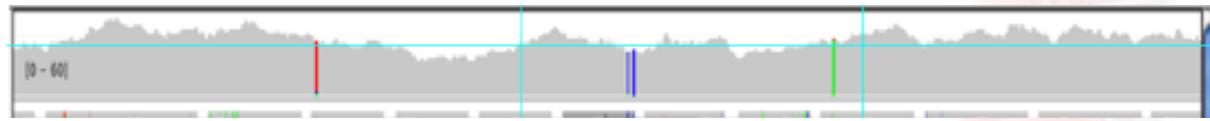
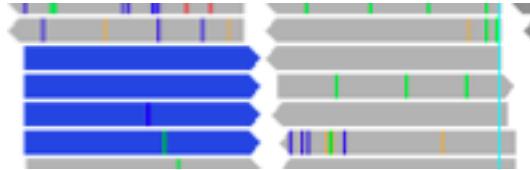


缩放大小与分辨率

全染色体—概要信息, e.g. 如覆盖度



~ 50-100 kb -- 局部信息, 如基因表达, 染色体结构变化, SNPs等



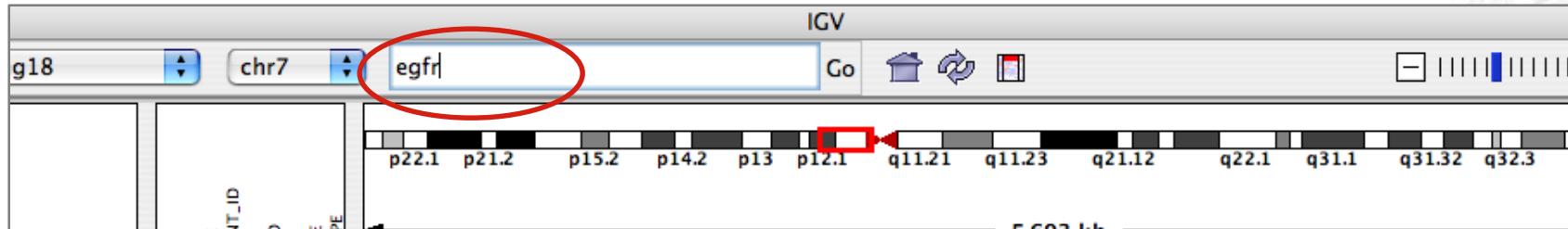
~ 500 bp – 单碱基水平



数据浏览

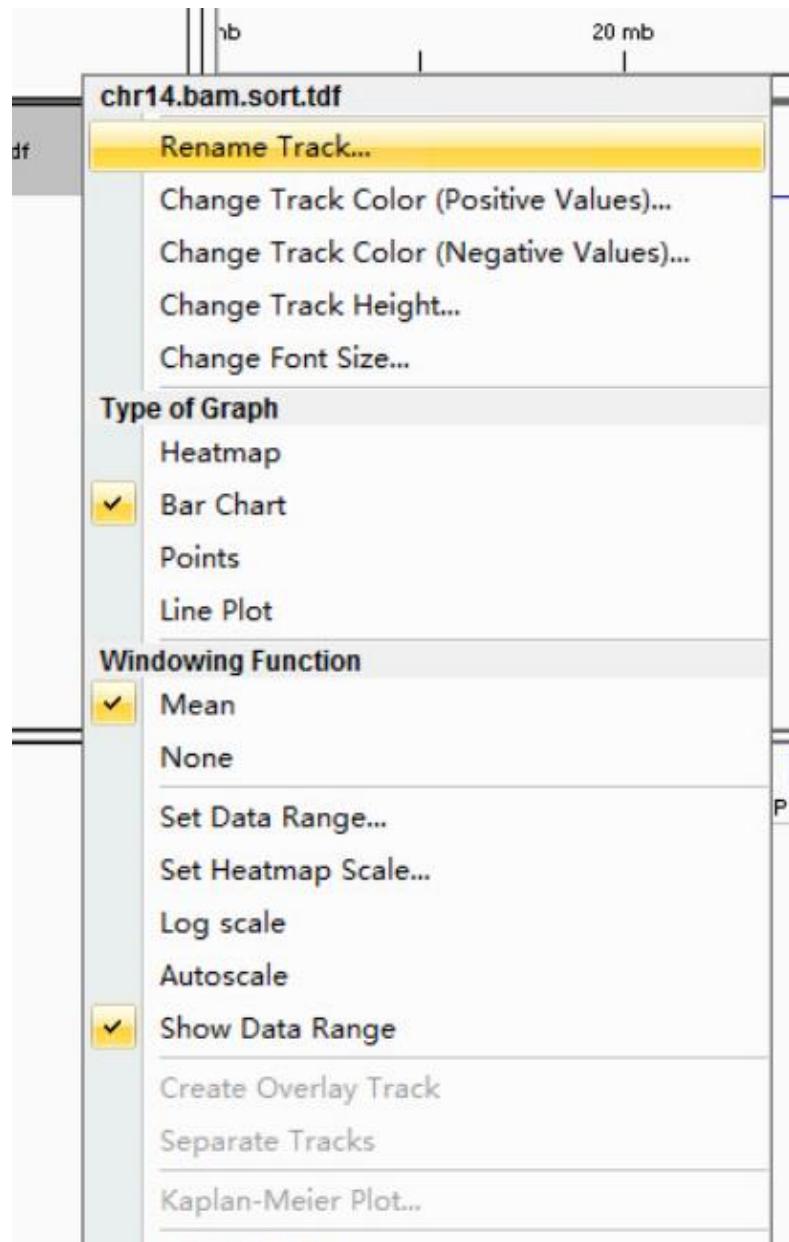


直接跳跃到某个基因的位置(填写基因ID)



- 在检索栏可以输入基因的ID来搜索对应的区域。
 - With or without zoom (View > Preferences > General)
- Click on a feature track (e.g. gene track, BED, GFF)
 - Ctrl+F = jump forward to next feature
 - Ctrl+B = jump backward to previous feature

