IX International Course of Massive Data Analysis FOR GENOMICS

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

INB

GBPA Genomics & Bioinformatics Platform of Andalusia

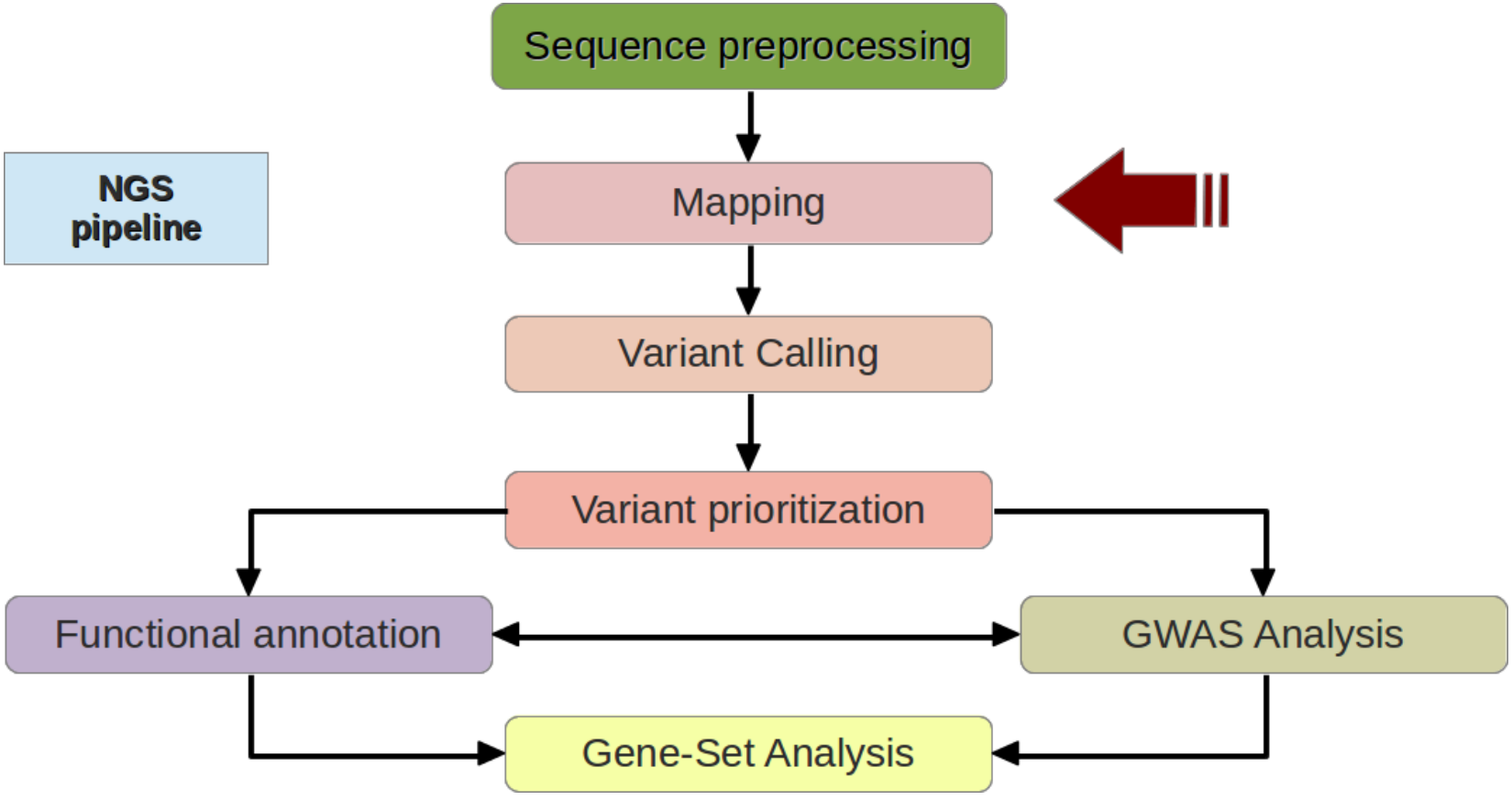
Fco. Javier López

Visualization of Mapped Reads: Integrative Genomics Viewer (IGV)

