Visualization of Data

Many slides courtesy Broad/MIT.

What is IGV

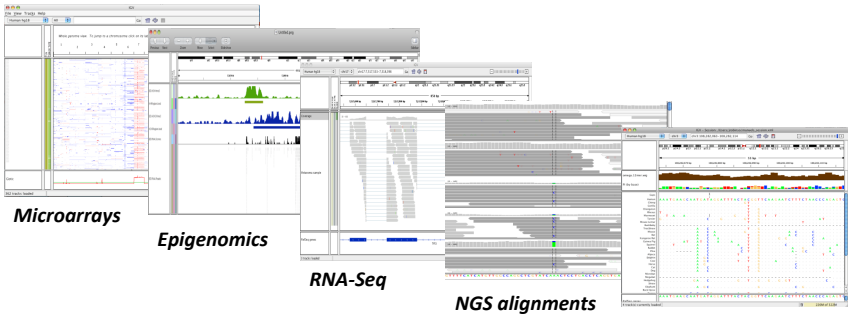
Integrative Genomics Viewer

A desktop/server application

for integrated visualization

of multiple data types and annotations

in the context of the genome



BROAD INSTITUTE

Motivation

Integrative Genomics Viewer

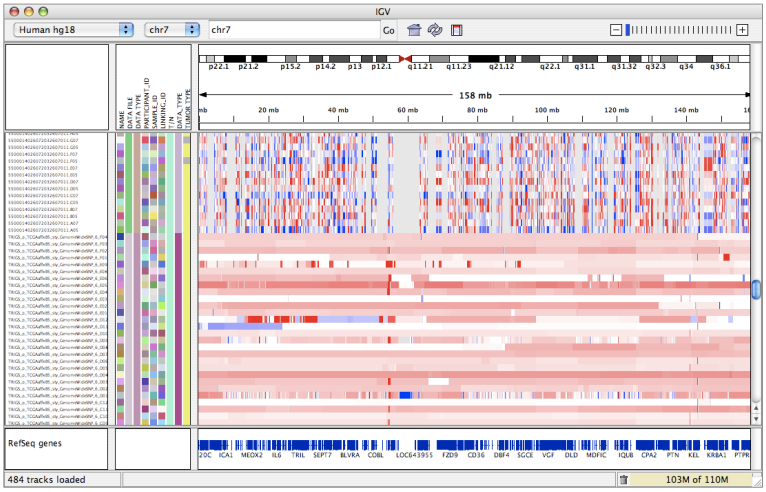
- Easily view investigator-generated datasets alongside publically available data
- Support integration of diverse data types and sample attribute information
- Handle large datasets

BROAD INSTITUTE

IGV layout

Integrative Genomics Viewer

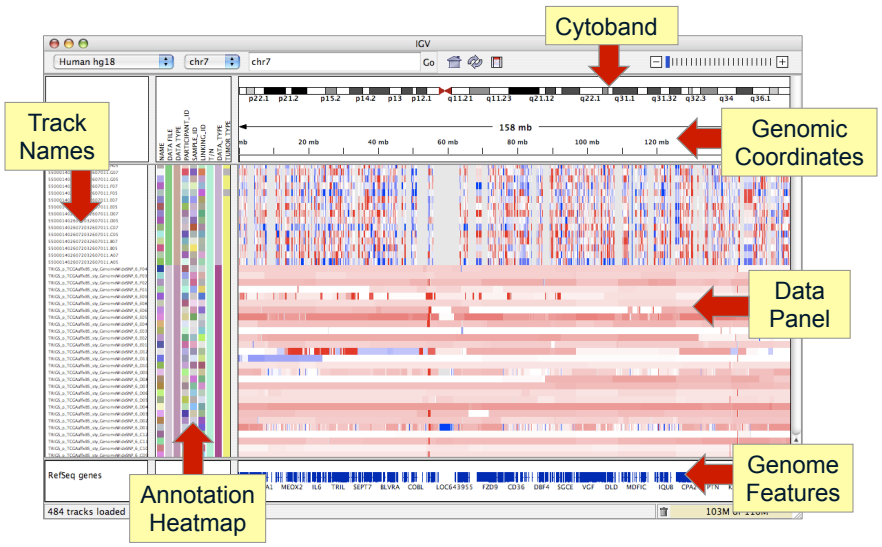
Expression and copy number data



BROAD INSTITUTE

IGV layout

Integrative Genomics Viewer

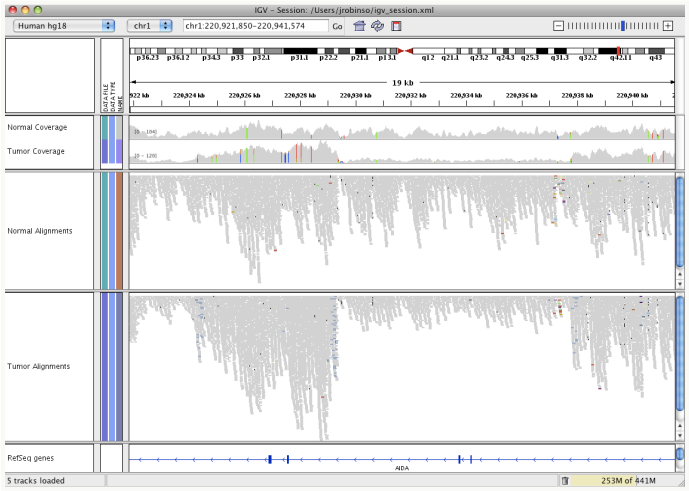


BROAD INSTITUTE

IGV layout

Integrative Genomics Viewer

NGS data

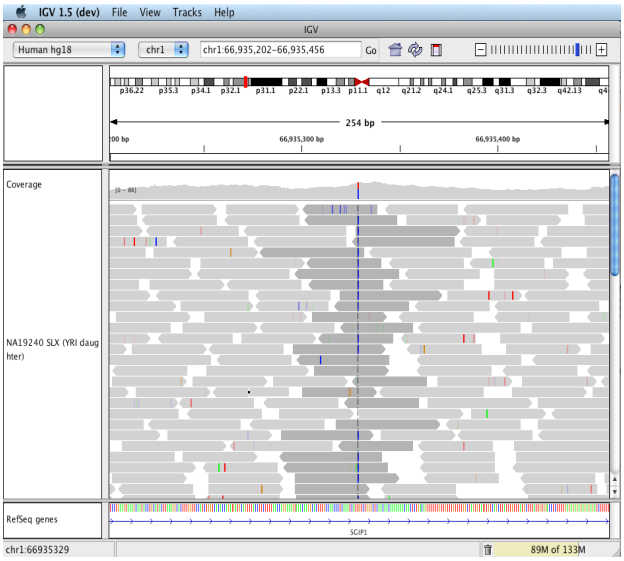


BROAD INSTITUTE

IGV layout

Integrative Genomics Viewer

NGS data



BROAD INSTITUTE

UI basics

Integrative Genomics Viewer

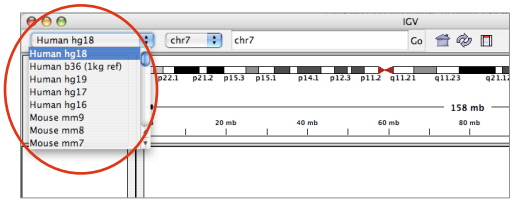
- Selecting a reference genome
- Loading data
- Navigating through the data
- Setting track attributes