(4) 设置track 属性

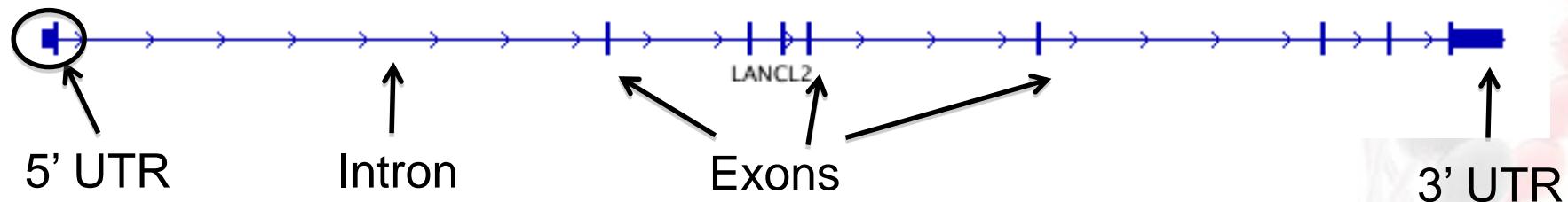


右键点击track,出现的下拉菜单可以显示选择项：

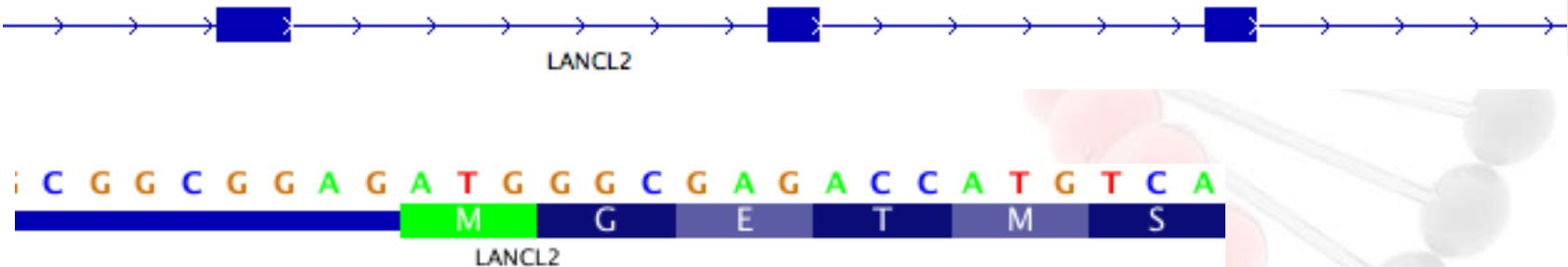
1. Track颜色，高度等；
2. 呈现形式：热图、柱形图等；
3. 取值的方式（均值，标准化等）
4. 删除、保存图片

注释的track

基因注释



Zoomed in views

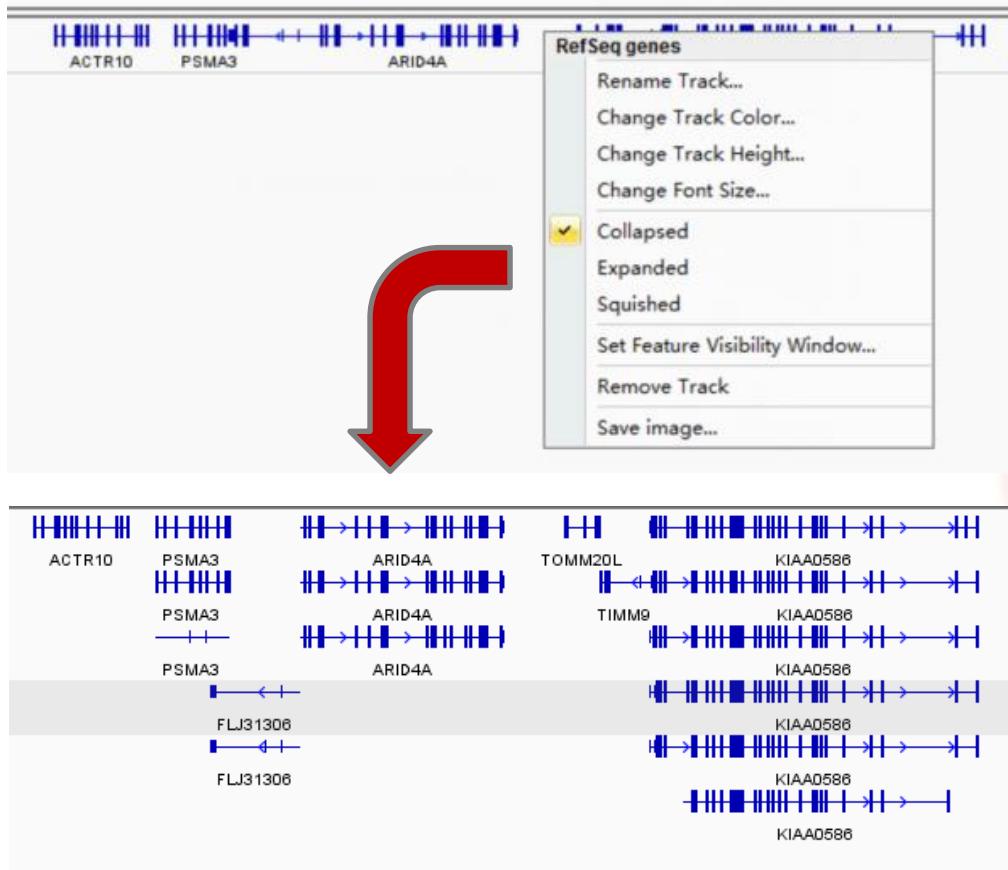


注释的展示模式

1. 注释的track, 默认是单行显示



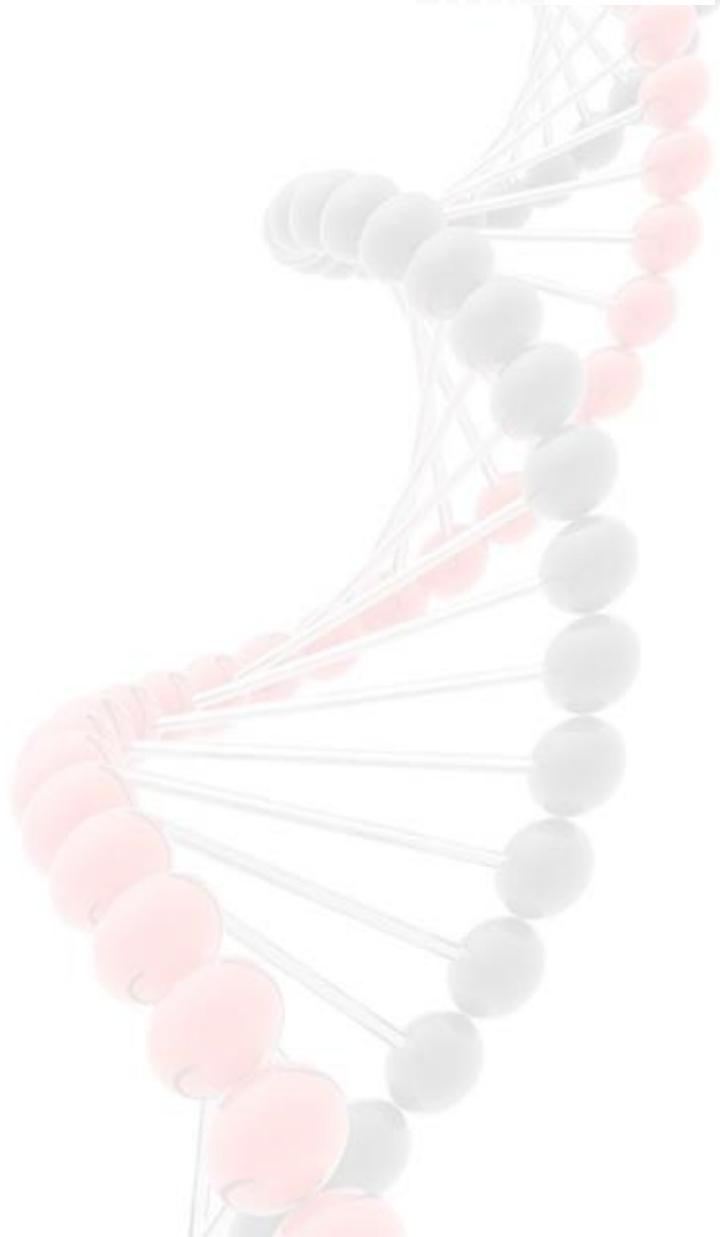
2. 右键点击, 选择“Expanded”或“squished”可以展开track。