GBPA Genomics & Bioinformatics Platform of Andalusia

GBPA Genomics & Bioinformatics Platform of Andalusia

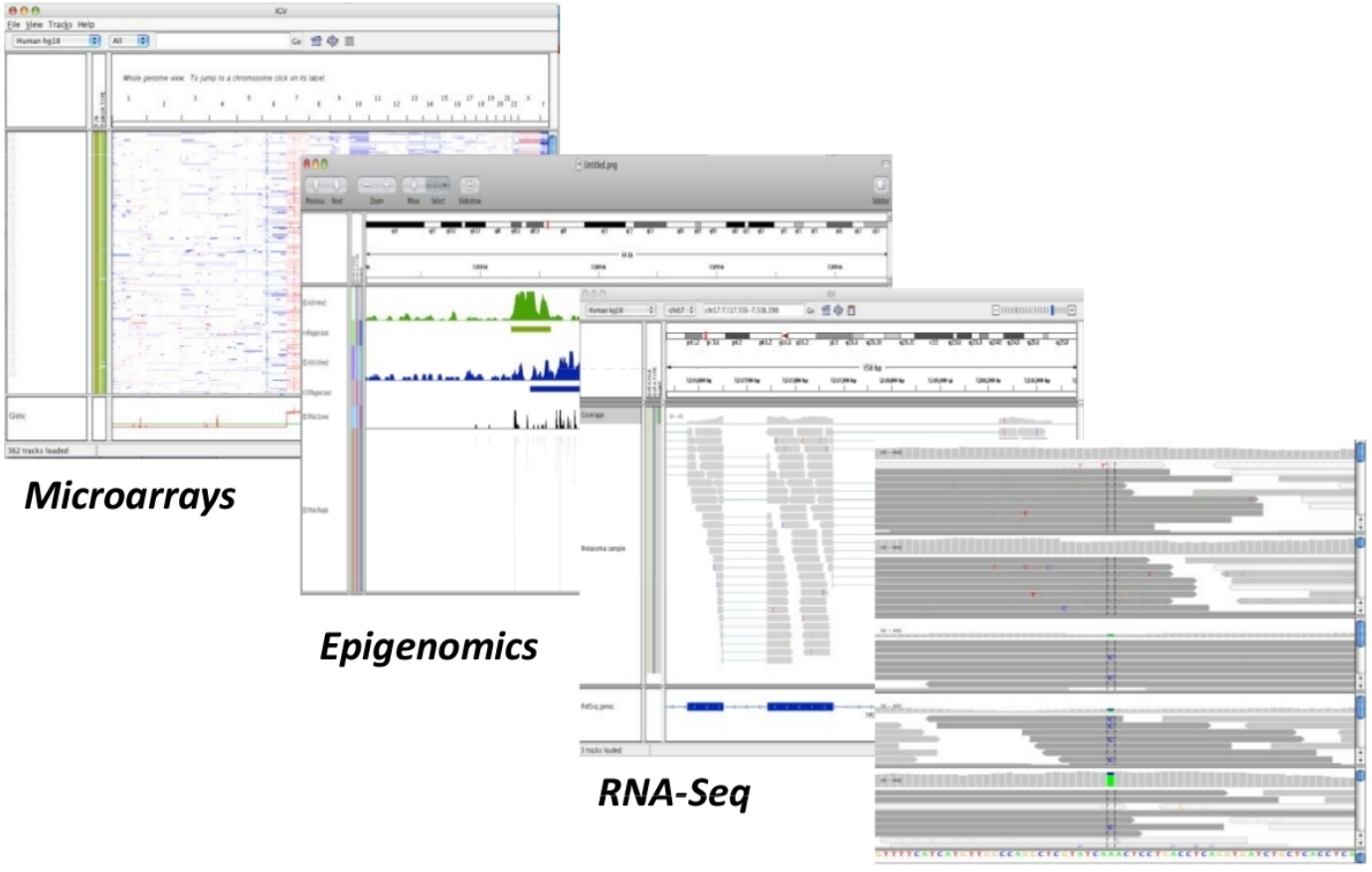
Where are we?



GBPA Genomics & Bioinformatics Platform of Andalusia

Why IGV?

- IGV is an integrated visualization tool of large data types



Why IGV?

- Integrate different data types simultaneously
- View large datasets easily
- Fast navigation
- Run it locally on desktop
- Easy to use interface

GBPA Genomics & Bioinformatics Platform of Andalusia

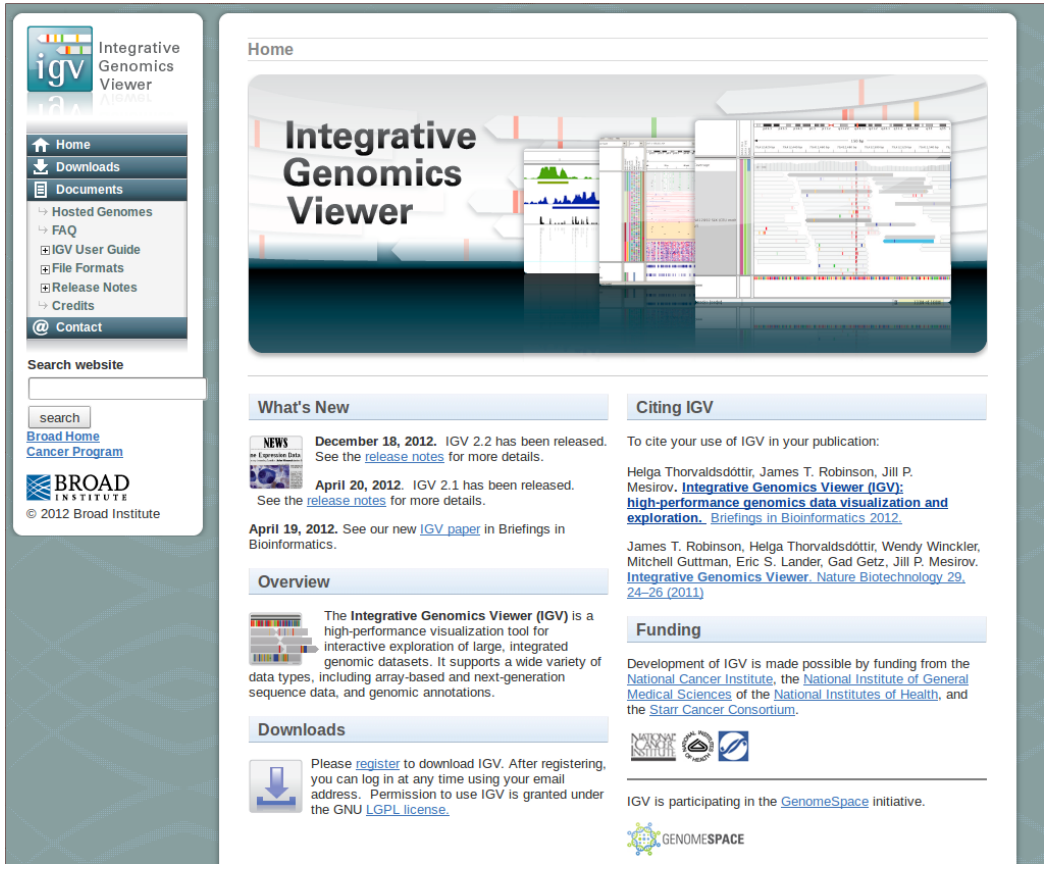
Why IGV?

