Visualization of mapped reads

Valencia, 28-30 Sep 2015

Alejandro Alemán
aaleman@cipf.es

Genomics Data Analysis CIBERER

Outline

1) Why do we need to visualize results?

2) Different formats: BAM, SAM, BED, GFF

3) Genomics viewers: IGV, Genome Maps

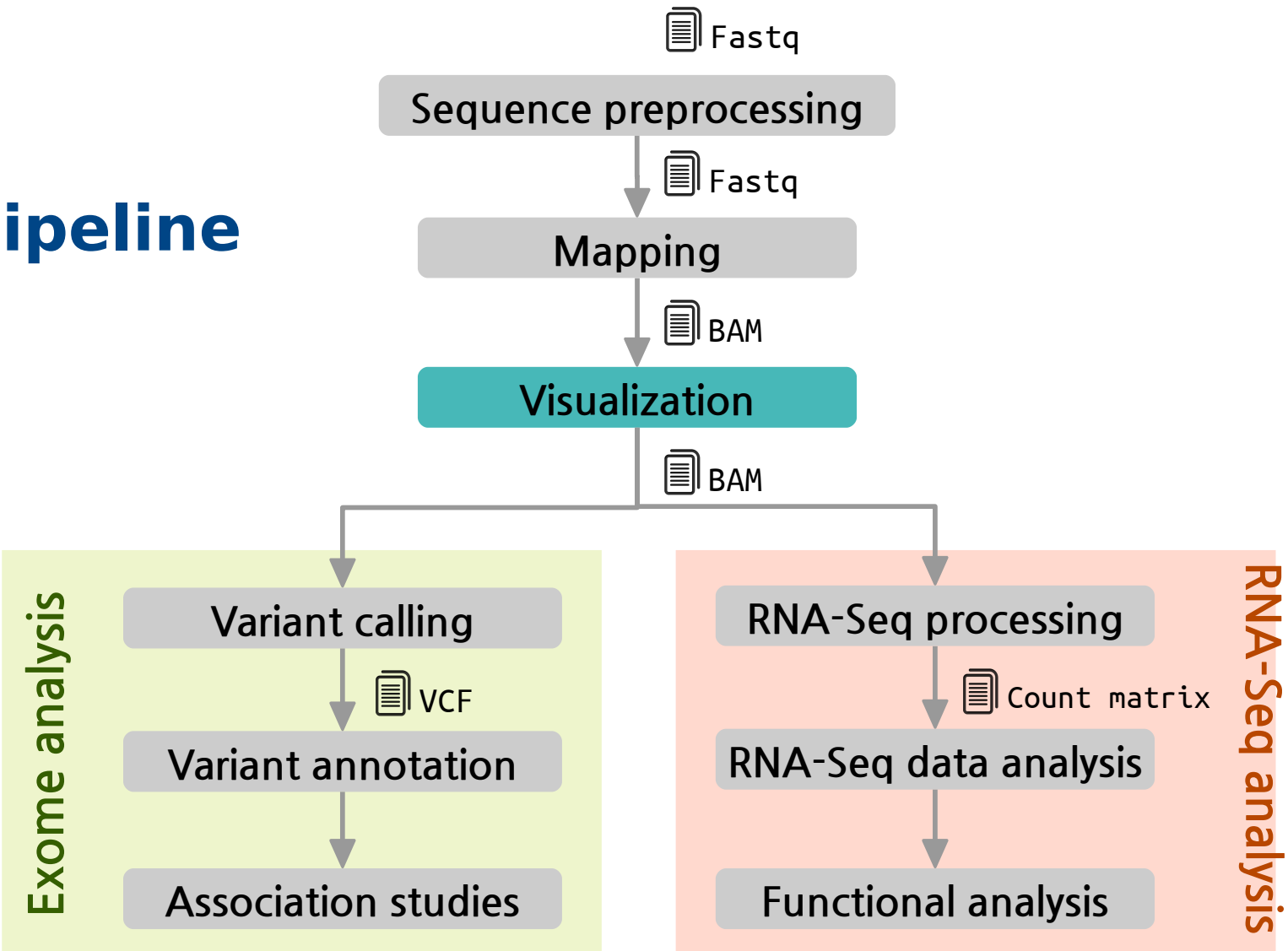
Outline

1) Why do we need to visualize results?

2) Different formats: BAM, SAM, BED, GFF

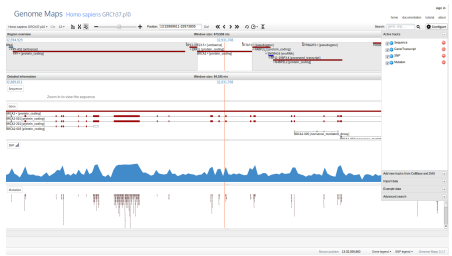
3) Genomics viewers: IGV, Genome Maps

The pipeline



Why visualization?

- **Large** quantities of genomic **data** (NGS, array based methods, etc)
- **Human interpretation** and judgment using visualization can help complex biological relationships
- Two **Genomics Viewers**:
  - Integrative Genomics Viewer (IGV)
  - Genome Maps (http://genomemaps.org/)



Outline

1) Why do we need to visualize results?

