



Cold  
Spring  
Harbor  
Laboratory

# Advanced Sequencing Technologies & Applications

<http://meetings.cshl.edu/courses.html>

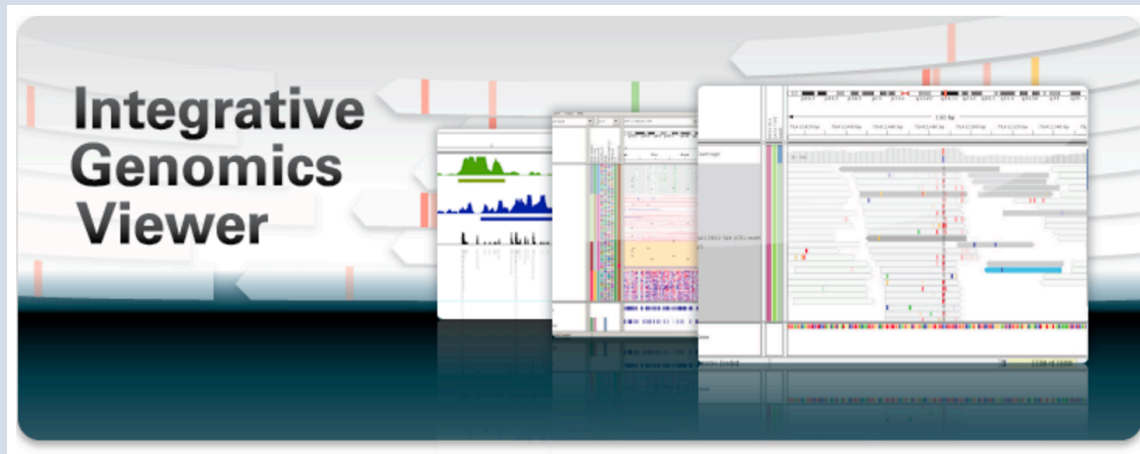


Cold  
Spring  
Harbor  
Laboratory

# Introduction to IGV The Integrative Genomics Viewer

Kelsy Cotto, Obi Griffith, Malachi Griffith,  
Alex Wagner, Jason Walker

Advanced Sequencing Technologies & Applications  
November 6- 18, 2018



# Visualization Tools in Genomics

- there are **over 40 different genome browsers**, which to use?
- depends on
  - task at hand
  - kind and size of data
  - data privacy

# HT-seq Genome Browsers



Integrative  
Genome  
Viewer



UCSC  
Genome Browser  
Cancer Genome Browser



Trackster  
(part of Galaxy)



Savant  
Genome  
Browser

- task at hand : visualizing HT-seq reads, especially good for inspecting variants
- kind and size of data : large BAM files, stored locally or remotely
- data privacy : run on the desktop, can keep all data private
- UCSC Genome Browser has been retro-fitted to display BAM files
- Trackster is a genome browser that can perform visual analytics on small windows of the genome, deploy full analysis with Galaxy