- Large-Scale projects using IGV
  - The Cancer Genome Atlas
    http://cancergenome.nih.gov/
  - Multiple Myeloma Research Consortium
    http://themmrc.org/
  - 1000 Genomes Project
    http://www.1000genomes.org/

GBPA Genomics & Bioinformatics Platform of Andalusia

Getting IGV

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



GBPA Genomics & Bioinformatics Platform of Andalusia

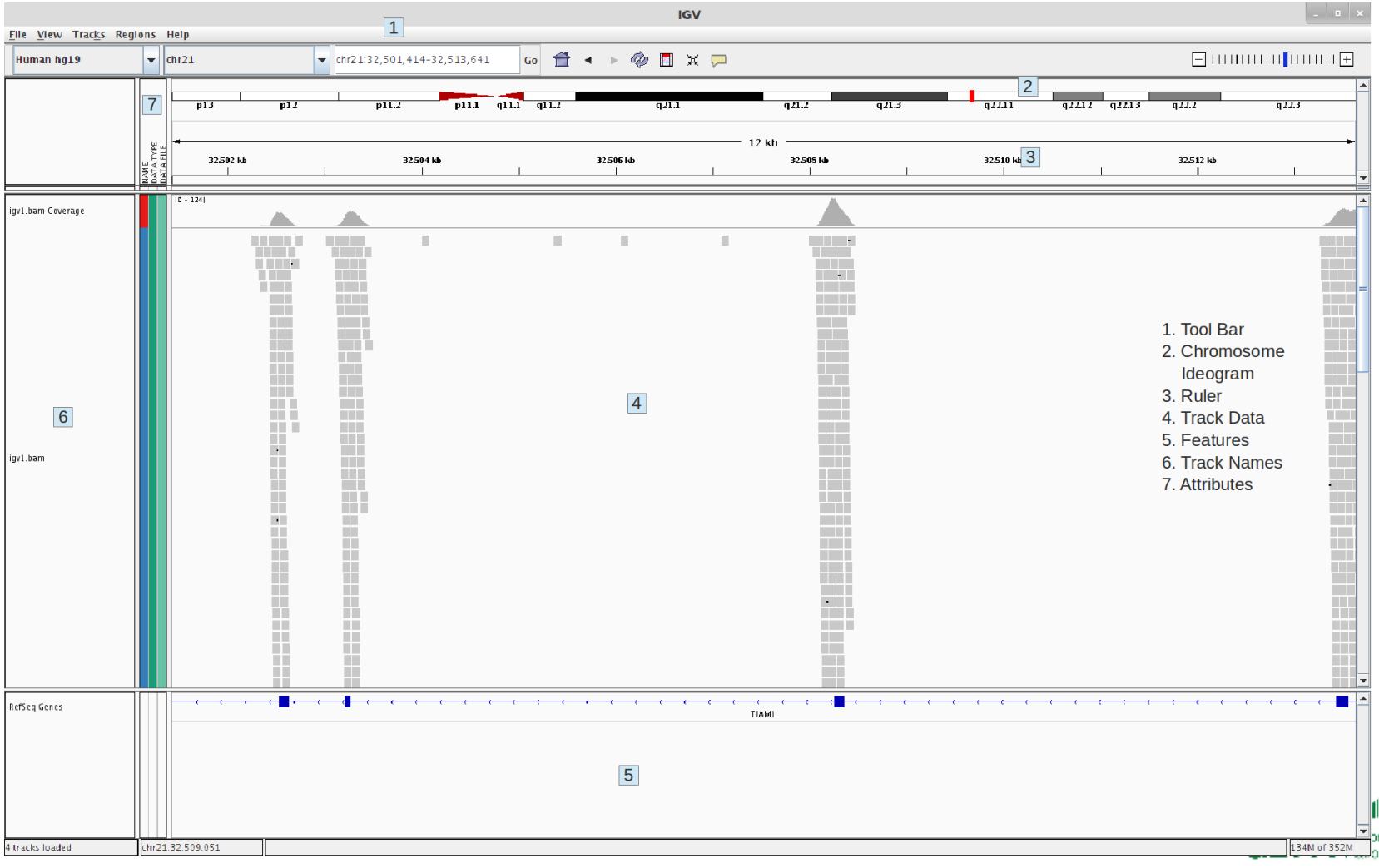
History and Usage

- First release was in August 2008
- Current version: 2.2
- Open source and freely available

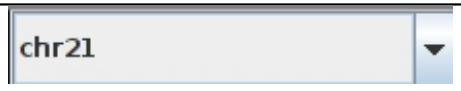
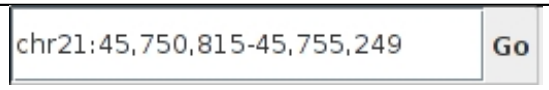
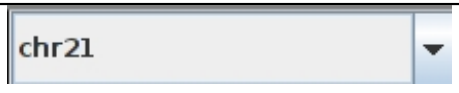
GBPA Genomics & Bioinformatics Platform of Andalusia