2) Different formats: SAM, BAM, BED, GFF

3) Genomics viewers: IGV, Genome Maps

SAM file format

- Text file that stores large nucleotide sequence alignments:

Header

```
@HD    VN:1.0  SO:coordinate
@SQ    SN:chr1 LN:249250621
@PG    ID:TopHat    VN:2.0.8    CL:/opt/soft/ngs/tophat/tophat-2.0.8.Linux_x86_64/tophat -p 4 -o
    /clinicfs/projects/3.ENCODE/mappings/Gm12878/Gm12878_Rp1_pair --no-coverage-search -r 300 --mate-std-dev 200 --
    library-type fr-unstranded /clinicfs/common/reference-genomes/homo_sapiens/bt2/hg19_ucsc/hg19_ucsc
    /clinicfs/projects/3.ENCODE/reads/Gm12878_Rp1_1.fastq /clinicfs/projects/3.ENCODE/reads/Gm12878_Rp1_2.fastq
```

Alignments

```
61PKHAAXX_HWUSI-EAS627_0007.68122391   337   chr1   10536   1    76M    =    173766  163
    TACCACCGAAATCTGTGCAGAGGAGAACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGC
    E5@>>>?B?A:BCB@??BAEBBCEC?@EDE@FEEFFEC=:BFFFFFAEE=EEDFFFFFDFFDFFFFEGGGDFEFFF   AS:i:-5 XN:i:0  XM:i:1
    XO:i:0  XG:i:0  NM:i:1  MD:Z:24C51    YT:Z:UU NH:i:3  CC:Z:=  CP:i:10536    HI:i:0
61PKHAAXX_HWUSI-EAS627_0007.68122391   113   chr1   10536   1    76M    chr16   90195094    0
    TACCACCGAAATCTGTGCAGAGGAGAACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGC
    E5@>>>?B?A:BCB@??BAEBBCEC?@EDE@FEEFFEC=:BFFFFFAEE=EEDFFFFFDFFDFFFFEGGGDFEFFF   AS:i:-5 XN:i:0  XM:i:1
    XO:i:0  XG:i:0  NM:i:1  MD:Z:24C51    YT:Z:UU NH:i:3  CC:Z:=  CP:i:10536    HI:i:1
```

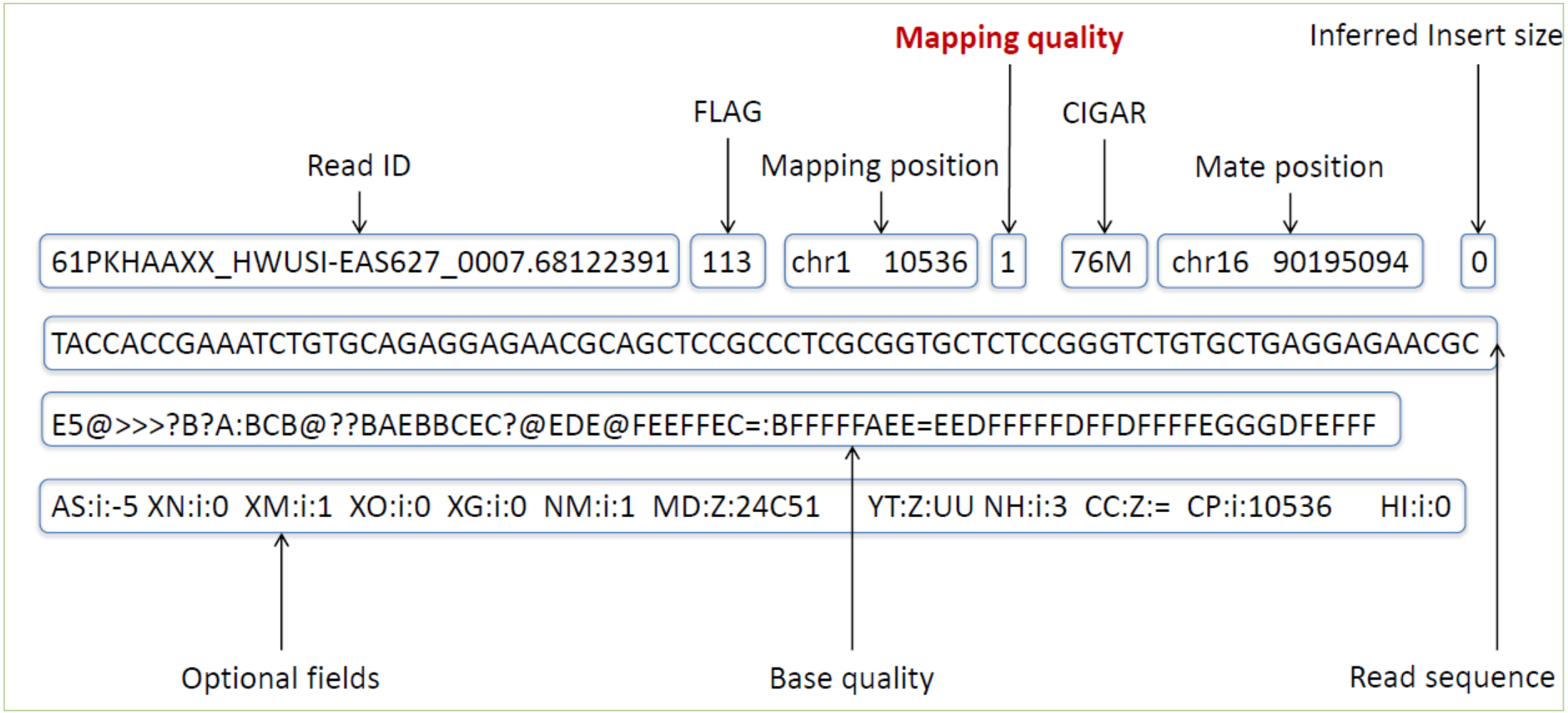
SAM file format

- Mandatory fields:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Optional fields depending on the aligner used