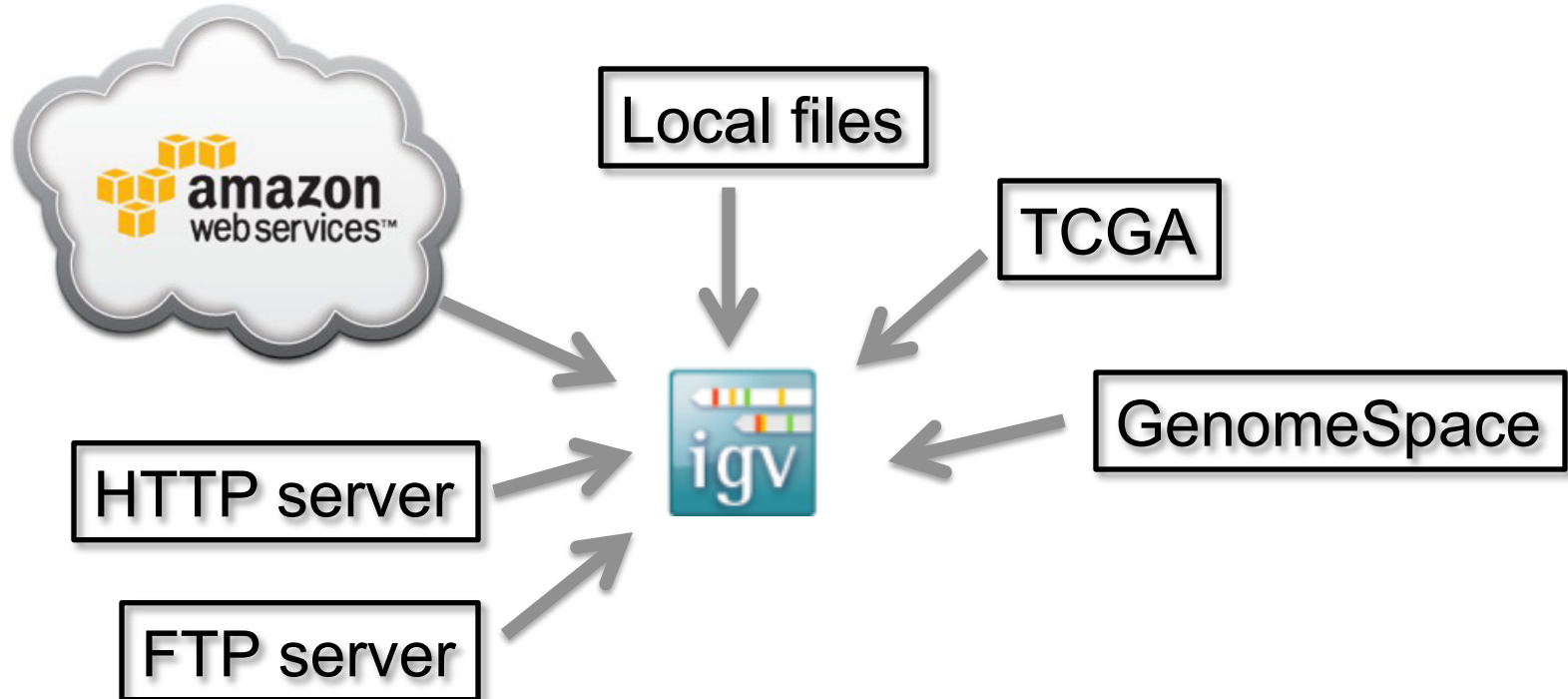


# Features

With IGV you can...

- Explore large genomic datasets with an intuitive, easy-to-use interface.
- Integrate multiple data types with clinical and other sample information.
- View data from multiple sources:
  - local, remote, and “cloud-based”.
- Automation of specific tasks using command-line interface