GBPA Genomics & Bioinformatics Platform of Andalusia

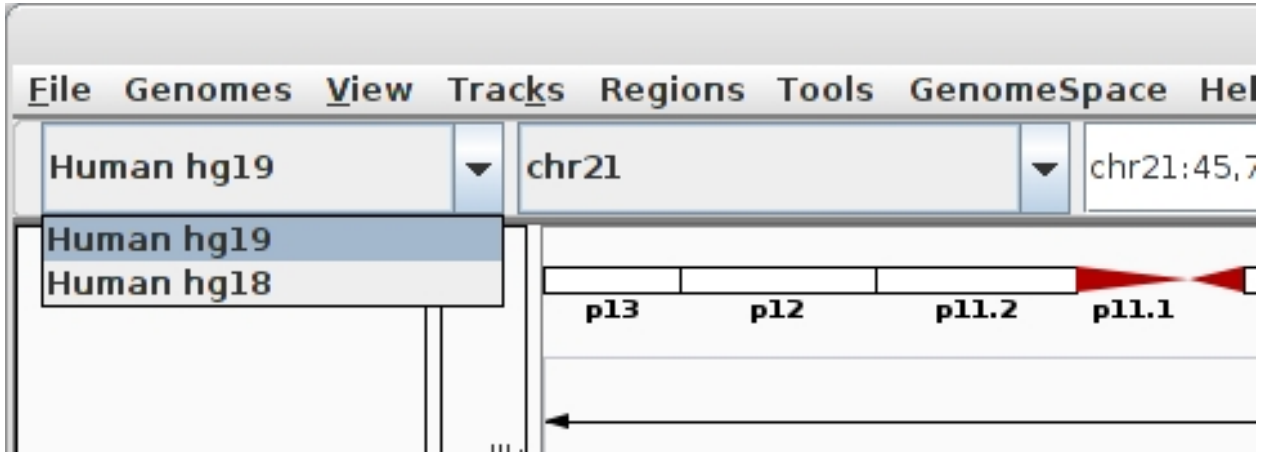
Main Window


IGV
File View Tracks Regions Help
Human hg19
chr21
chr21:32,501,414-32,513,641
Go
1
2
3
4
5
6
7
p13 p12 p11.2 p11.1 q11.1 q11.2 q21.1 q21.2 q21.3 q22.11 q22.12 q22.13 q22.2 q22.3
12 kb
32.502 kb 32.504 kb 32.506 kb 32.508 kb 32.510 kb 32.512 kb
igv1.bam Coverage
igv1.bam
RefSeq Genes
TIAM1
4 tracks loaded
Chr21:32.509.051
134M of 352M
1. Tool Bar
2. Chromosome Ideogram
3. Ruler
4. Track Data
5. Features
6. Track Names
7. Attributes

Tool Bar

Genome drop-down box 	Loads a genome
Chromosome drowp-down box 	Zooms to a chromosome
Search box 	Displays a genomic location
Whole genome view 	Zooms to whole genome view
	Moves backward and forward
Refresh 	Refreshes the display
Define a region 	Defines a region of interest
	Fits all tracks to the window
	Toggles on/off pop-up information
Zoom slider 	Moves to a chromosome

Available Genomes



- Human, Mouse, S. cerevisiae, C. Elegans, D. melanogaster, and others

http://www.broadinstitute.org/software/igv/Genomes

GBPA Genomics & Bioinformatics Platform of Andalusia

GBPA Genomics & Bioinformatics Platform of Andalusia

Data types

- General characteristics
  - Any data related to genome coordinates
  - Sample annotation/attributes
  - Genome annotations

GBPA Genomics & Bioinformatics Platform of Andalusia

Data types

- IGV supports multiple file formats

Source Data	Recommended File Formats
Sequence alignment data	SAM (must be sorted/indexed) BAM format (must be indexed)
Genome annotations	GFF or GFF3 format, BED format
ChIP-Seq, RNA-Seq	TDF format. Use igvtools to generate a binary read count. Load the resulting TDF file into IGV.
Any numeric data	IGV format, TAB format WIG format
Gene expression data	GCT format RES Format

GBPA Genomics & Bioinformatics Platform of Andalusia

Loading a BAM file

- BAM format: SAM binary. Reduces disk space and access time.
  - For each read, provides the position(s) where it maps and information about the alignment.
  - BAM files need to be indexed (samtools). SAM files need to be sorted by start position and indexed.

Index an example bam file

```
samtools index /home/biouser/mda13/mqc-igv/igv1.bam
ls -la /home/biouser/mda13/mqc-igv/igv1.*
```