SAM file format



SAM file format

- Mapping quality
  - In the SAM specification
    - 0 → Higher probability of mapping wrong
    - 255 → Lowest probability of mapping wrong

SAM file format

CIGAR

- It contains information about indels, junctions...

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

BAM file format

- SAM file under BGZF compression format.

- Important things:
  - Binary file.
  - Save disk (~80% of compression)
  - Indexing for efficient random access.
  - Easy to convert to one another using SAMtools
  - Accepted by most of the available software

GTF/GFF format

- Tab-delimited text file that defines a feature track
- Annotation files are normally in this format

Chromosome	Source	Feature	Start	End	Score	Strand
chr1	hg19_ensGene	Exon	100	250	0	+
chr1	hg19_ensGene	Start_codon	100	102	0	+
chr1	hg19_ensGene	CDS	100	250	0	+

BED format

- Tab-delimited text file that defines a feature track

Chromosome	Start	End	Feature_ID	Score	Strand
chr1	941	942	Peak_1	12,67	+
chr1	2276	2277	Peak_2	14,55	+
chr1	2718	2719	Peak_3	36,44	+

Mandatory fields ← → Optional fields

Outline

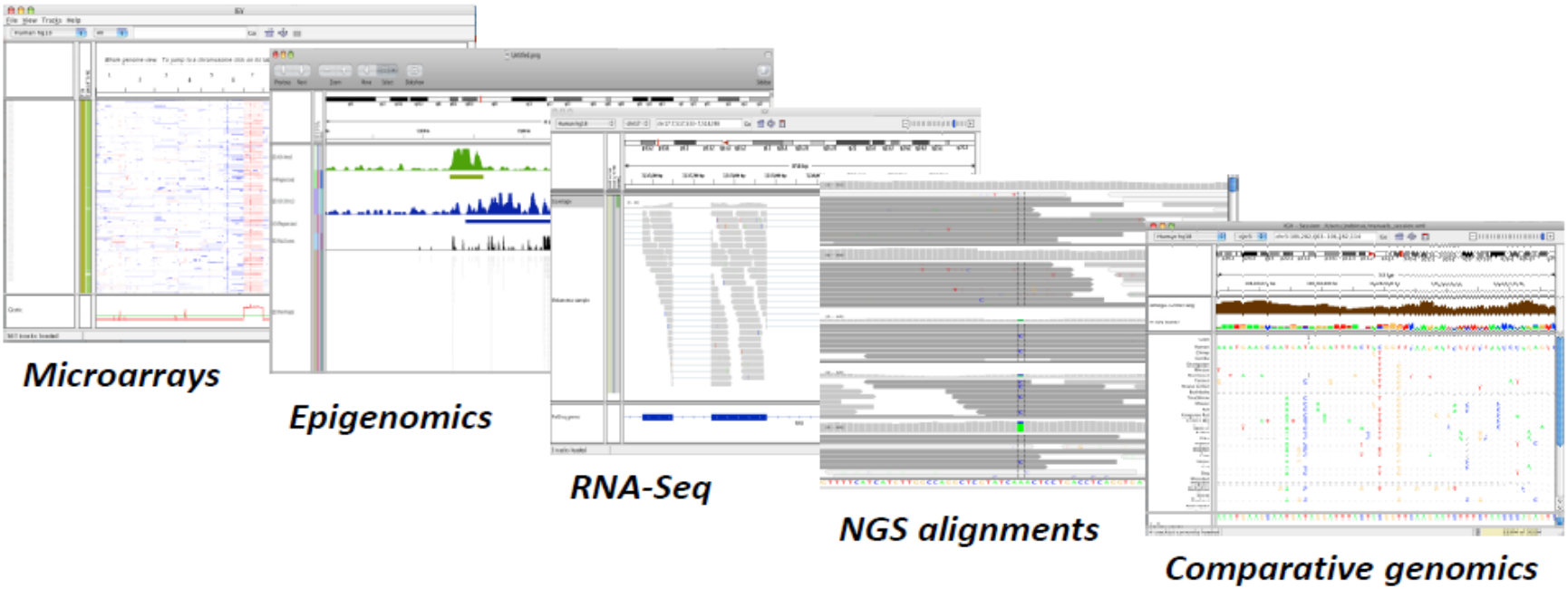
1) Why do we need to visualize results?

2) Different formats: BAM, SAM, BED, GFF

3) Genomics viewers: IGV, Genome Maps

What is IGV?