BROAD INSTITUTE

Selecting a reference genome

Integrative Genomics Viewer

- Select one of the hosted genomes from the pull-down menu



- For more information see www.broadinstitute.org/igv/Genomes
- You can import other genomes if you have the sequence data

BROAD INSTITUTE

Loading data

Integrative Genomics Viewer

Types of data
- Any data tied to genomic coordinates
- Genome annotations
- Sample attributes/annotations

File formats
- Many different file formats supported
- See www.broadinstitute.org/igv/FileFormats

BROAD INSTITUTE

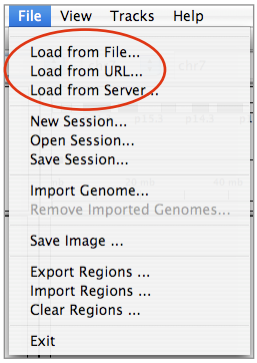
Tracks

- Two generic types:
  - data (continuous valued data)
  - annotation (features)
- Specialized types include
  - alignments
  - mutations
  - multiple alignments

- Type is defined by file format, and can be overridden by the user

- IGV uses type to determine
  - initial placement in a panel
  - display options and options for other track attributes

Loading data



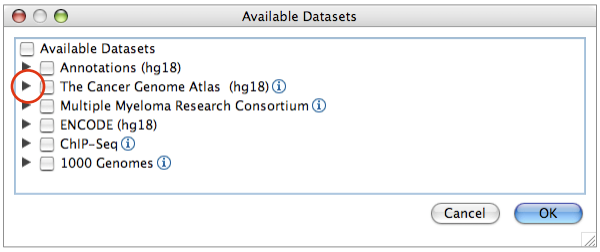
#1 : Load local file

#2 : Load from URL

#3 : Load from server

(Broad IGV data server, other data server)

"Load from server" menu



What you see depends on :

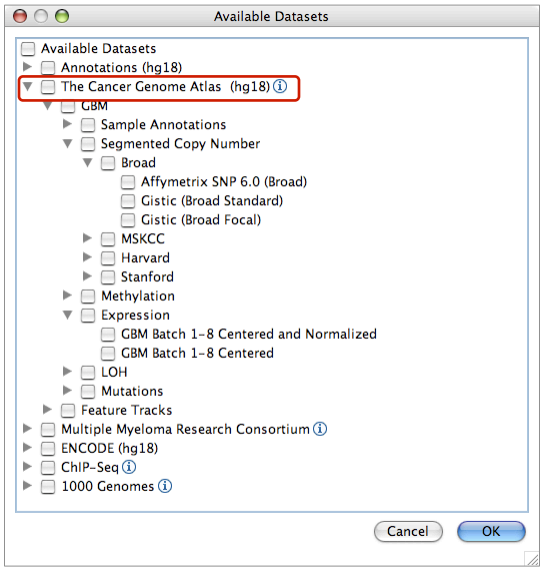
(1) which server you selected – default is Broad server

(2) which reference genome you've selected

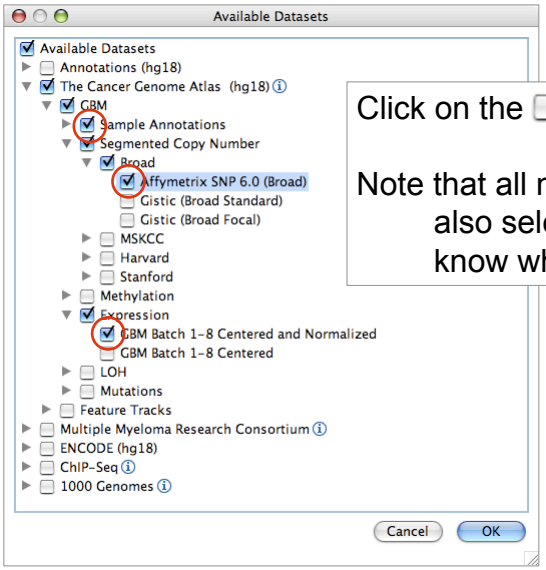
Click on the ⓘ for more information about the data source

Click on the ▶ ɔ expand the sub-menus

"Load from server" menu



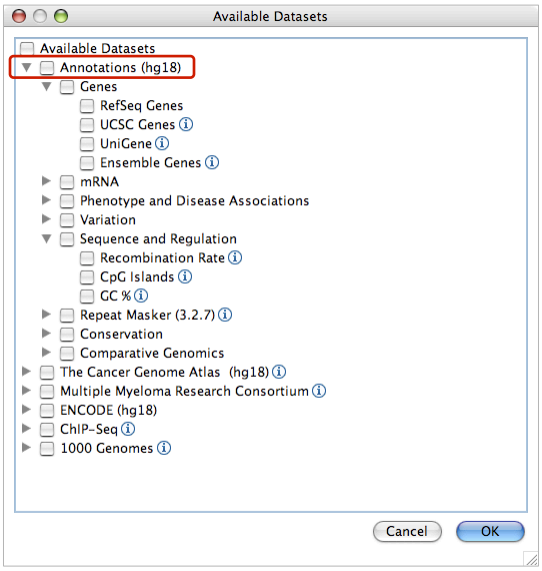
"Load from server" menu



Click on the ☐ to select datasets

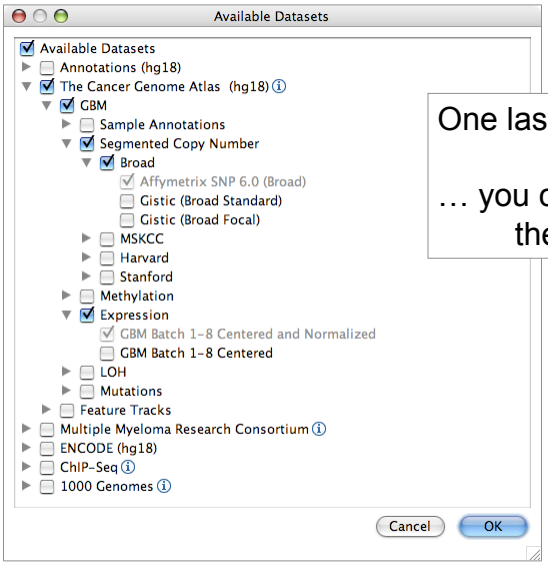
Note that all nested datasets are also selected – make sure you know what you've selected