GBPA Genomics & Bioinformatics Platform of Andalusia

Loading a BAM file

Open IGV

```
igv
```

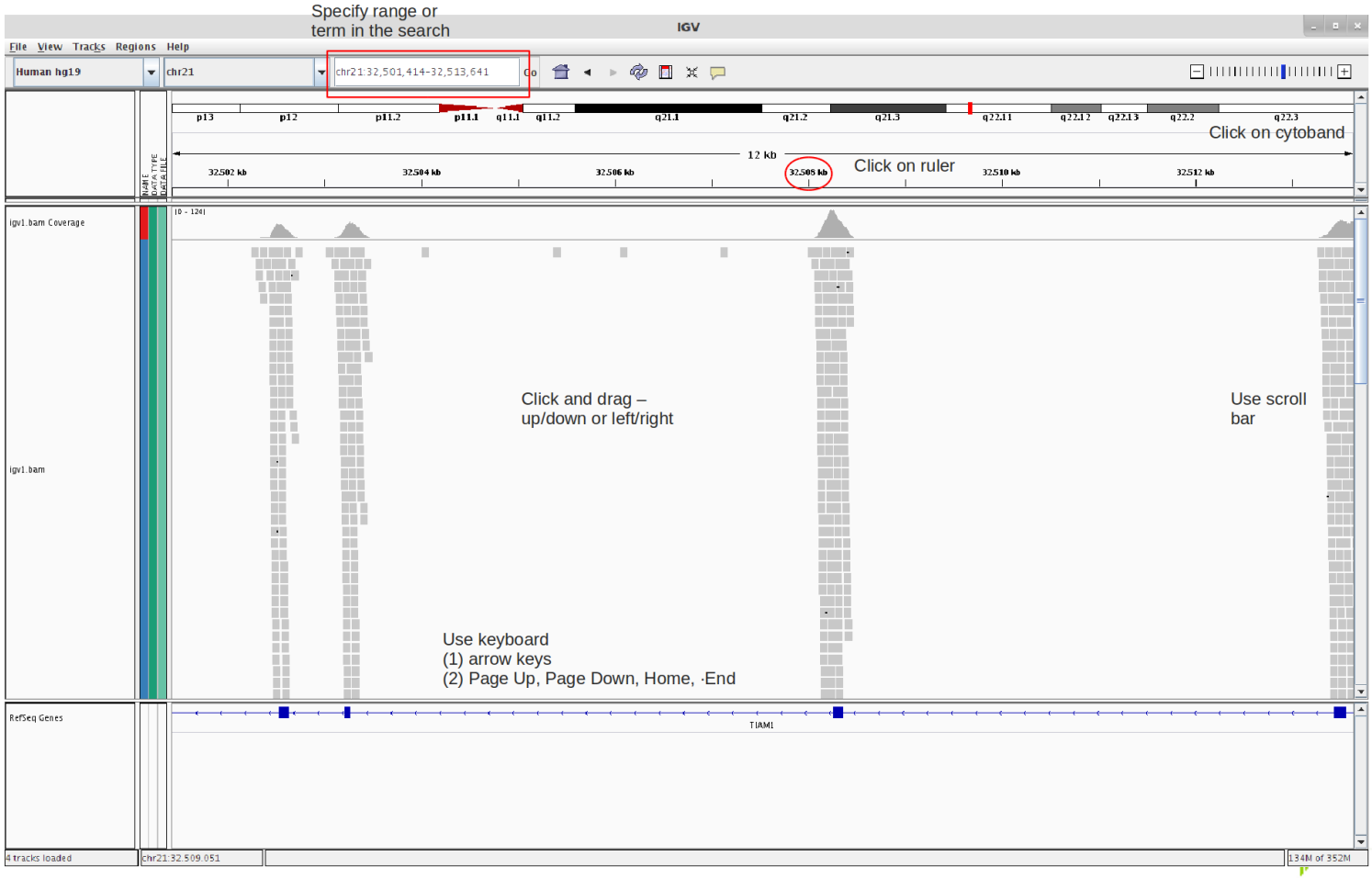
Load an example bam file

- File → Load from file → /home/biouser/mda13/mqc-igv/igv1.bam

GBPA Genomics & Bioinformatics Platform of Andalusia

GBPA Genomics & Bioinformatics Platform of Andalusia

Moving around



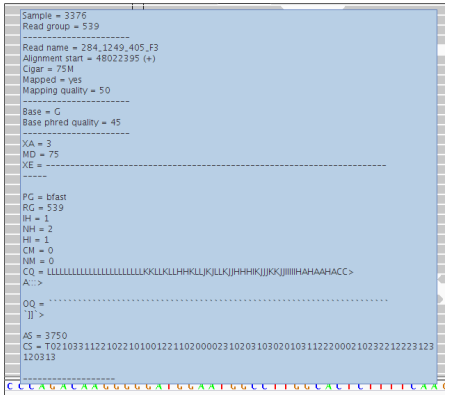
GBPA Genomics & Bioinformatics Platform of Andalusia