- High performance visualization tool for **interactive exploration** of large genomic datasets (Microarray, RNA-Seq, epigenomics...)



http://www.broadinstitute.org/igv

Motivations of IGV

- Integrative Genomics Viewer (**IGV**)
  - **Integrate** different data types simultaneously
  - View **large datasets** easily
  - Faster navigation or browsing
  - Runs **locally** on your desktop
  - Used by large-scale projects
  - Open source and **freely available**

THE CANCER GENOME ATLAS

1000 Genomes
A Deep Catalog of Human Genetic Variation

MMRC
Multiple Myeloma Research Consortium
Powerful Collaboration Accelerates Results℠

Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov
Integrative Genomics Viewer (IGV): high-peformance genomics data visualization and exploration
Brief Bioinform (2013) 14 (2): 178-192

IGV Data types

- Any data related to **genome coordinates**
- **Sample annotations or attributes**
- **Genome** annotations

Recommended file formats:

Source data	Recommended File Formats
Sequence alignment data	SAM (must be sorted/indexed) BAM (must be indexed)
Genome annotations	GFF or GFF3 format BED format
Variant data	VCF
Any numeric data	IGV format, TAB format WIG format
Gene expression data	GCT format RES format

IGV Indexing a BAM file

- BAM format: Binary **SAM** file → Reduces disk space and time
- BAM/SAM files need to be **indexed** (using **samtools**) → SAM files will be sorted by start position and indexed
- Index files must reside in the **same directory** as the BAM or SAM file

Index the example BAM file

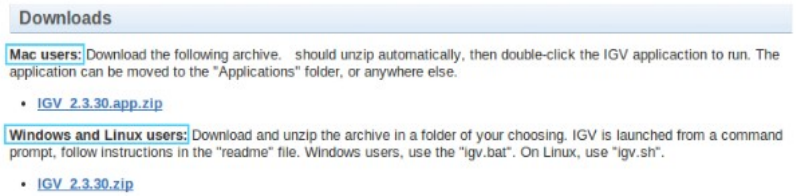
```
samtools index igv1.bam
```

IGV Registration and download

1. Be sure that **Java 6 or later** is installed on your machine
2. Go to the IGV website:

   http://www.broadinstitute.org/igv/home
3. Click **Downloads** at the left panel
4. Click to register and fill the form

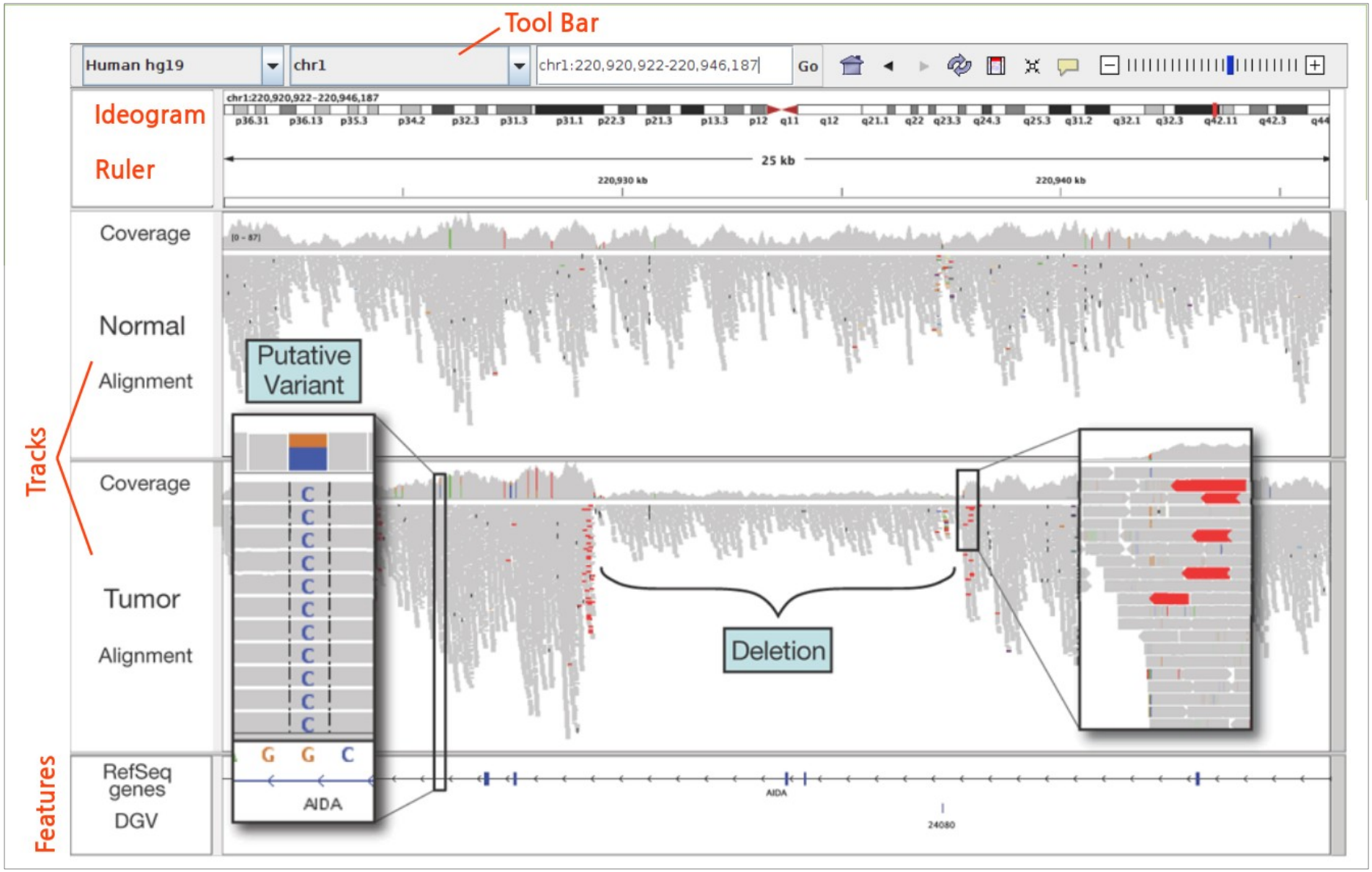
5. Download the most suitable file for your system