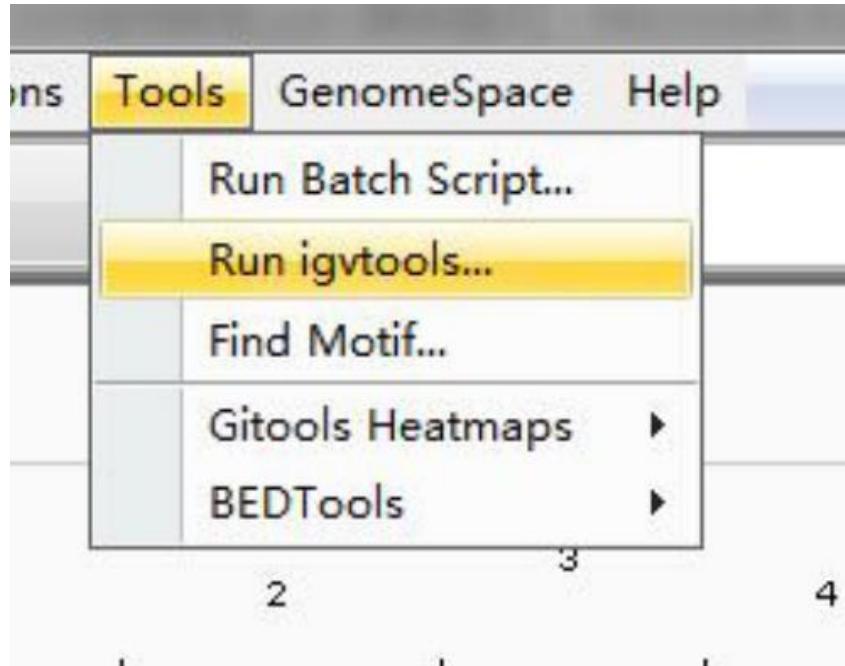
提纲



- 软件介绍与安装启动
- 数据导入与文件格式介绍
- IGV tools
- 数据练习



某些文件输入前需要预处理，以便IGV可以正确读取。



有4个功能。

count: 计算将已排序的比对文件，转化为包含深度信息的TDF文件。

支持的输入格式 : sam, **.bam**, .aligned, .sorted.txt, .bed

备注：count的输出文件也是TDF格式的。这个命令主要针对比对文件设计。

sort: 将输入注释或比对文件，根据行的起始位点排序。

支持的输入格式 .cn, .igv, .sam, .aligned, and .bed.

index: 给输入文件加索引。

支持的输入格式: .sam, .aligned, .sorted.txt

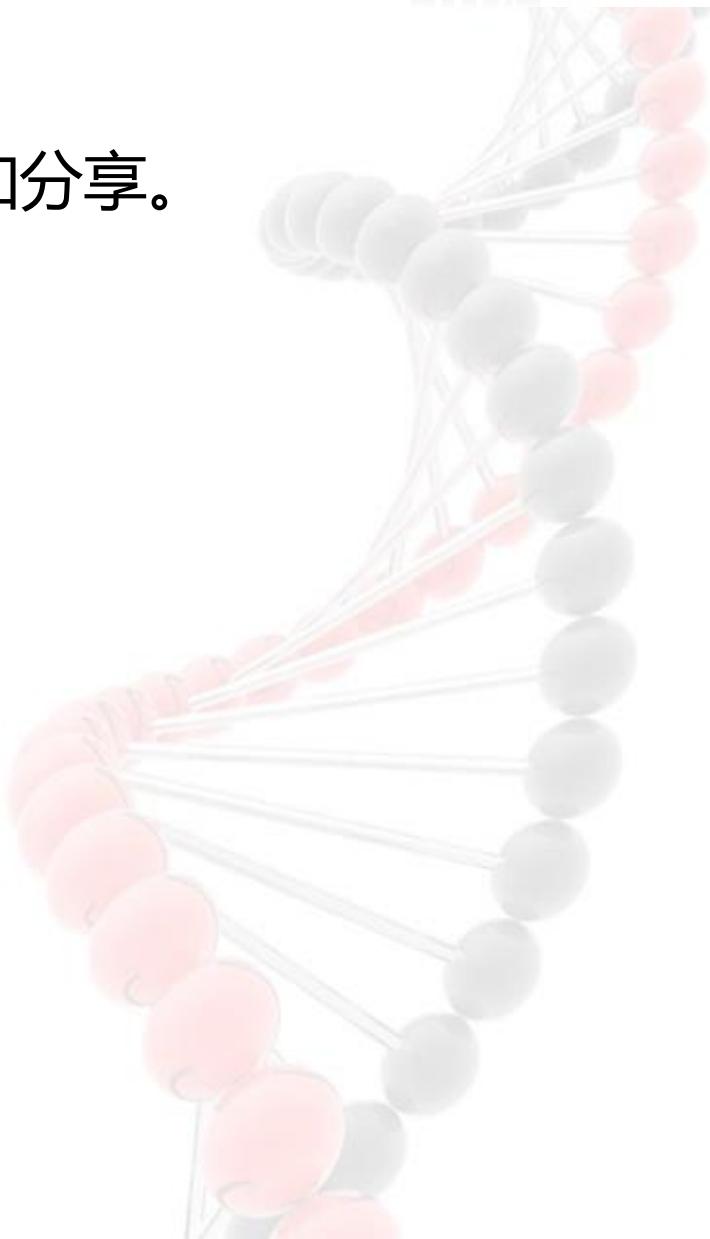
toTDF: 将sorted后的文件，转为二进制的TDF文件 (.tdf)