# IGV data sources

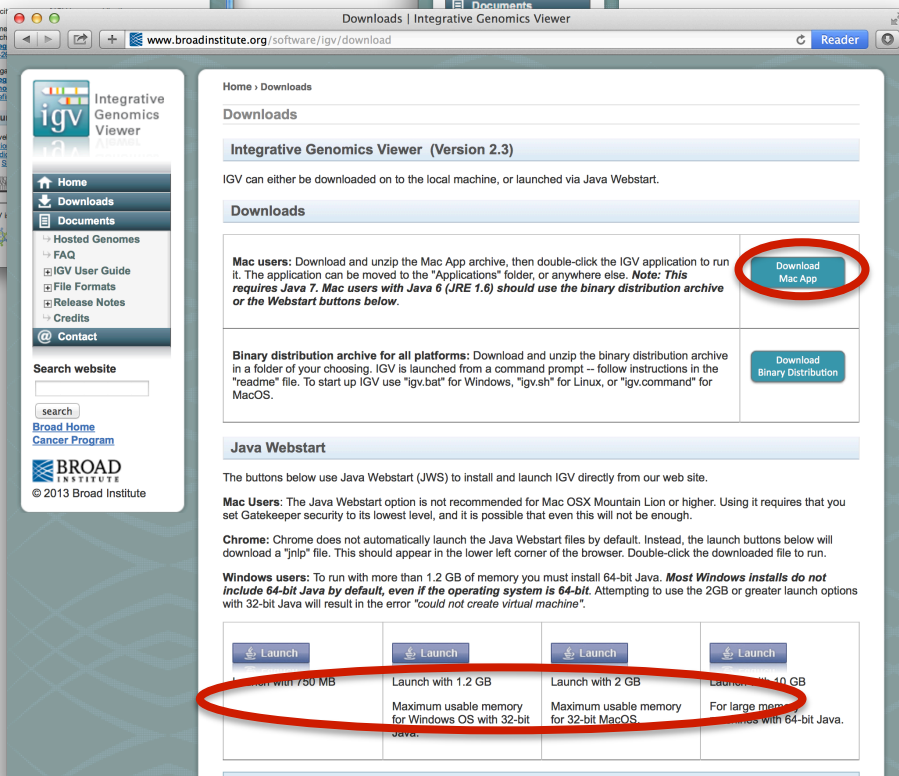
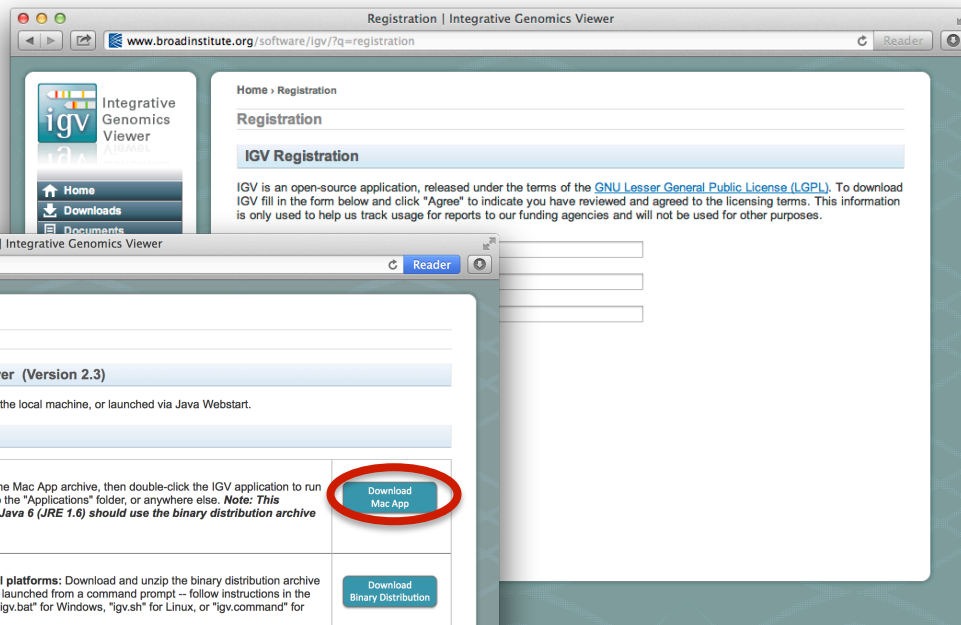
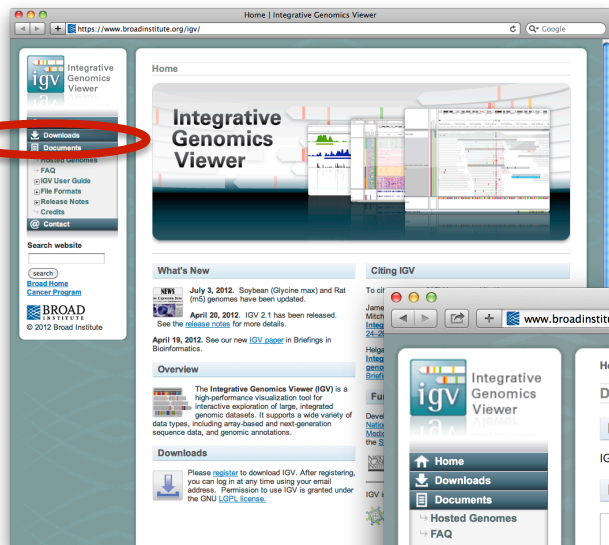


- View **local** files without uploading.
- View **remote** files without downloading the whole dataset.