Moving around

Type genomic coordinates 21:32508271 in the search box

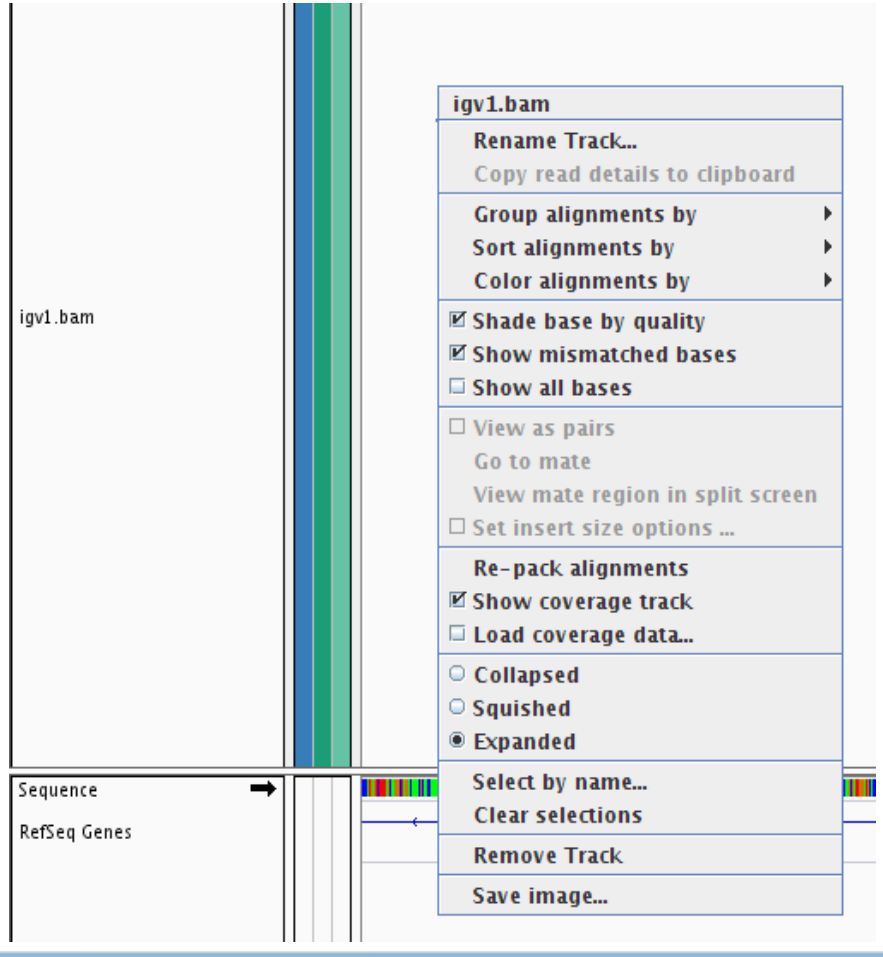
- At low resolution only coverage is shown
- At higher resolution, reads are shown including where bases differ