支持的输入格式: .wig, .cn, .snp, .igv, .gct

TDF的优势

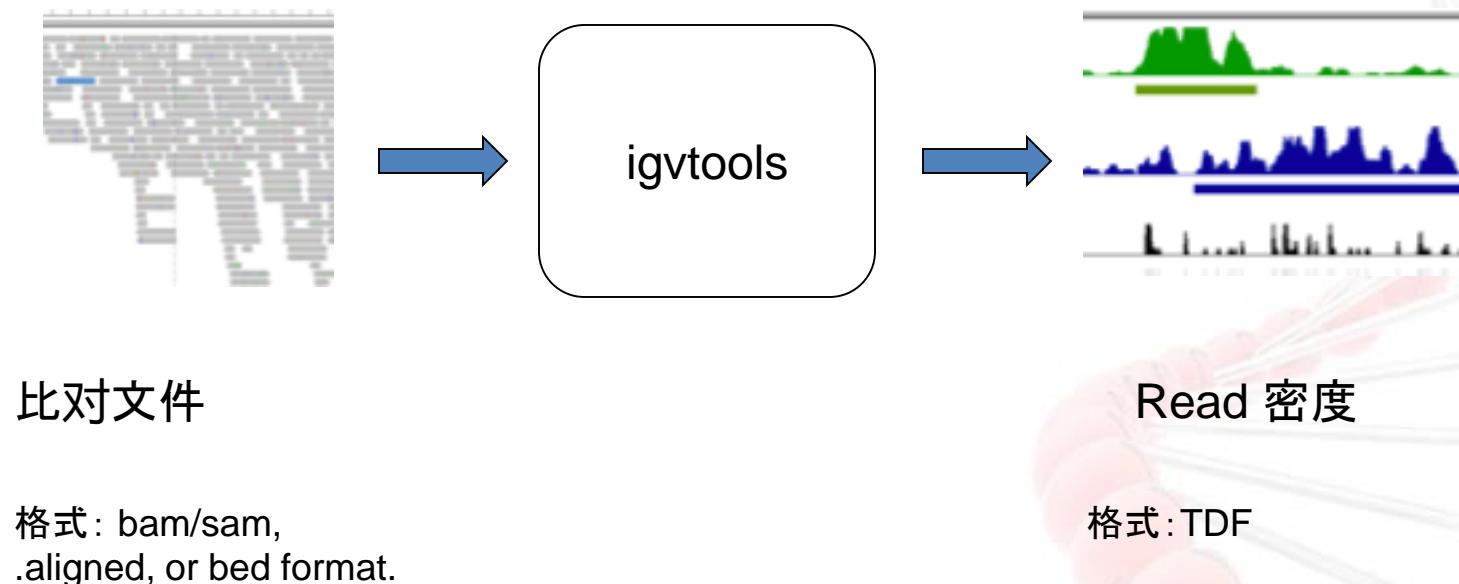


1.数据被压缩，文件更小，便于导入和分享。



Count 得到的TDF文件

Count命令主要用于比对文件转化为只含有密度信息的TDF文件,. 例如 ChIP-Seq数据、RNA-Seq 等.
这个命令才是对我们有用命令。



IGVTools index



可以给注释文件加索引。

如果是比对文件，只能对SAM文件加索引。(not BAM)