6. Run IGV

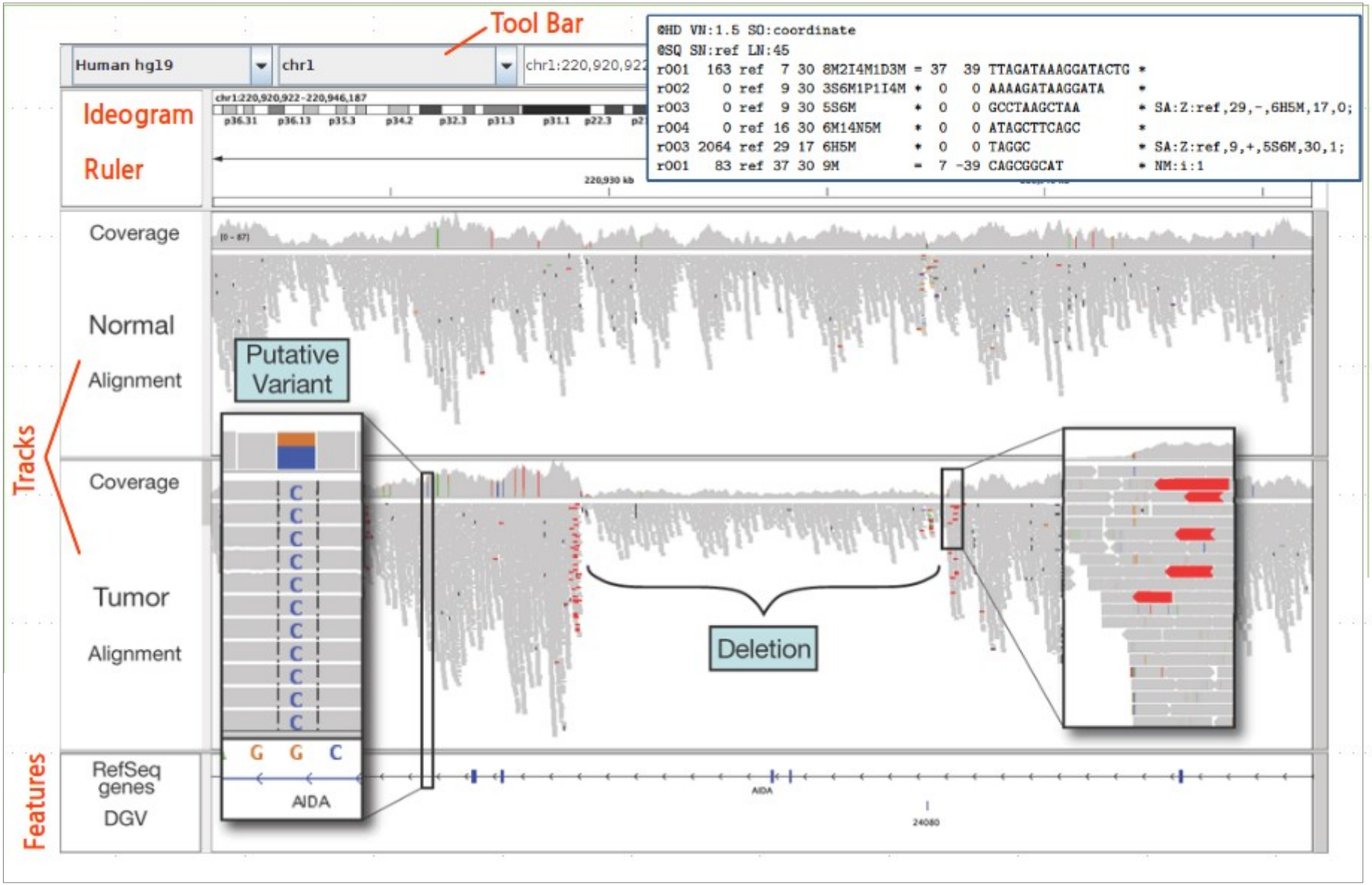
```
./igv.sh
```

IGV Interface



Visualization of mapped reads

IGV interface



Visualization of mapped reads

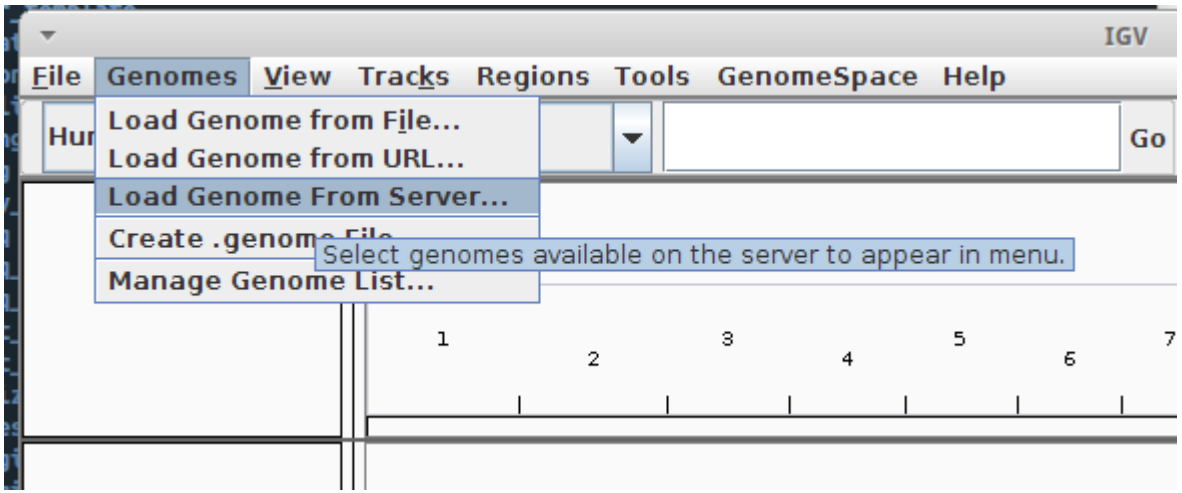
IGV Download genome

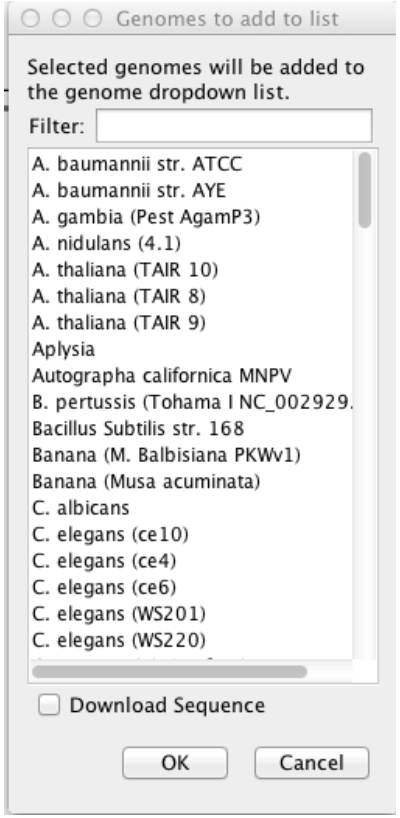
Initially, the genome drop-down list contains a single item, "**Human hg18**"