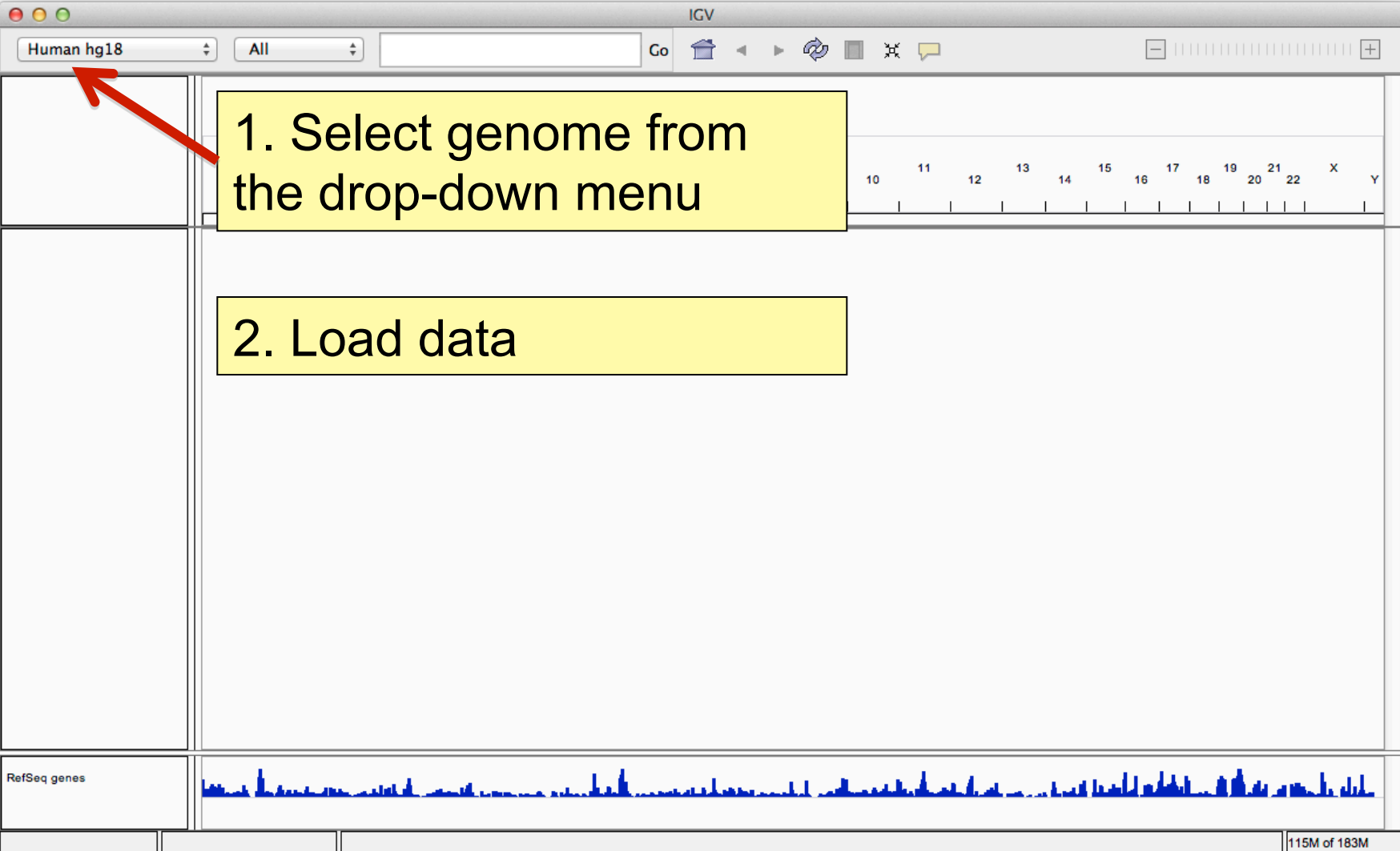
# Using IGV: the basics

- Launch IGV
- Select a reference genome
- Load data
- Navigate through the data
  - WGS data
    - SNVs
    - structural variations

# Launch IGV



# Launch IGV



The screenshot shows the IGV web interface. At the top, there is a header with the text "IGV". Below the header, there is a navigation bar containing a dropdown menu with "Human hg18" selected, a "Go" button, and several navigation icons. A red arrow points to the "Human hg18" dropdown menu. Below the navigation bar, there is a large white area with a yellow box containing the text "1. Select genome from the drop-down menu". Below this, there is another yellow box containing the text "2. Load data". At the bottom of the interface, there is a track labeled "RefSeq genes" showing a blue bar chart. In the bottom right corner, there is a status bar showing "115M of 183M".