"Load from server" menu



"Load from server" menu

Integrative Genomics Viewer



One last thing …

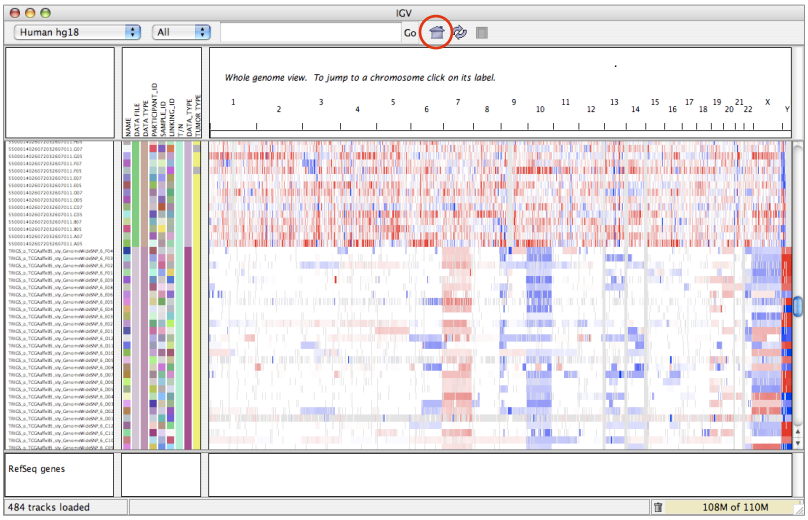
… you cannot **unload** using the checkboxes

BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Whole genome view

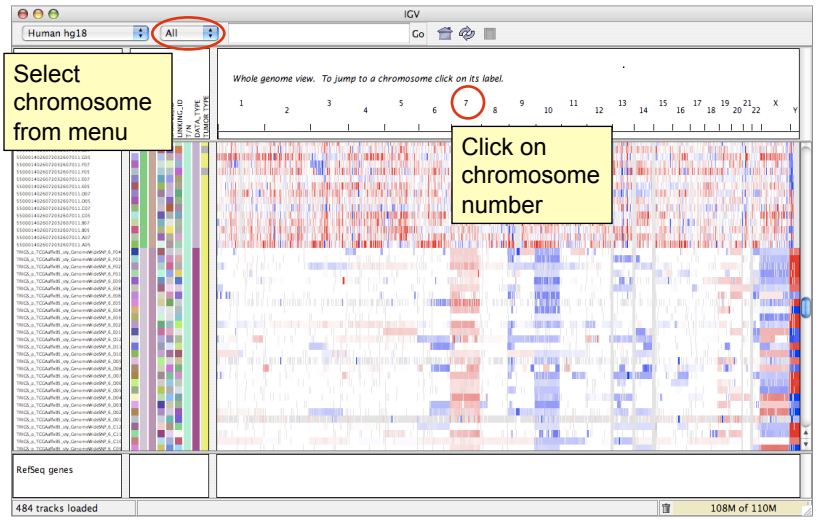


BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Zooming in to the chromosome level