IGV provides **a number of genomes that are hosted** on a server at the Broad Institute

List of genomes hosted: http://www.broadinstitute.org/software/igv/Genomes

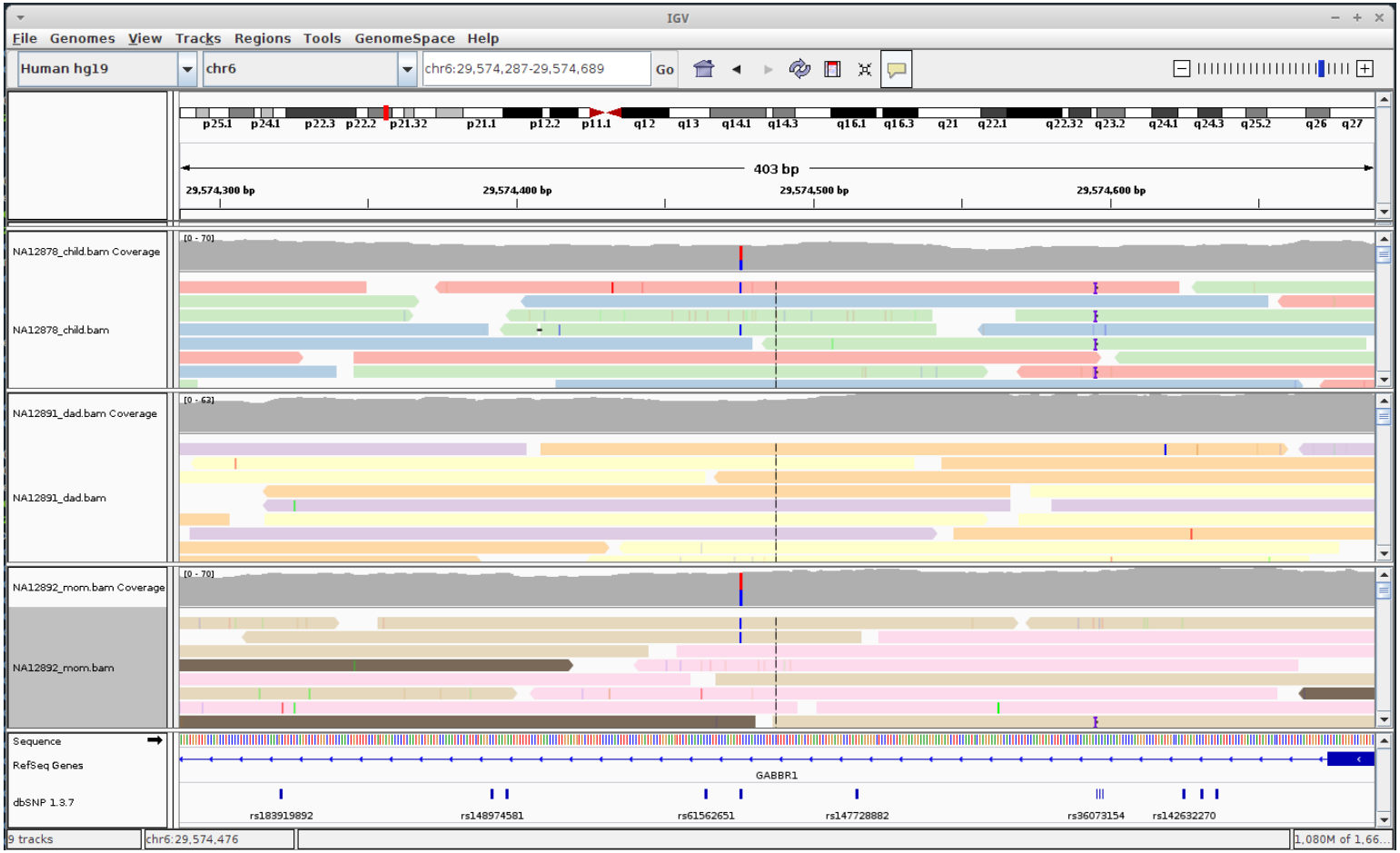
- Genomes → Load genome from server...
  Select Human hg 19