1. Select genome from the drop-down menu

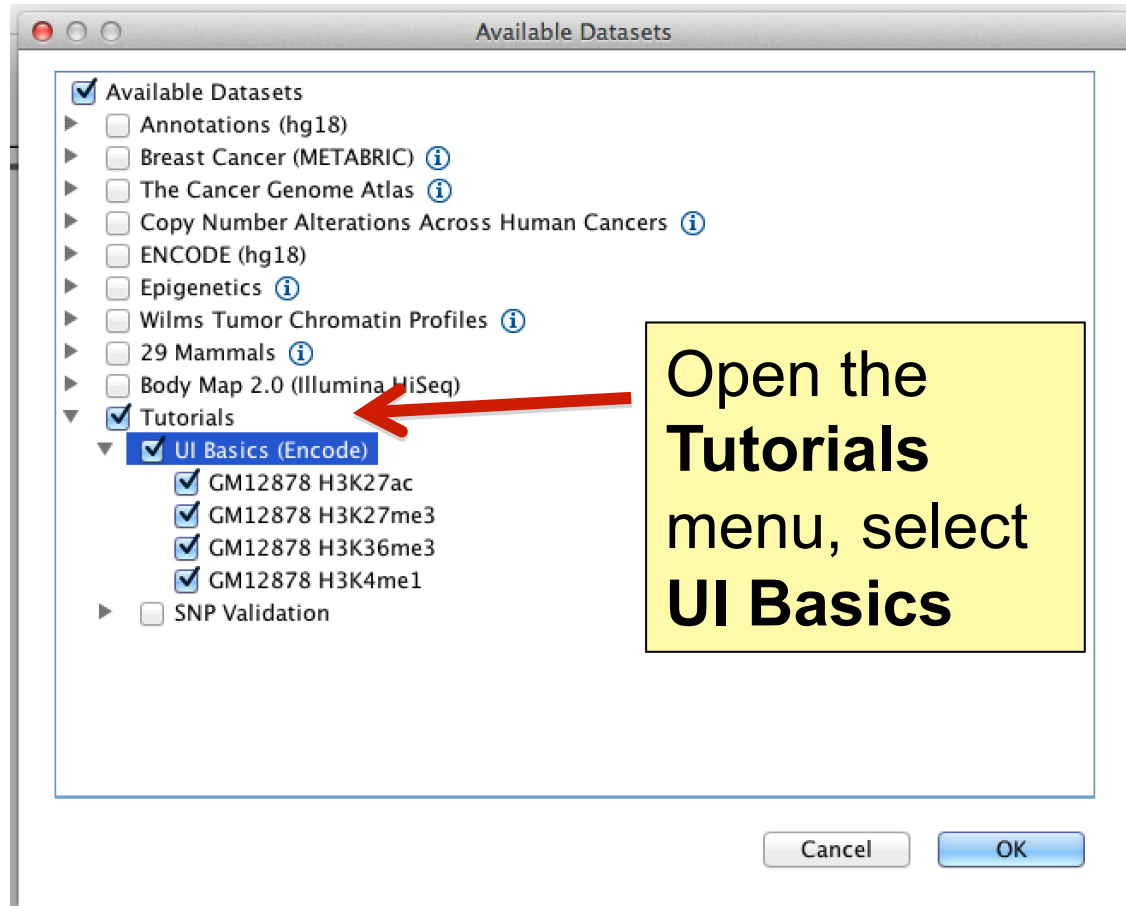
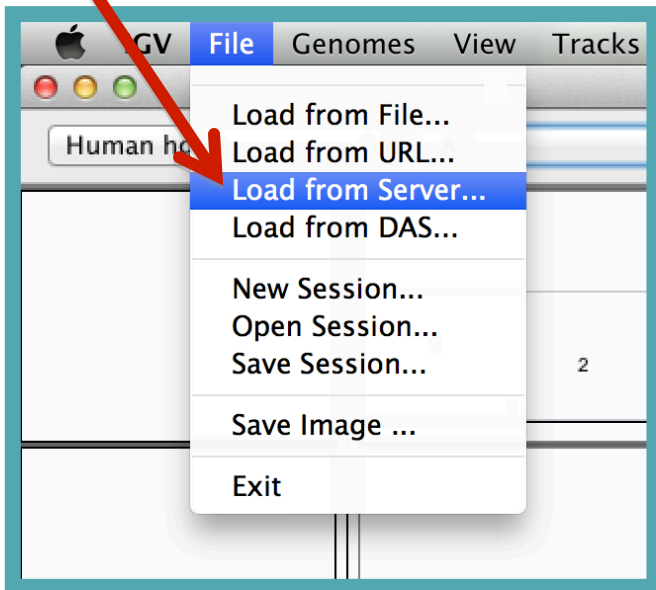
2. Load data