Note: 不要与 *samtools* index 混淆。Samtools用于给 BAM 文件加索引。

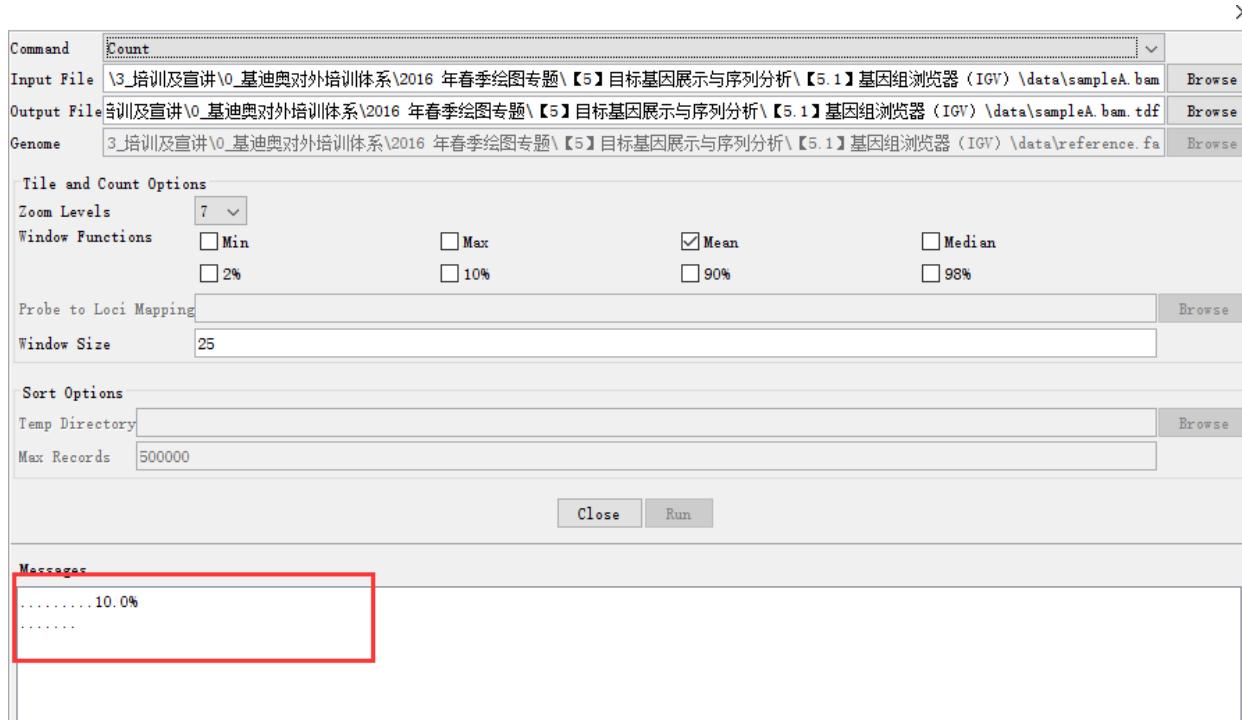
SAM => igvtools

BAM => samtools

演示



这里我们给出了这个两个样本比对结果的bam文件，只有count命令可以处理（其他命令只能处理sam文件）

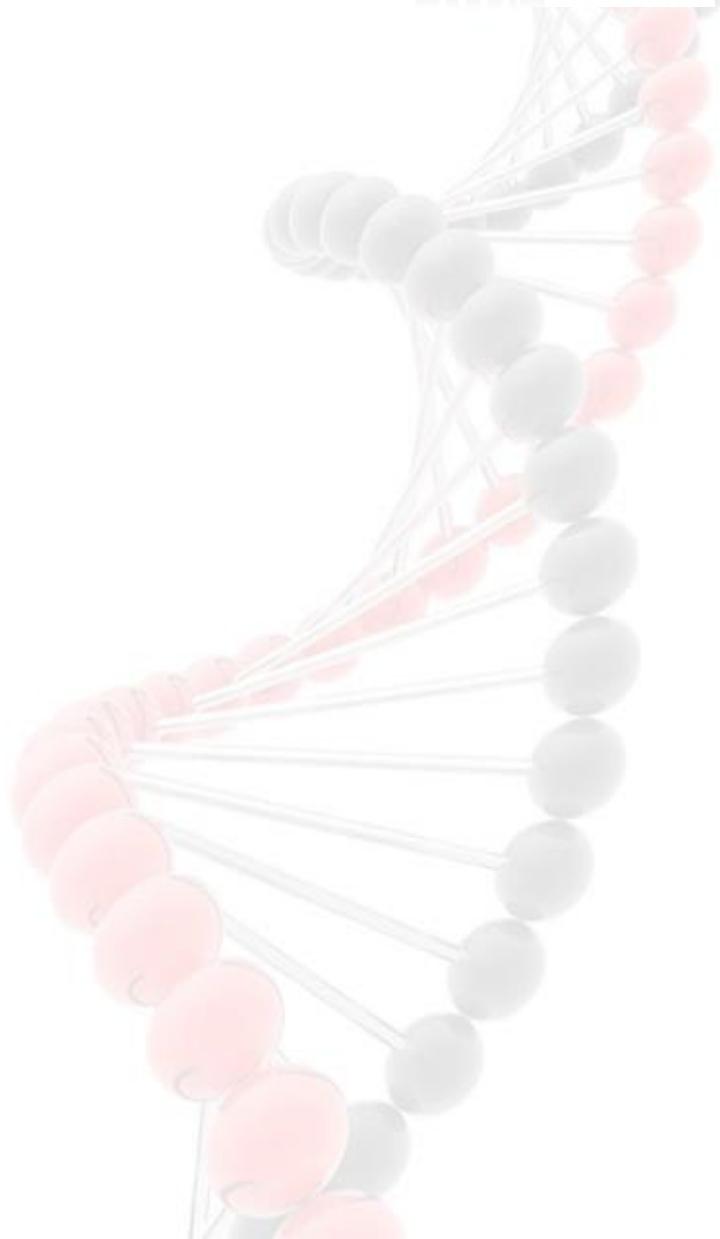


填写好输入文件的bam文件，按照默认设置，就可以输出TDF文件了

提纲



- 软件介绍与安装启动
- 数据导入与文件格式介绍
- IGV tools
- **数据练习**



样本信息与操作



1. 参考序列 reference.fa (文件夹中有)
2. 其他文件
 - 2.1 注释文件 (gtf文件)
 - 2.2 比对结果 (bam文件)
 - 2.3 甲基化数据 (igv文件)
3. 找到目标基因
4. 优化图片

名称	修改日期
reference.fa	2016/3/23 22:
reference.gff	2016/3/23 22:
sampleA.bam	2016/3/23 10:
sampleA.igv	2016/3/23 22:
sampleB.bam	2016/3/23 10:
sampleB.igv	2016/3/23 22:

操作步骤



1. 导入参考基因组

Genomes → load genome from file → 选择
“reference.fa” 文件

2. 导入注释文件

file → load genome from file → 选择
“reference.gff” 文件

备注：由于gff文件已经是按坐标顺序排列的，因此
不需要重新排序。

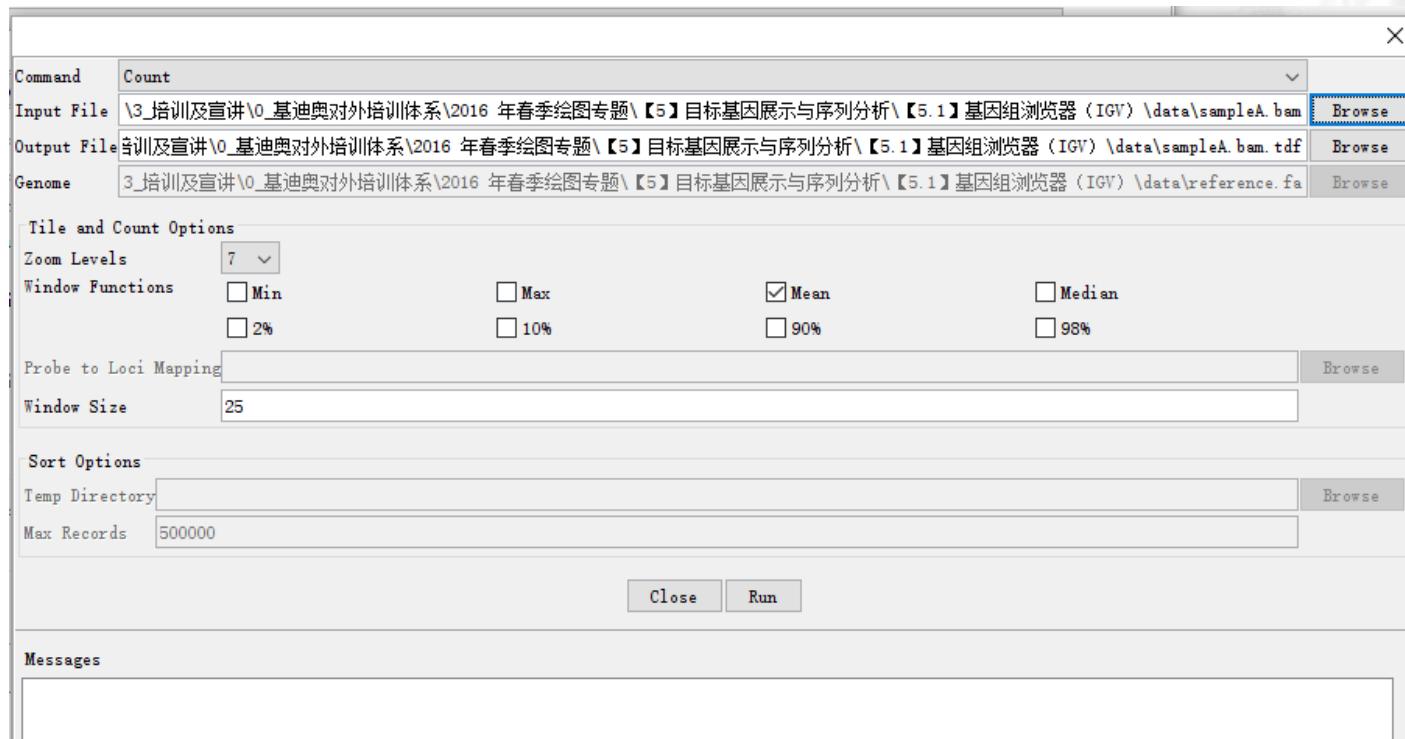
操作步骤



3. 将bam文件转化为TDF文件

Tools → run IGV tools → command界面内选择

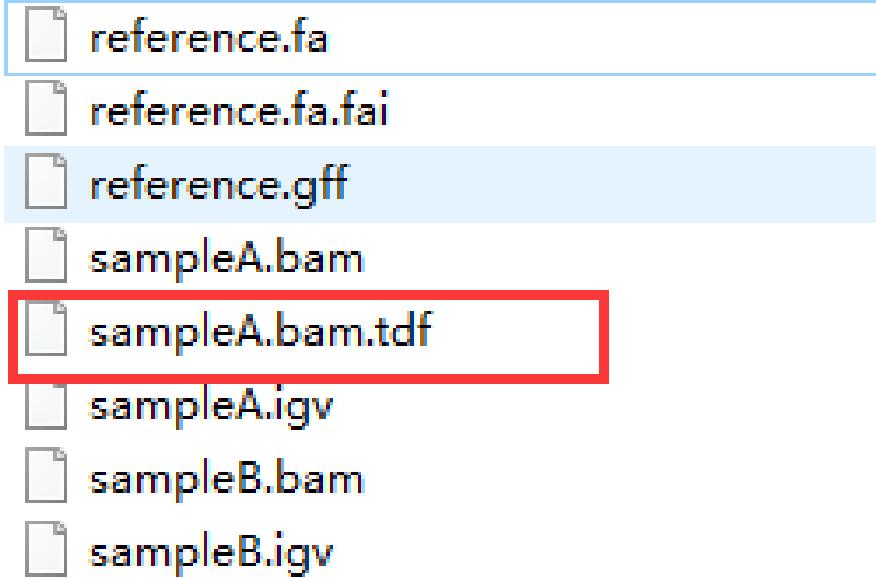
“count” → input界面选择输入bam文件→ 点击run



操作步骤

3. 将bam文件转化为TDF文件

- 目录下多出tdf文件。依此生成另一个样本的tdf文件。
- 然后通过file → load from file 导入。



操作步骤



3. 输入甲基化数据 (IGV格式)

建议先打开IGV文件（使用notepad ++），理解数据的内容。

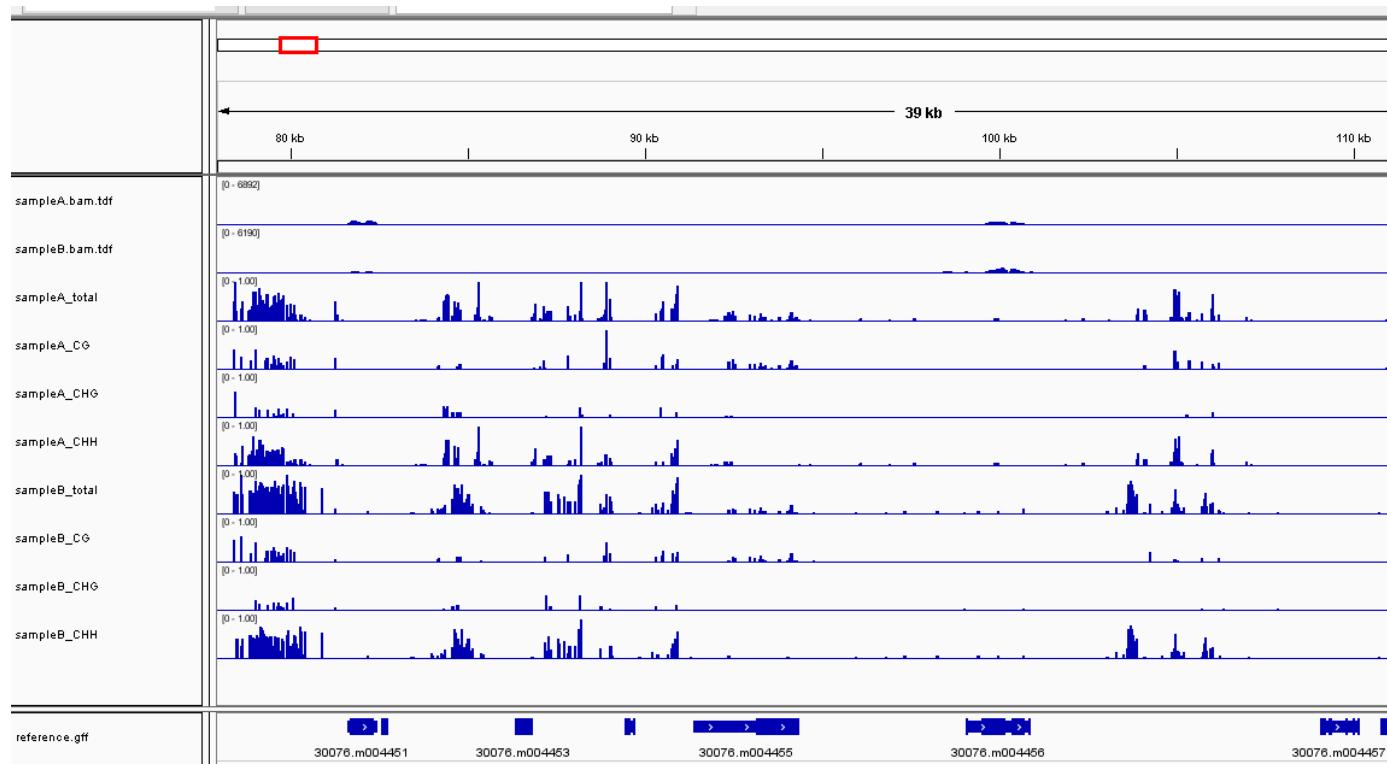
1	Chromosome	Start	End	Feature	sampleA_total	sampleA(CG)	sampleA(CHG)	sampleA(CHH)
2	30076	3	3	methyl	0.00	0.00	0.00	0.00
3	30076	7	7	methyl	0.00	0.00	0.00	0.00
4	30076	14	14	methyl	0.00	0.00	0.00	0.00
5	30076	18	18	methyl	0.00	0.00	0.00	0.00
6	30076	19	19	methyl	0.000	0.00	0.000	0.00
7	30076	21	21	methyl	0.00	0.00	0.00	0.00
8	30076	24	24	methyl	0.00	0.00	0.00	0.00
9	30076	25	25	methyl	0.00	0.00	0.00	0.00
10	30076	30	30	methyl	0.00	0.00	0.00	0.00
11	30076	41	41	methyl	0.00	0.00	0.00	0.00

文件包含这个scaffolding 所有C位点的甲基化率。后四列，分别是所有C, CG位点, CHG位点, 和CHH位点的甲基化率。因为，在植物中非CG的位点甲基化，依然起到很重要的作用。