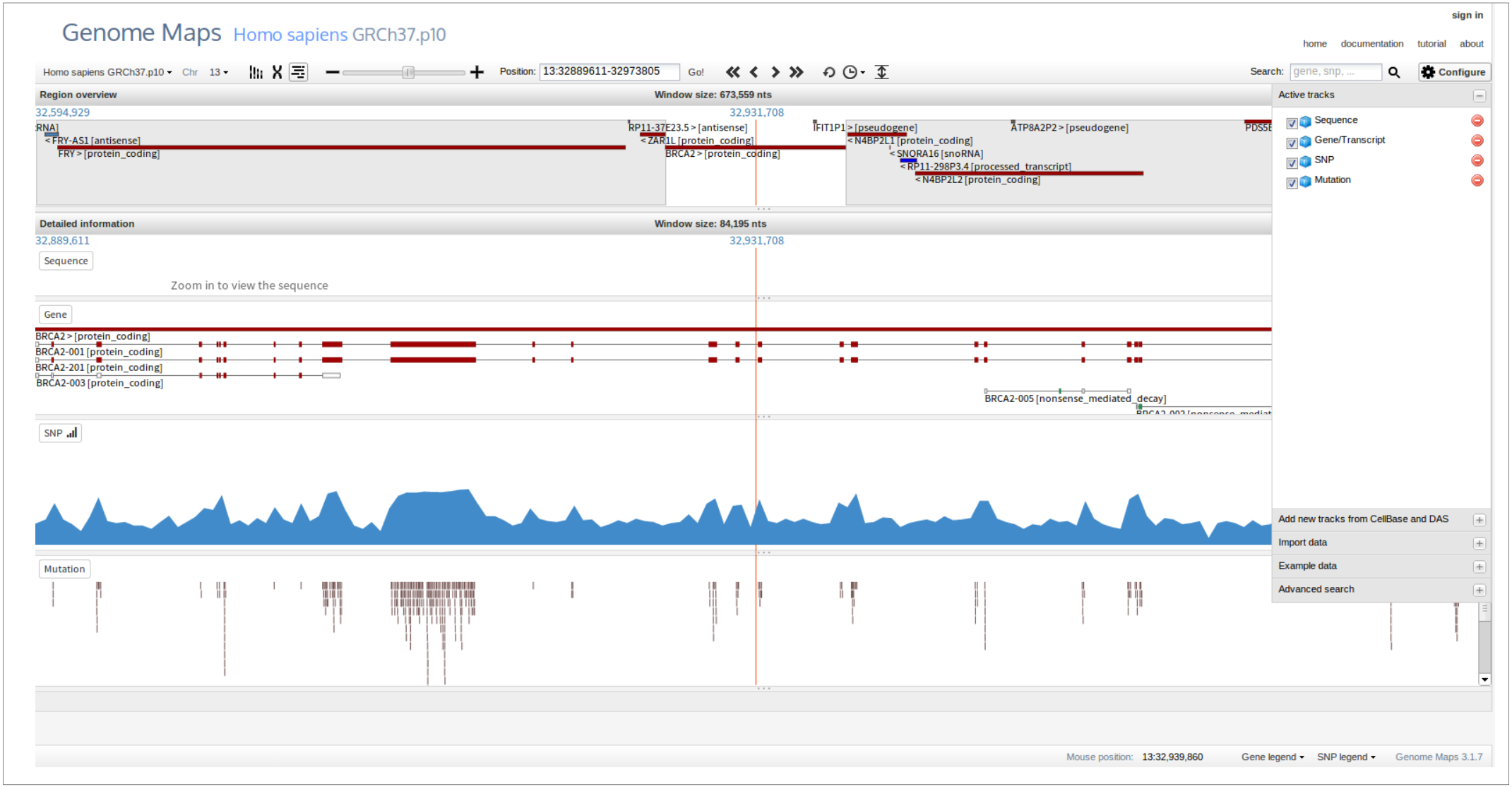
IGV Loading and browsing files

File → Load from file...
Select `NA12878_child.bam`, `NA12891_dad.bam` and `NA12892_mom.bam`