RefSeq genes

115M of 183M

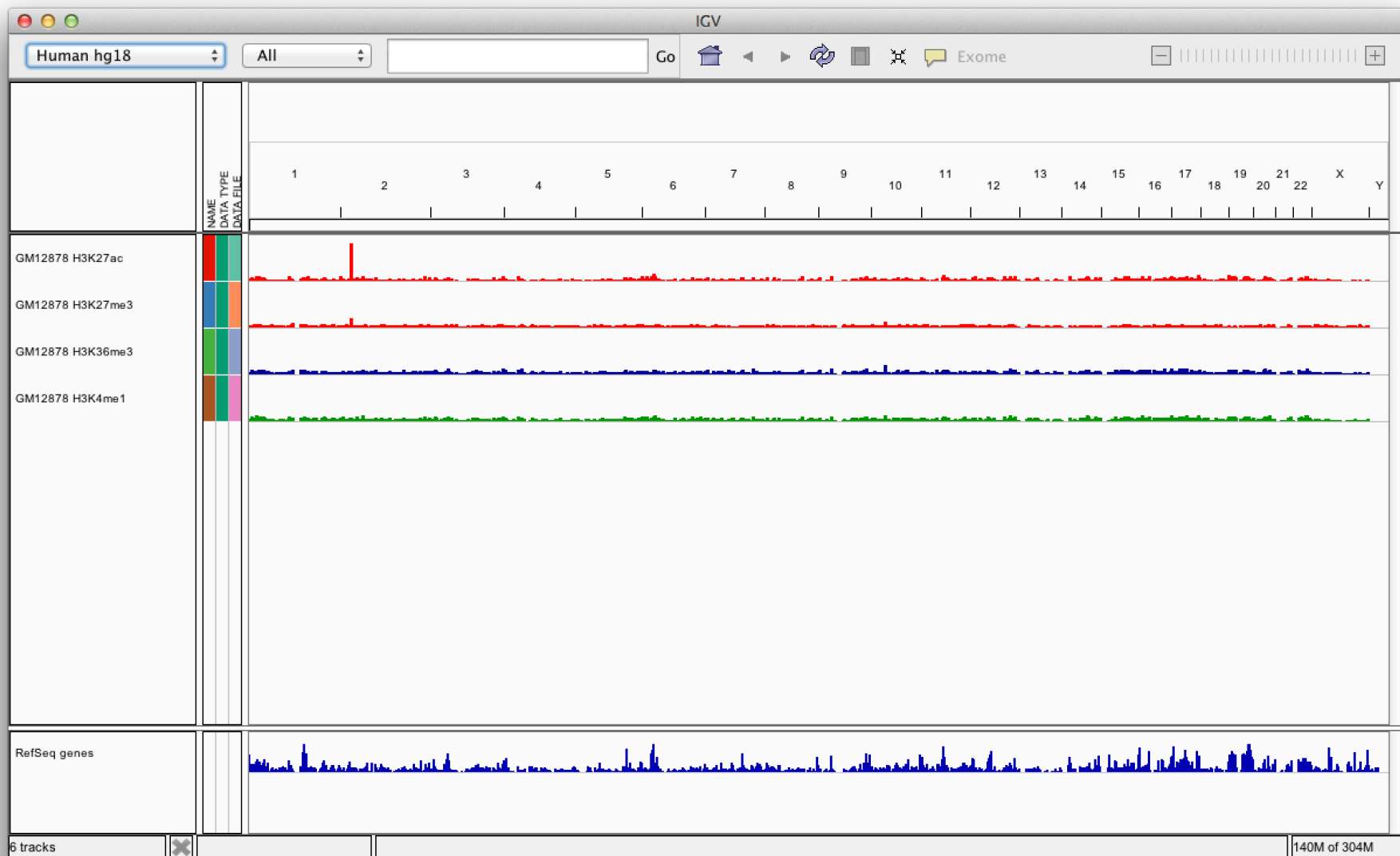
# Load data

Select File > Load from Server...