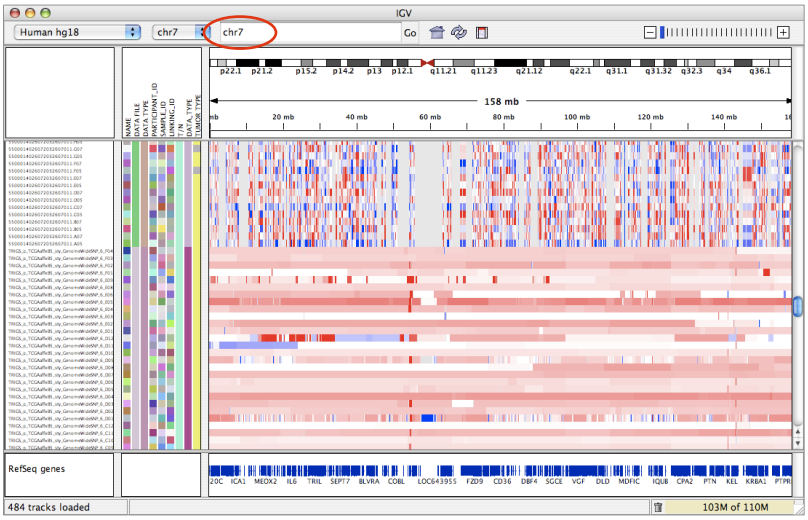


BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Chromosome view

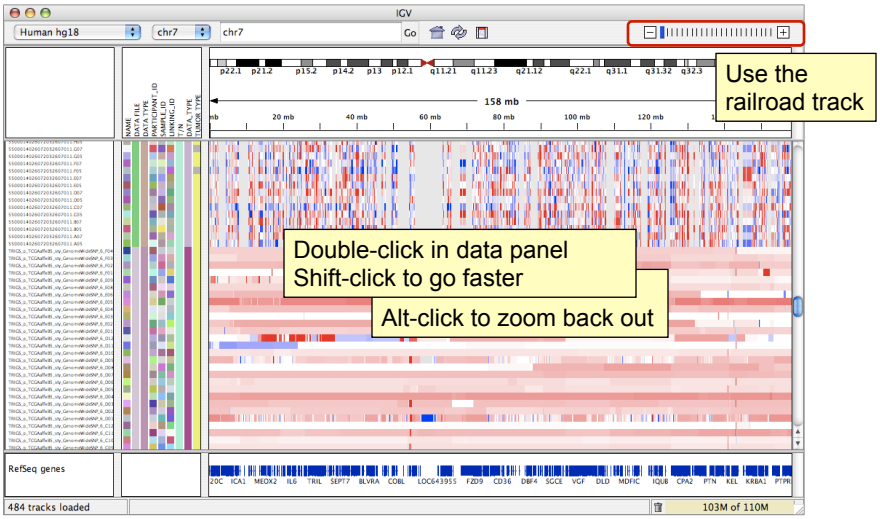


BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Zooming further in

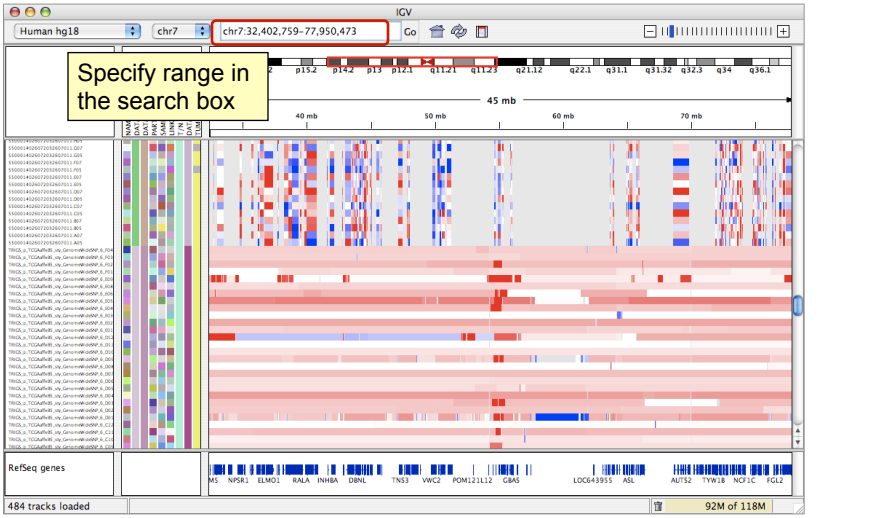


BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Zooming further in

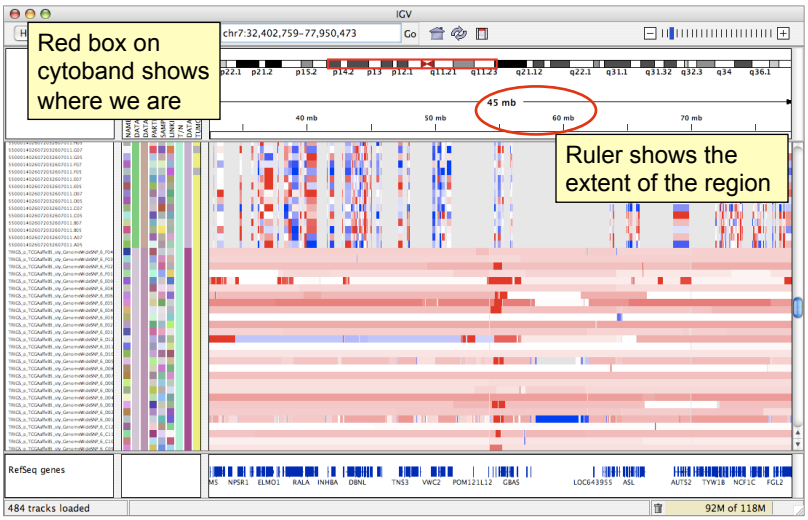


BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Zooming further in



BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Scroll or jump to location at same zoom level

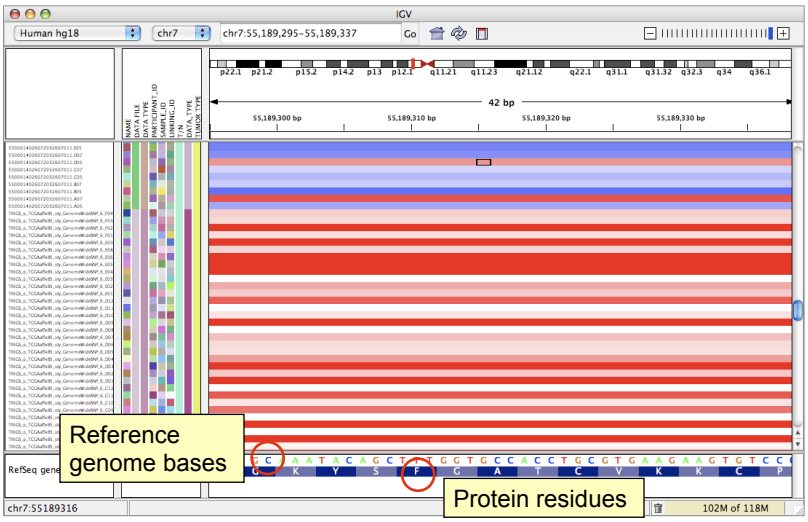
Use keyboard
(1) arrow keys
(2) Page Up, Page Down, Home, End

BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Zoomed in to base pair view

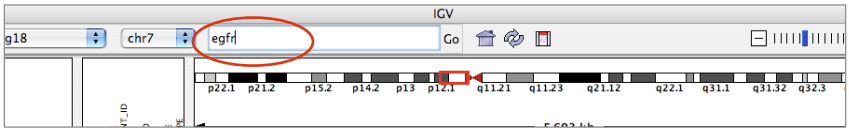


BROAD INSTITUTE

Navigating through the data

Integrative Genomics Viewer

Jump to feature



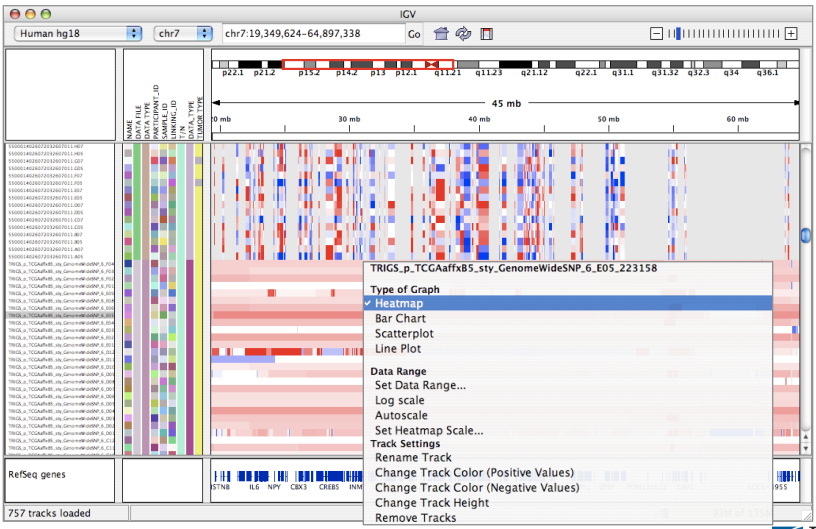
- Enter name of feature in search box
  - With or without zoom (View > Preferences > General)
- Click on a feature track (e.g. gene track, BED, GFF)
  - Ctrl+F = jump forward to next feature
  - Ctrl+B = jump backward to previous feature