操作步骤

3. 输入甲基化数据 (IGV格式)

然后通过file → load from file 导入，结果如下图：



目标基因查询



- 查询以下四个基因，看其有什么特点：

30076.m004707

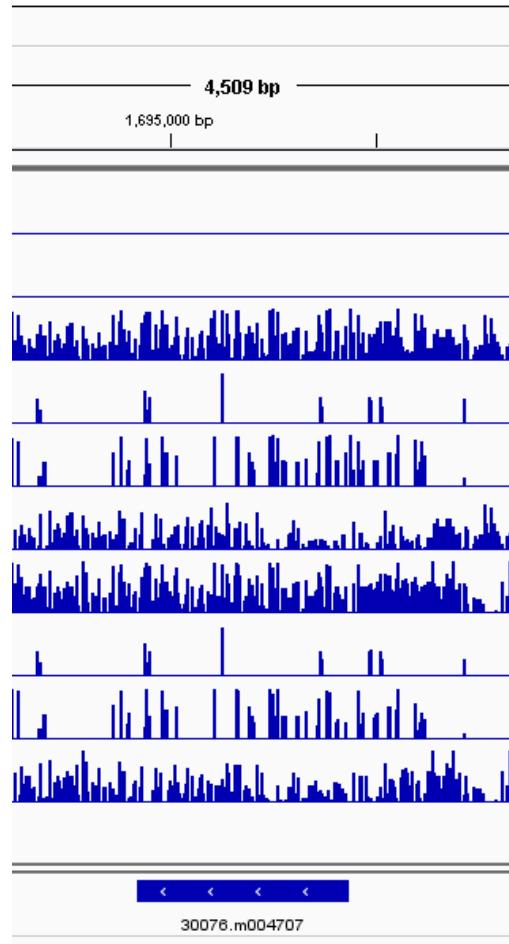
30076.m004684

30076.m004513

30076.m004451

目标基因查询

● 30076.m004707



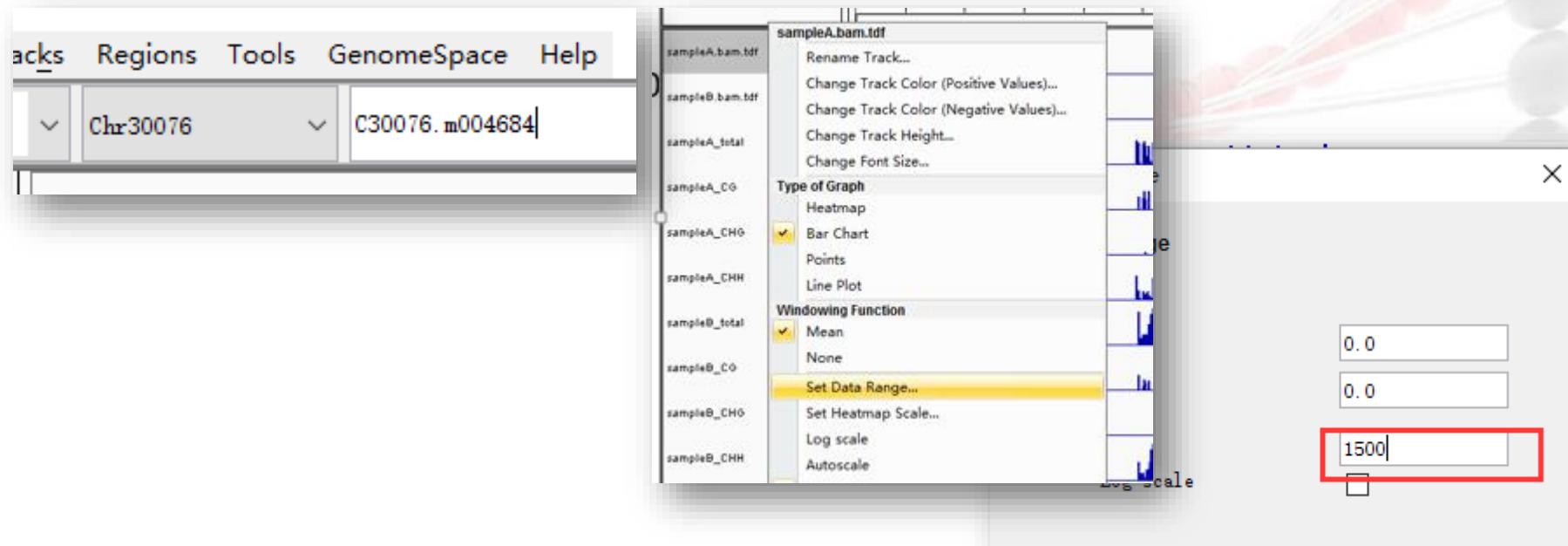
这个基因在两个样本完全不表达，
其甲基化状态如何？

目标基因查询

- 基因30076.m004684:

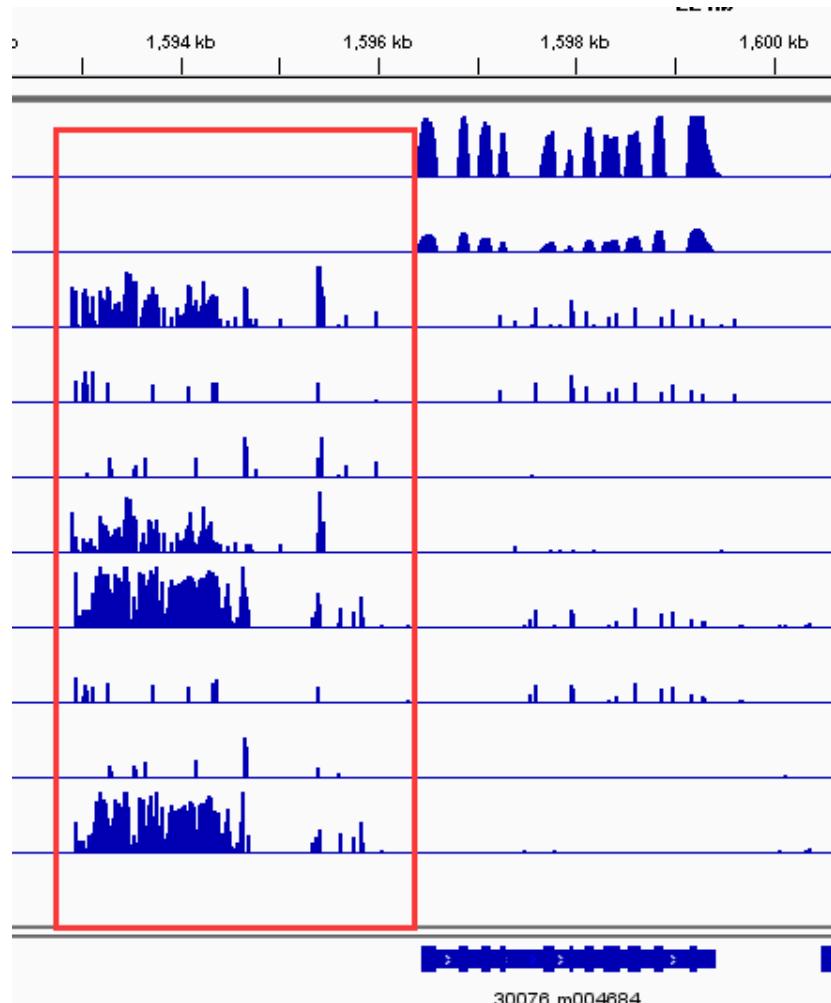
在地址栏直接填写基因名称跳到相应位置。

由于表达量过低，需要修改TDF的显示范围。右键点击sampleA.tdf.bam的名称，选择set data range，改到1500.