Leave the cursor on a read



Track options

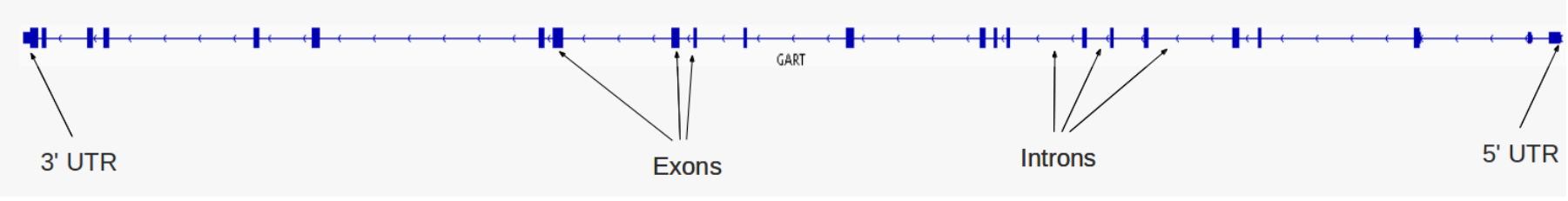
Right click on track



GBPA Genomics & Bioinformatics Platform of Andalusia

RefSeq track

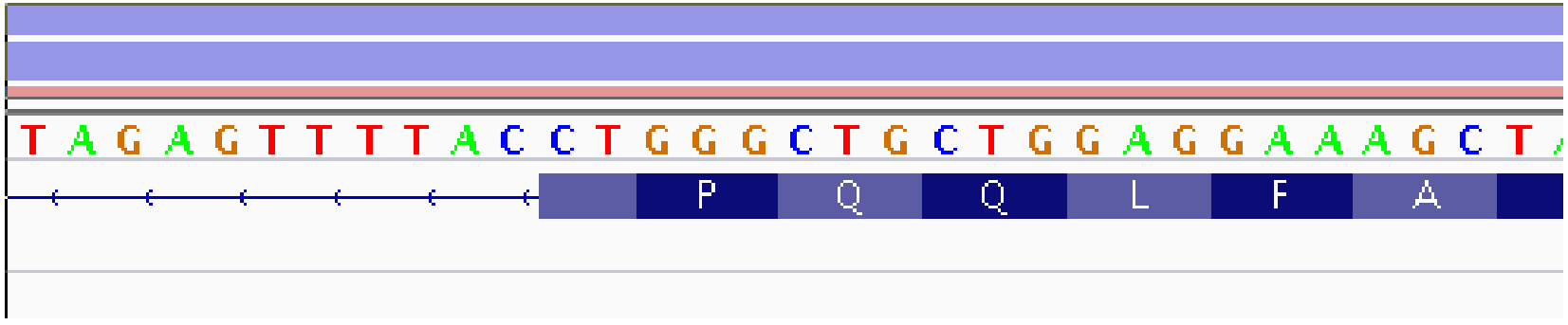
Type GART in the search box



GBPA Genomics & Bioinformatics Platform of Andalusia

RefSeq track

Zoom in to focus on an exon

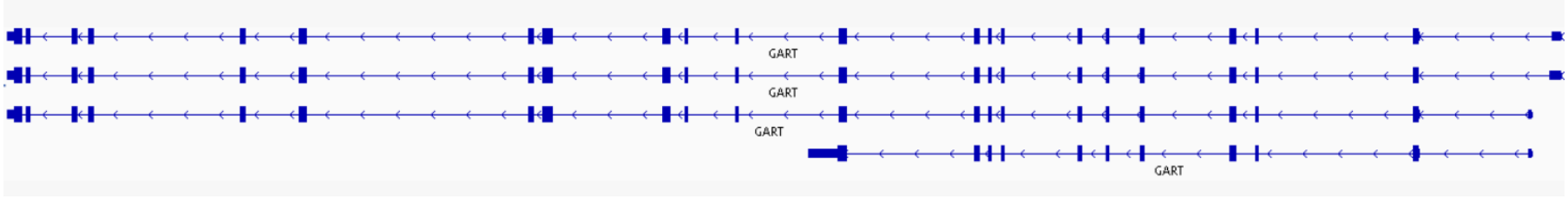


- Reference nucleotide sequence
- Predicted aminoacid sequence

GBPA Genomics & Bioinformatics Platform of Andalusia

RefSeq track

Right click on the RefSeq track → Expanded



Zoom out to get a general view of the mapping profile

- Reads are grouped around exons. Is that a coincidence?
  No. This bam is part of an exome sequencing experiment.

GBPA Genomics & Bioinformatics Platform of Andalusia

Loading a BED file

- Let's load capture regions
- BED format: to store a list of genomic regions. Text file with the list of regions. Each line contains one region with three required fields separated by tabs: chromosome, start coordinate, end coordinate

  http://genome.ucsc.edu/FAQ/FAQformat.html

```
less /home/biouser/mda13/mqc-igv/igv.bed
```

Load a bed file

- File → Load from file → igv.bed
  Another track appears: blue boxes indicate target regions

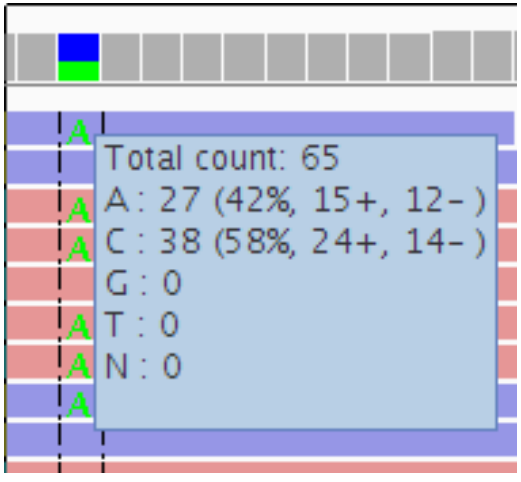
GBPA Genomics & Bioinformatics Platform of Andalusia

Visualizing variants

Move to 21:48022375

- Variant: ref:C, alt:A, heterozygosis

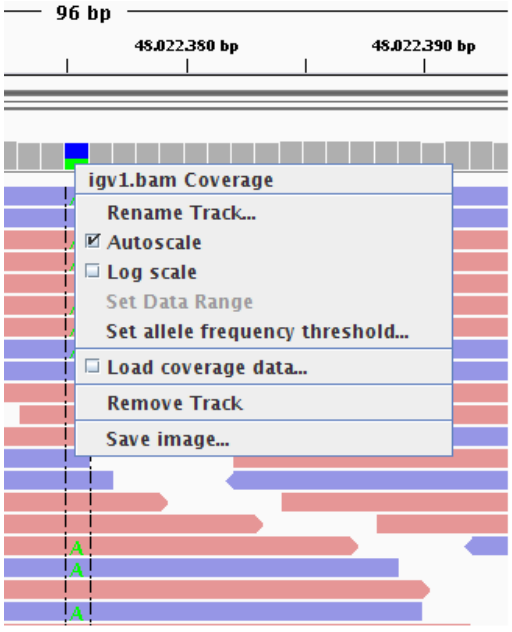
Leave the cursor over that position on the histogram (top of the track)

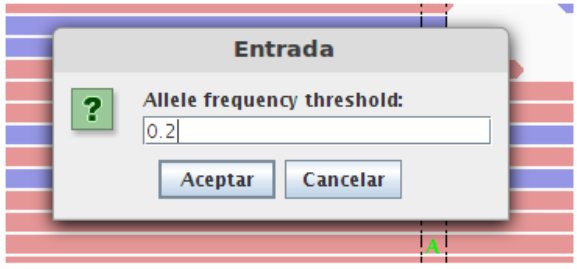


GBPA Genomics & Bioinformatics Platform of Andalusia

Visualizing variants

Right click over that position on the histogram





Move to 21:47786817

- Variant: ref:C, alt:G, homozygosis

GBPA Genomics & Bioinformatics Platform of Andalusia

Visualizing variants

- Files containing a variant list can also be loaded: VCF files (Variant Calling Format).

```
less /home/biouser/mda13/mqc-igv/igv.vcf
```

- VCF format:

  http://genome.ucsc.edu/FAQ/FAQformat.html

  - Text file
  - Header including information about the mapping and variant calling processes: set of lines begining with ##
  - An additional header line begining with #: contains a table header with column identifiers
  - One line for each variant: chromosome, genomic position, reference and alternative bases

GBPA Genomics & Bioinformatics Platform of Andalusia

Visualizing variants

- VCF files need to be indexed before being loaded:

```
igvtools index /home/biouser/mda13/mqc-igv/igv.vcf
ls -la /home/biouser/mda13/mqc-igv/igv.vcf*
```