BROAD INSTITUTE

Setting track attributes

Integrative Genomics Viewer

Right-click popup menu

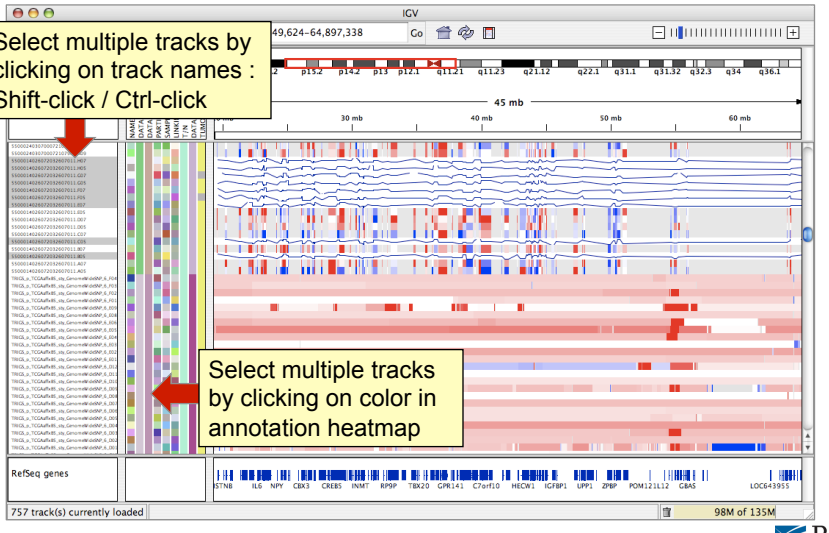


BROAD INSTITUTE

Setting track attributes

Integrative Genomics Viewer

Multiple tracks

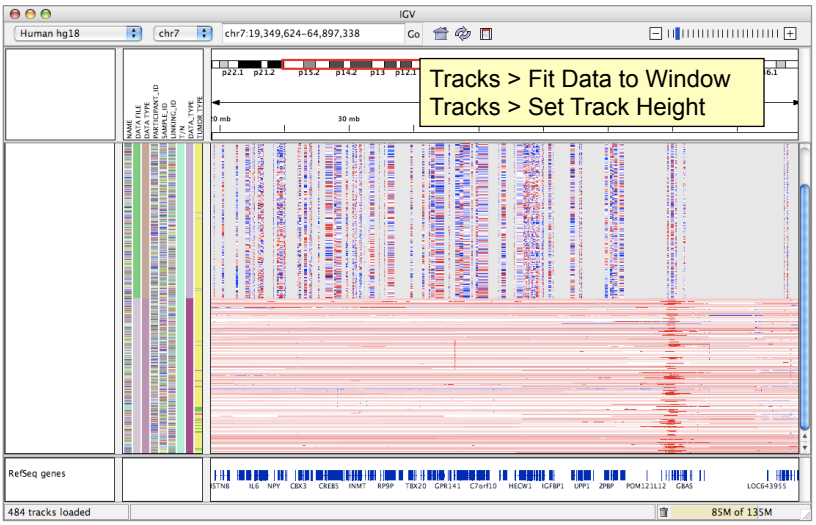


BROAD INSTITUTE

Setting track attributes

Integrative Genomics Viewer

Global attributes

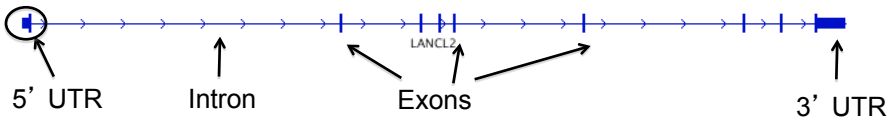


BROAD INSTITUTE

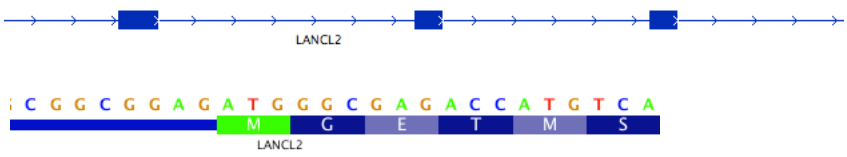
Annotation track

Integrative Genomics Viewer

Gene representation



Zoomed in views



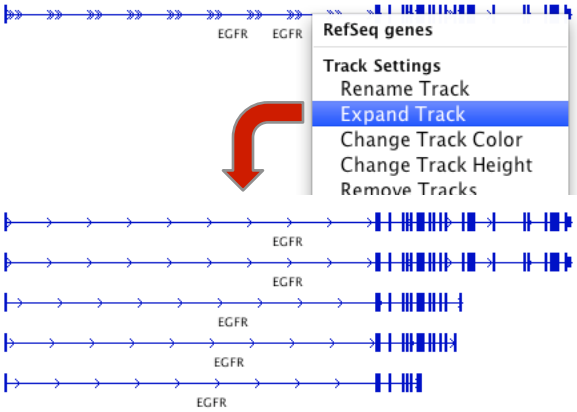
BROAD INSTITUTE

Annotation display mode

Integrative Genomics Viewer

1. Features are drawn in a single row, by default

2. Expand the track using the popup menu

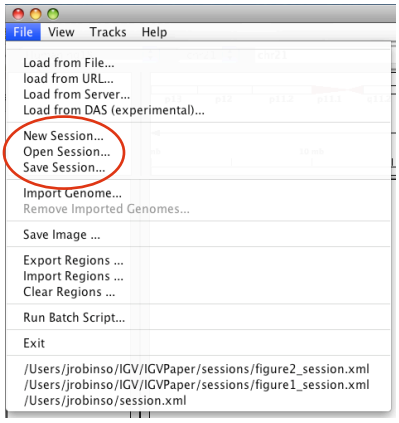


BROAD INSTITUTE

Sessions

Integrative Genomics Viewer

- Save current state of IGV to a named session file.
- Use to
  - restore the same state
  - share session with colleagues



BROAD INSTITUTE

File Formats

BROAD INSTITUTE

Integrative Genomics Viewer

File formats

Integrative Genomics Viewer

- Annotation File Formats
- Data File Formats
- Track Line
- Genomes and FASTA Files

BROAD INSTITUTE

Annotation File formats

•BED - UCSC standard format. Useful for displaying any feature type from simple blocks to genes.
http://genome.ucsc.edu/FAQ/FAQformat.html#format1

•GFF – Two variants, GFF2 and GFF3. Can also be used all feature types, tends to be more verbose and slower to parse than BED. File sizes can be significantly larger.
http://www.sequenceontology.org/gff3.shtml

•Note: BED file coordinates are "zero-based half-open". This means an interval spanning the first base is represented as 0-1. GFF files are "one-based open". An interval spanning the first base is represented as 1-1. This difference is responsible for many off-by-one bugs.