目标基因查询

- 基因30076.m004684:



- 这个基因在两个样本表达量是否有差异？
- 基因上游是否有甲基化的差异，属于哪种甲基化类型（CG、CHG or CHH）

目标基因查询



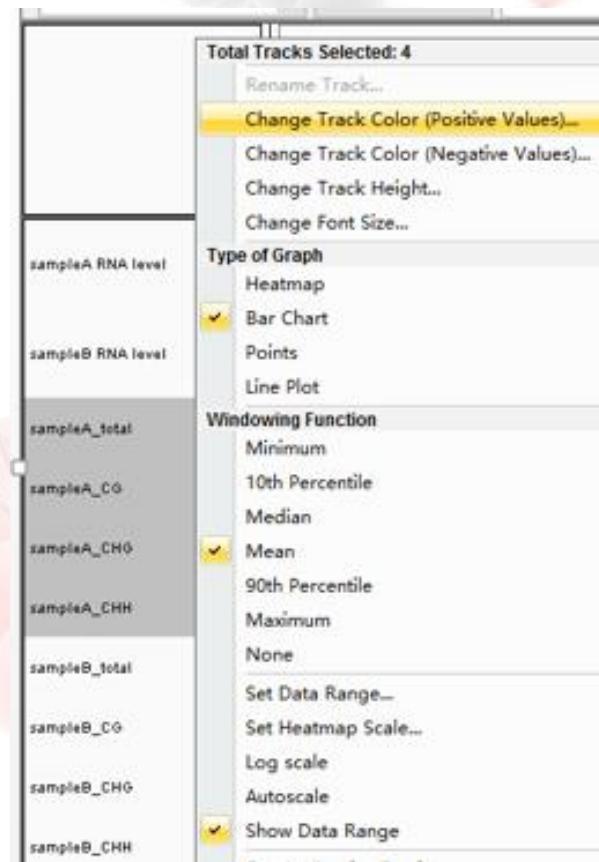
- 类似的，查询以下两个基因，其基因表达与甲基化是否潜在存在关联？（重点查看5'启动子端）

30076.m004513

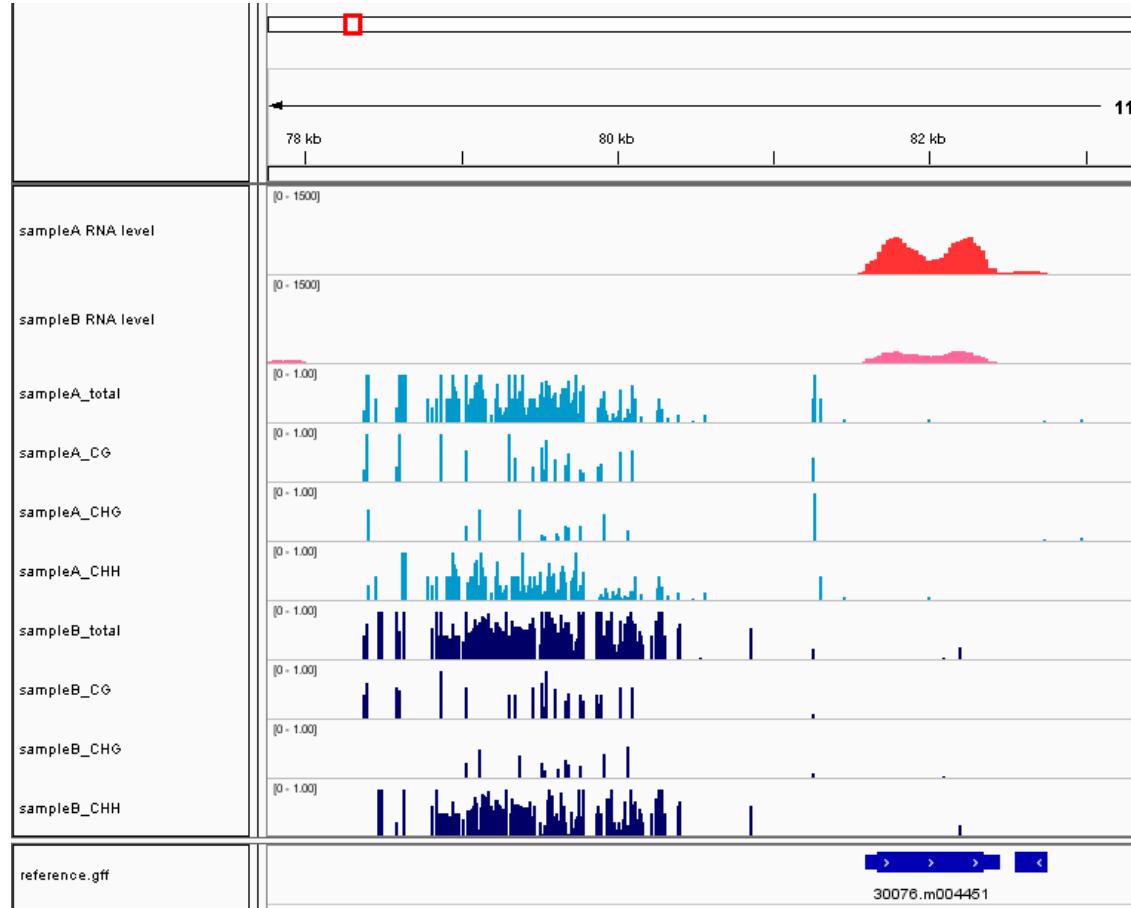
30076.m004451

修改属性

1. 找到30076.m004451这个基因。
2. Bam文件的track名称改为 “sample* RNA level” ,
track height改为60。
3. RNA表达改为红色，甲基化改为蓝色。
4. 高表达组采用深色，低表达组改用浅色。
5. 输出图片。



效果图



- 最终反映了甲基化和基因表达潜在的负调控关系。
- CHH的甲基化在这里起到主导作用。

谢谢&问题