Load a VCF file

- File → Load from file → igv.vcf A new track is added

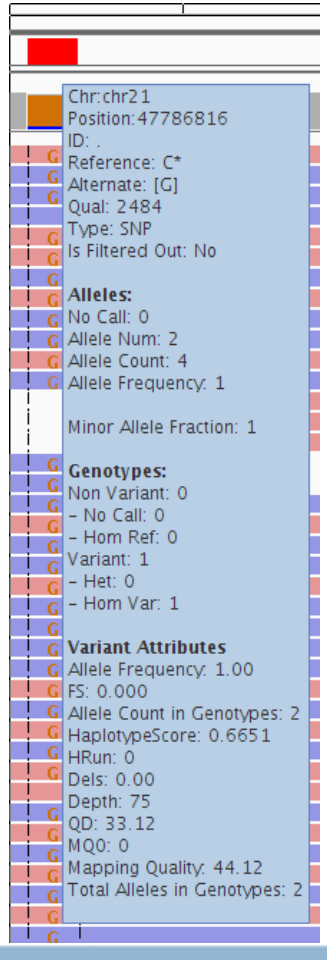
Move to 21:47786817

- A peak appears indicating the variant position.

GBPA Genomics & Bioinformatics Platform of Andalusia

Visualizing variants

- Leave the cursor over that position on the vcf track:

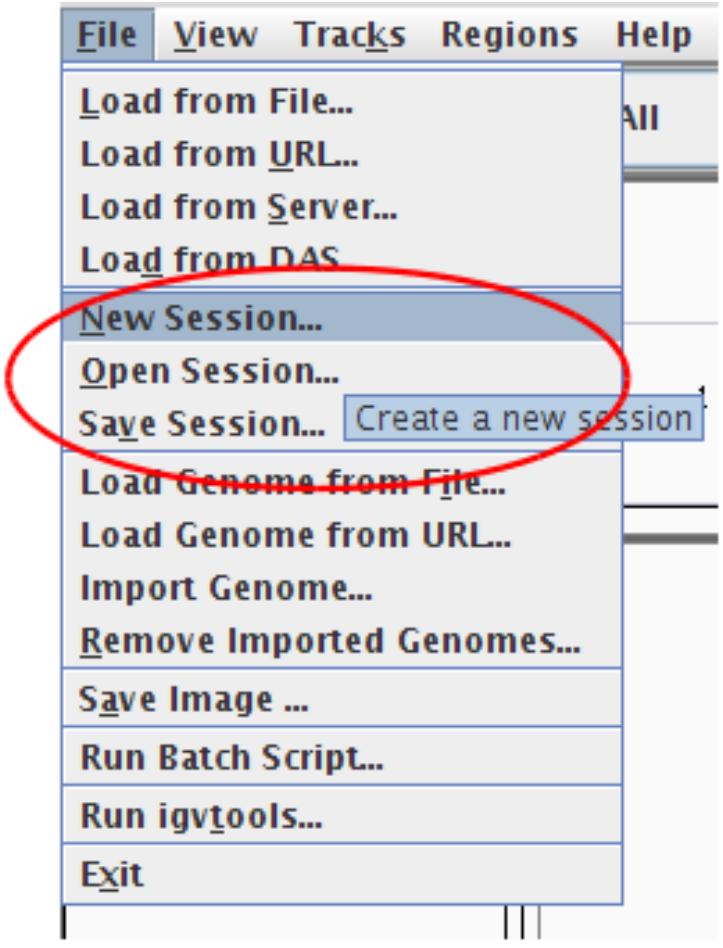


GBPA Genomics & Bioinformatics Platform of Andalusia

GBPA Genomics & Bioinformatics Platform of Andalusia

Create/Open/Save Sessions

- Your current session can be saved



GBPA Genomics & Bioinformatics Platform of Andalusia

Exercise 1

- Load /home/biouser/mda13/mqc-igv/igv2.bam
- Load /home/biouser/mda13/mqc-igv/igv1.bam
- Move to chr11:1,016,174-1,018,316
- Can you see any difference between both tracks?
- Is the igv2.bam mapping a good or a bad result?
- Can you think of the reasons that lead to such result?
- Can you think of any way of improving it?

GBPA Genomics & Bioinformatics Platform of Andalusia

Exercise 2

- Load /home/biouser/mda13/mqc-igv/igv1.bam
- Move to 21:47917047
- What is happenning in our sample in that position?

GBPA Genomics & Bioinformatics Platform of Andalusia

Exercise 3

- Load /home/biouser/mda13/mqc-igv/igv1.bam
- Move to 21:26973663
- What is happenning in our sample in that position?

GBPA Genomics & Bioinformatics Platform of Andalusia

Exercise 4

- Load /home/biouser/mda13/mqc-igv/igv1.bam
- Which of these variants would you trust? Why? Which is the sequence change in each case?
  1. 21:47821726
  2. 21:46596230
  3. 21:42848560
  4. 21:47917170

GBPA Genomics & Bioinformatics Platform of Andalusia

Exercise 5

- Considering that /home/biouser/mda13/mqc-igv/igv.bed contains the target regions of a given sequencing experiment and that /home/biouser/mda13/mqc-igv/igv1.bam is the mapping result, how well are the target regions of the gene MX2 covered?
- Focus on a particular target region and have a look at the histogram on the top of the igv1.bam track.
  - Which is the shape of the histogram? Can you explain this?
  - Which would be the desired shape?

GBPA Genomics & Bioinformatics Platform of Andalusia

Exercise 6

- Considering that /home/biouser/mda13/mqc-igv/igv.bed contains the target regions of a given sequencing experiment and that /home/biouser/mda13/mqc-igv/igv1.bam is the mapping result, look for unsequenced target regions of the PRMT2 gene.

GBPA Genomics & Bioinformatics Platform of Andalusia