BROAD INSTITUTE

Data File formats

•Single Track Formats

- WIG – for fixed or variable step data with fixed spans
  http://genome.ucsc.edu/goldenPath/help/wiggle.html
- BEDGraph – similar to BED format
  http://genome.ucsc.edu/goldenPath/help/bedgraph.html

BROAD INSTITUTE

UCSC track line

A track line can be used to control many aspects of the track display such as graph type, color, and scale.

Can be used with wig, bed, gff, igv, cn, and gct files.

Line begins with "track" for wig and bed, "#track" for other formats.

Track line consists of key=value pairs, separated by a single space

Example:
track name="my custom track" graphType=bar color=255,0,0

Viewing NGS Data

Next-generation sequencing

Integrative Genomics Viewer

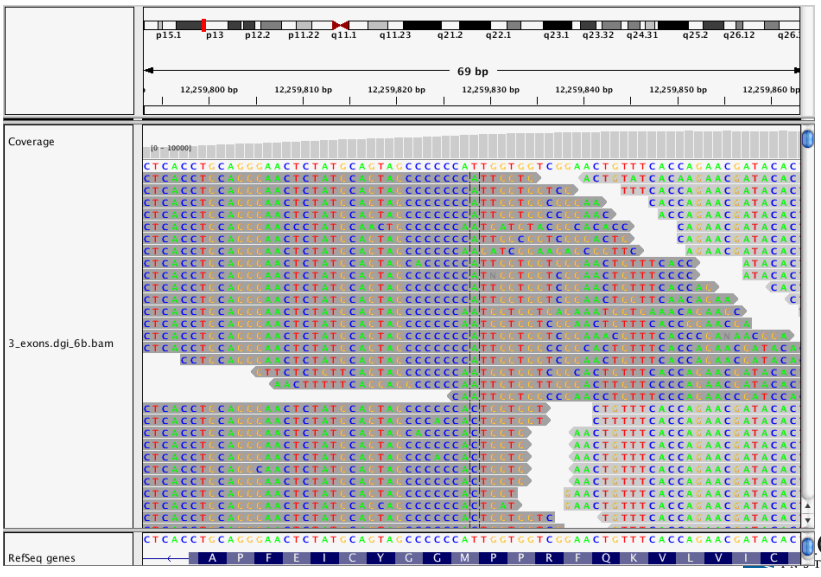
The size of NGS datasets presents many challenges, including:

- Implementation
  - Managing terabyte size files with modest compute resources (desktop computers).
- Visual design
  - Highlight events of interest
  - Deemphasize irrelevant details
  - Avoid information overload

BROAD INSTITUTE

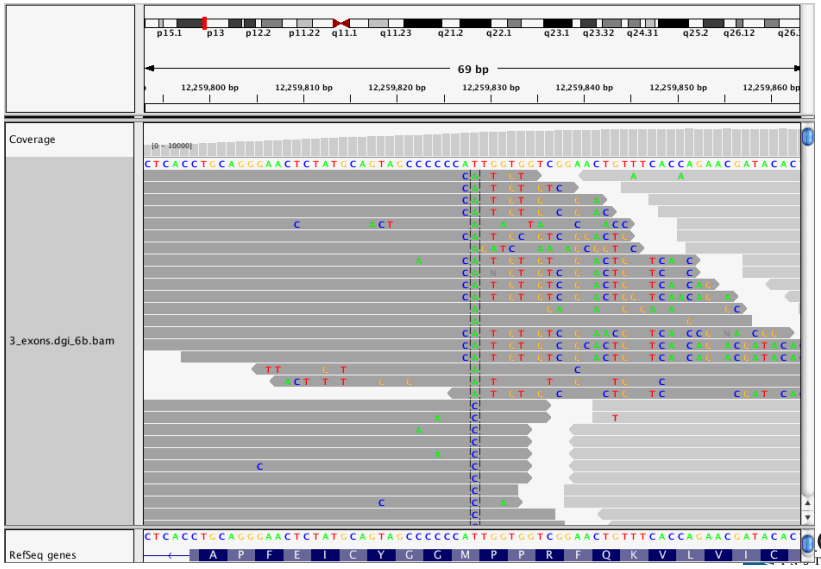
Aligned reads – all bases

Integrative Genomics Viewer

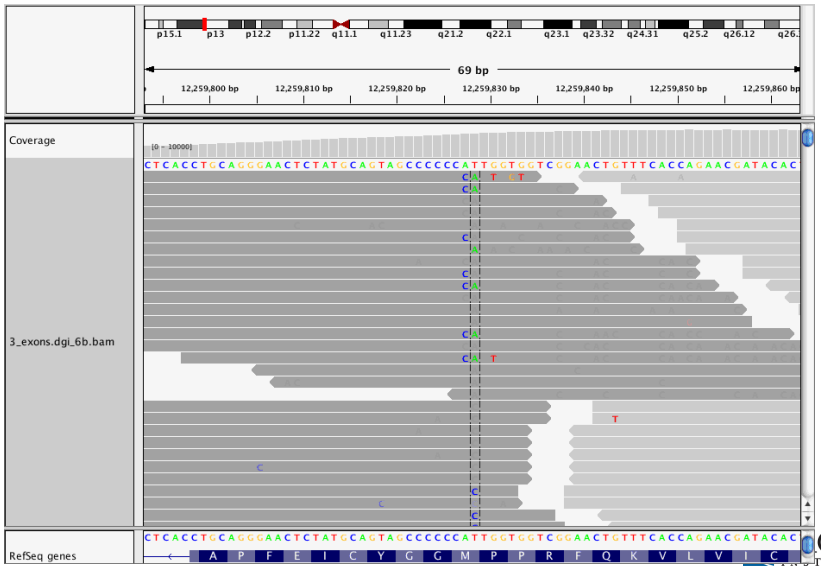


BROAD INSTITUTE

Aligned reads - mismatches



Aligned reads – base quality



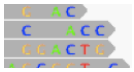
Vary view by resolution scale

Integrative Genomics Viewer

Whole chromosome -- calculated summary data, e.g. coverage.