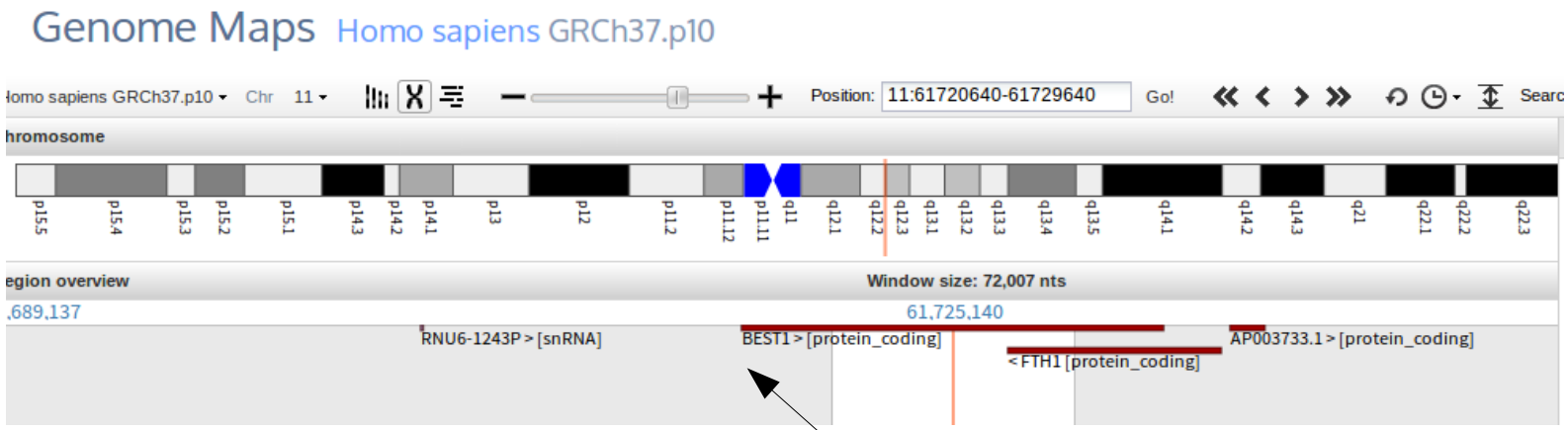
Genome Maps

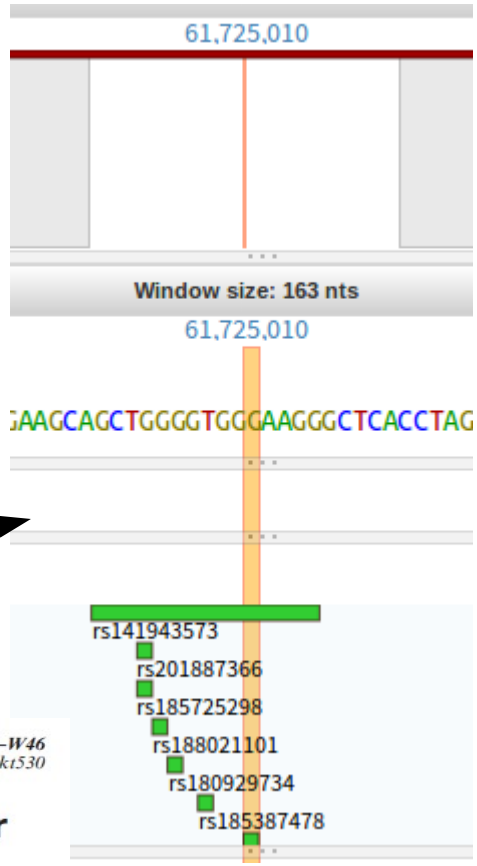
http://www.genomemaps.org/

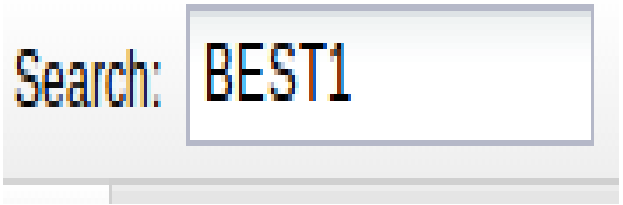


Visualization of mapped reads

Genome Maps








Published online 8 June 2013 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* ***W41–W46***
doi:10.1093/nar/gkt530

Genome Maps, a new generation genome browser

Ignacio Medina[1,*], Francisco Salavert[1,2], Rubén Sanchez[3], Alejandro de Maria[1], Roberto Alonso[1], Pablo Escobar[1], Marta Bleda[1,2] and Joaquín Dopazo[1,2,4,*]

[1]Department of Computational Genomics, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012, Spain, [2]CIBER de Enfermedades Raras (CIBERER), Valencia 46012, Spain, [3]Genometra S.L., Valencia, Spain and [4]Functional Genomics Node (INB) at CIPF, Valencia 46012, Spain

Any questions?