~ 50-100 kb -- putative rearrangements, SNPs

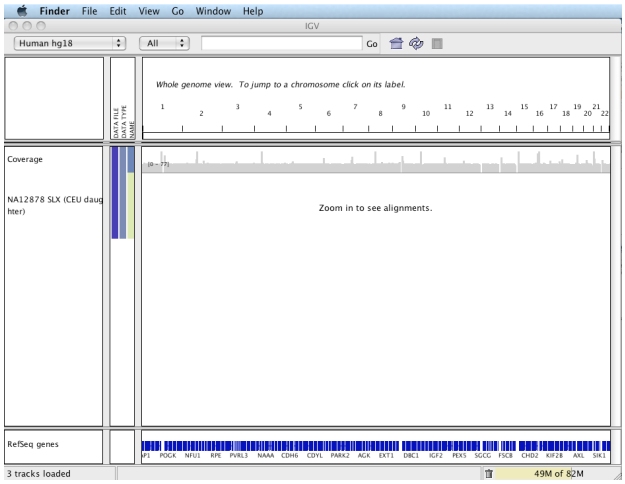
~ 500 bp -- bases



BROAD INSTITUTE

Viewing NGS data

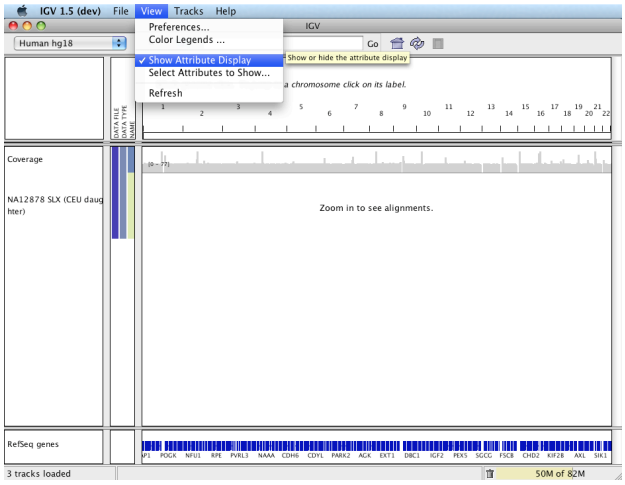
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

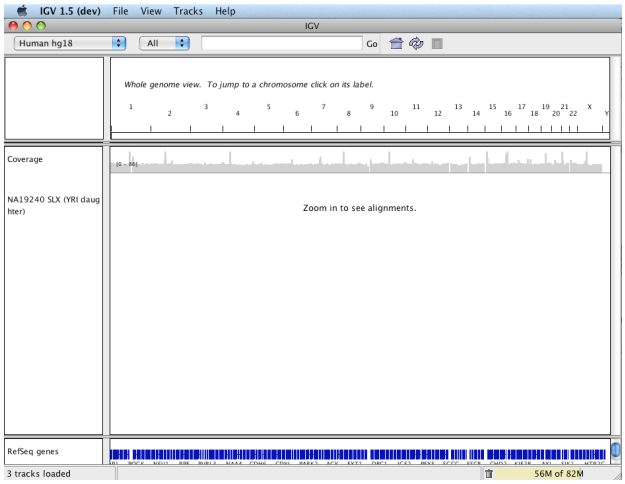
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

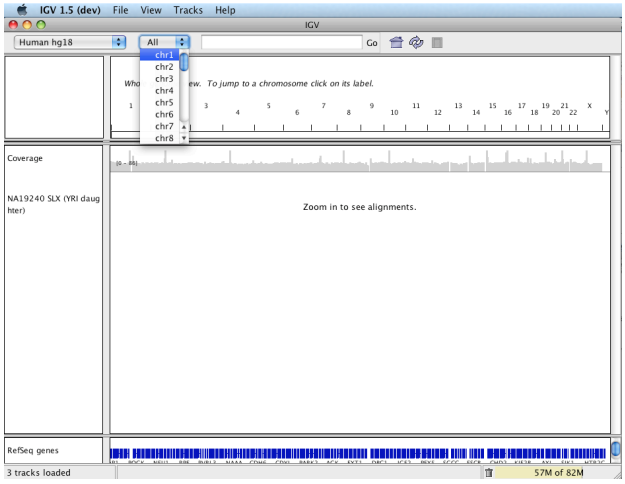
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

Integrative Genomics Viewer

BROAD INSTITUTE

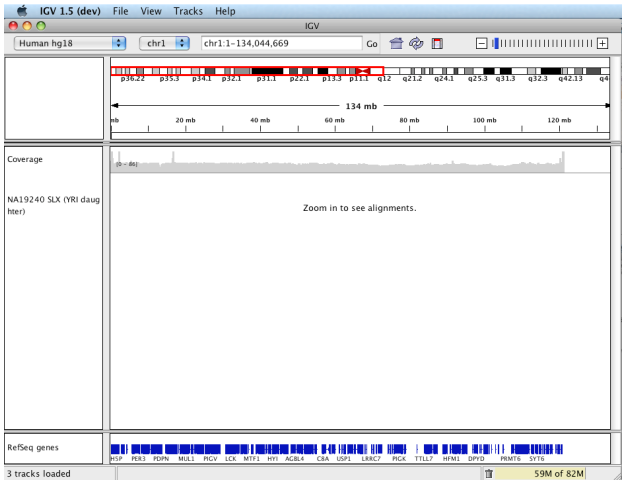
Viewing NGS data

Integrative Genomics Viewer

BROAD INSTITUTE

Viewing NGS data

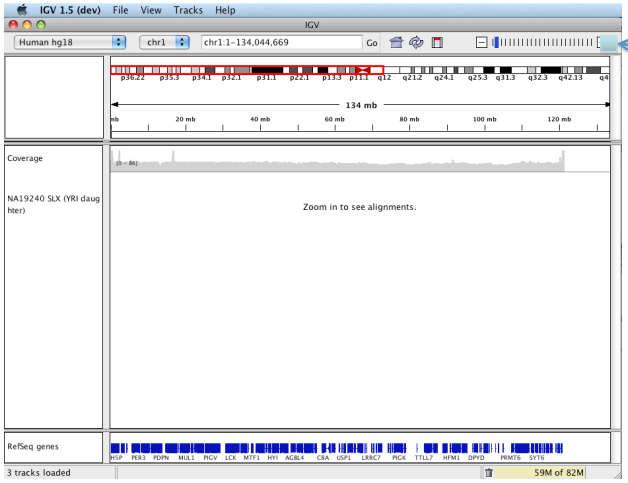
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

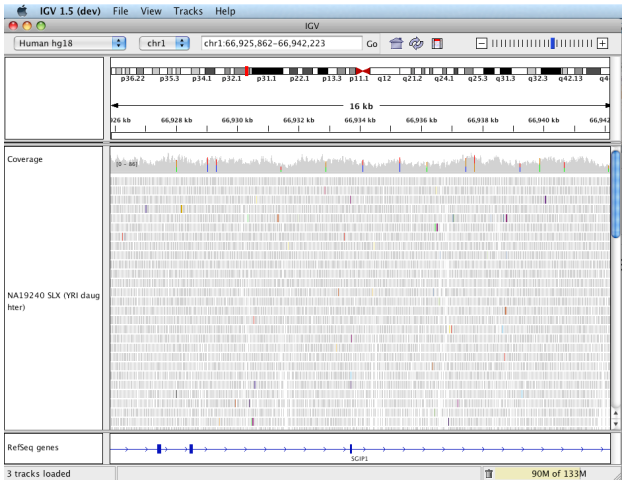
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

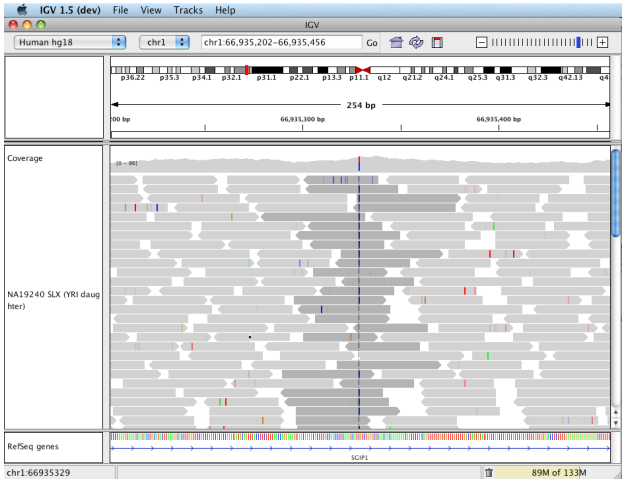
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

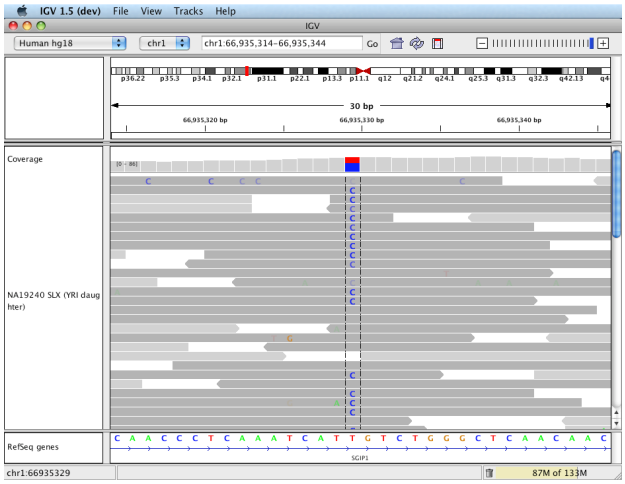
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

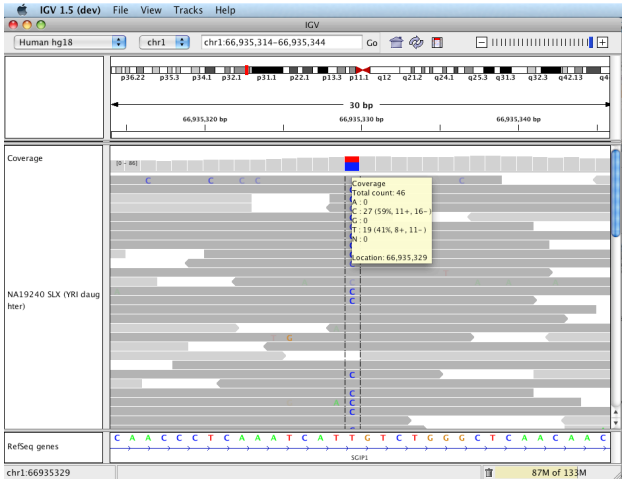
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

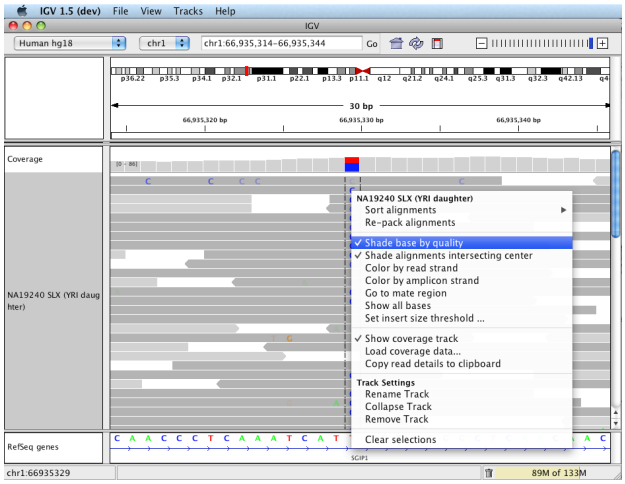
Integrative Genomics Viewer



BROAD INSTITUTE

Viewing NGS data

Integrative Genomics Viewer



BROAD INSTITUTE

IGVTools

BROAD INSTITUTE

Integrative Genomics Viewer

IGVTools

Integrative Genomics Viewer

IGVTools is a set of utilities for preparing large files for efficient display.

tile: converts a sorted data input file to a binary tiled data (.tdf) file.
Supported input file formats: .wig, .cn, .snp, .igv, .gct

count: computes average alignment or feature density for over a specified window size across the genome.
Supported input file formats: .sam, .bam, .aligned, .sorted.txt, .bed
sort: sorts the input file by start position. Supported input file formats: .cn, .igv, .sam, .aligned, and .bed.

index: creates an index file for an input ascii alignment file.
Supported input file formats: .sam, .aligned, .sorted.txt

BROAD INSTITUTE

IGVTools tile

The *tile* utility converts large ascii data files into tiled data format (.tdf) files. TDF files have the following advantages

1.Data is is indexed for efficient retrieval.

2.Data for zoomed out views are preprocessed.

3.TDF files are web friendly, large data files can be shared over the web. Only small slices of the file are actually transferred as needed.

IGVTools count

The count command is used to transform alignment files to read density TDF files, e.g. for ChIP-Seq, RNA-Seq, & similar alignment counting experiments.

Alignments

Alignments in bam/sam, .aligned, or bed format.

Read density

"Tiled Data File" indexed and optimized for fast retrieval at multiple resolution scales

IGVTools sort

This utility sorts IGV supported genomic formats by start position.

Example:
igvtools sort -m 1000000 –t ~/myTmpDir inputFile.sam outputFile.sorted.sam

The sort command uses a combination of memory and disk to handle large files.

-m = maximum # of lines to hold in memory. When this number is exceeded a temporary file is created.

-t = directory used to create temporary files during sorting.

BROAD INSTITUTE

IGVTools index

Used to create an index file for viewing SAM (not BAM) files

Note: to be confused with the *samtools* index, which is used to create an index for BAM files

SAM => igvtools
BAM => samtools

Example: igvtools index inputFile.sam

Result inputFile.sam.sai

The index file must remain with the sam file to be found, IGV just appends .sai to the end.

BROAD INSTITUTE