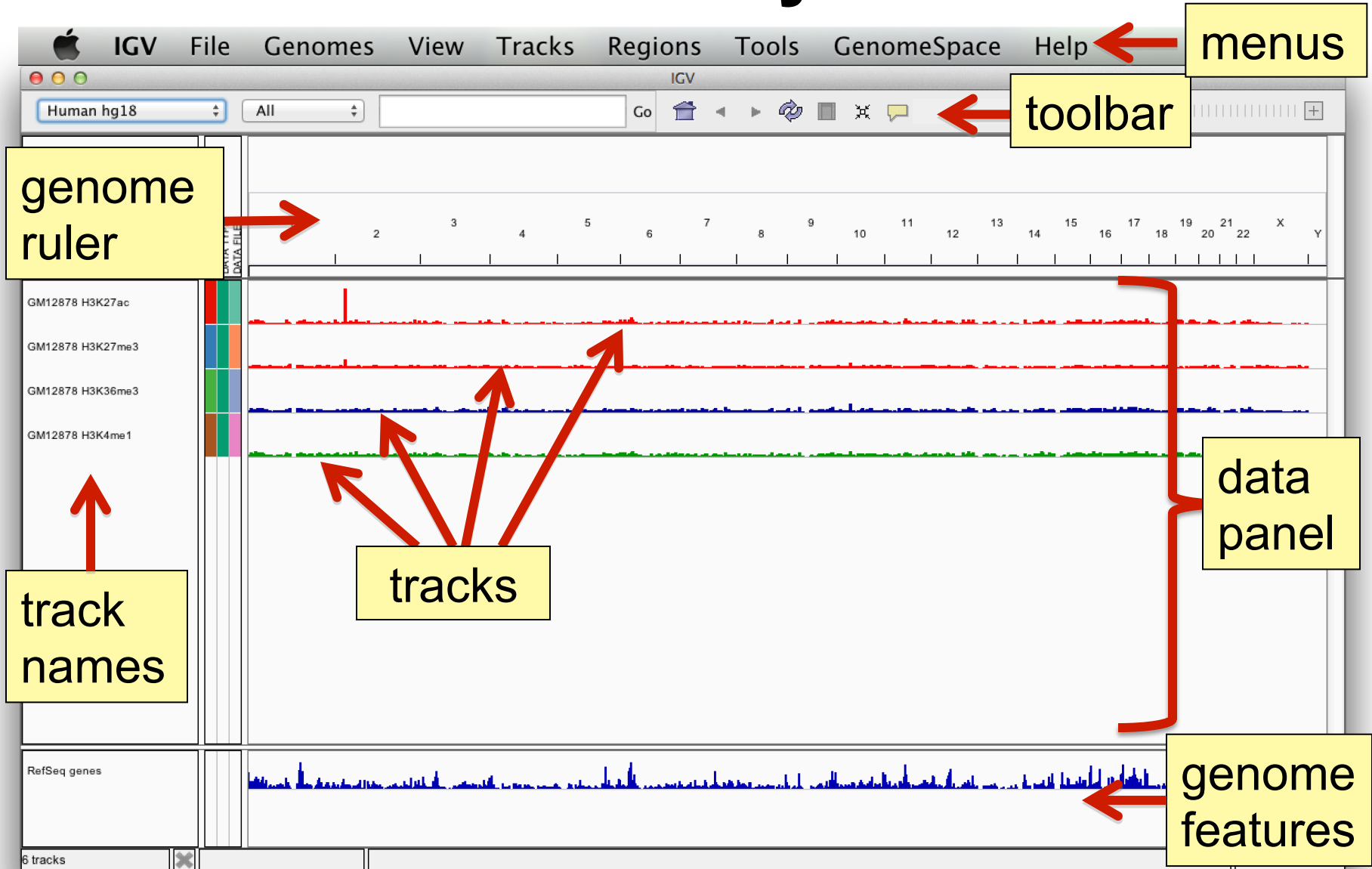
Open the Tutorials menu, select UI Basics

# Screen layout





# Screen layout

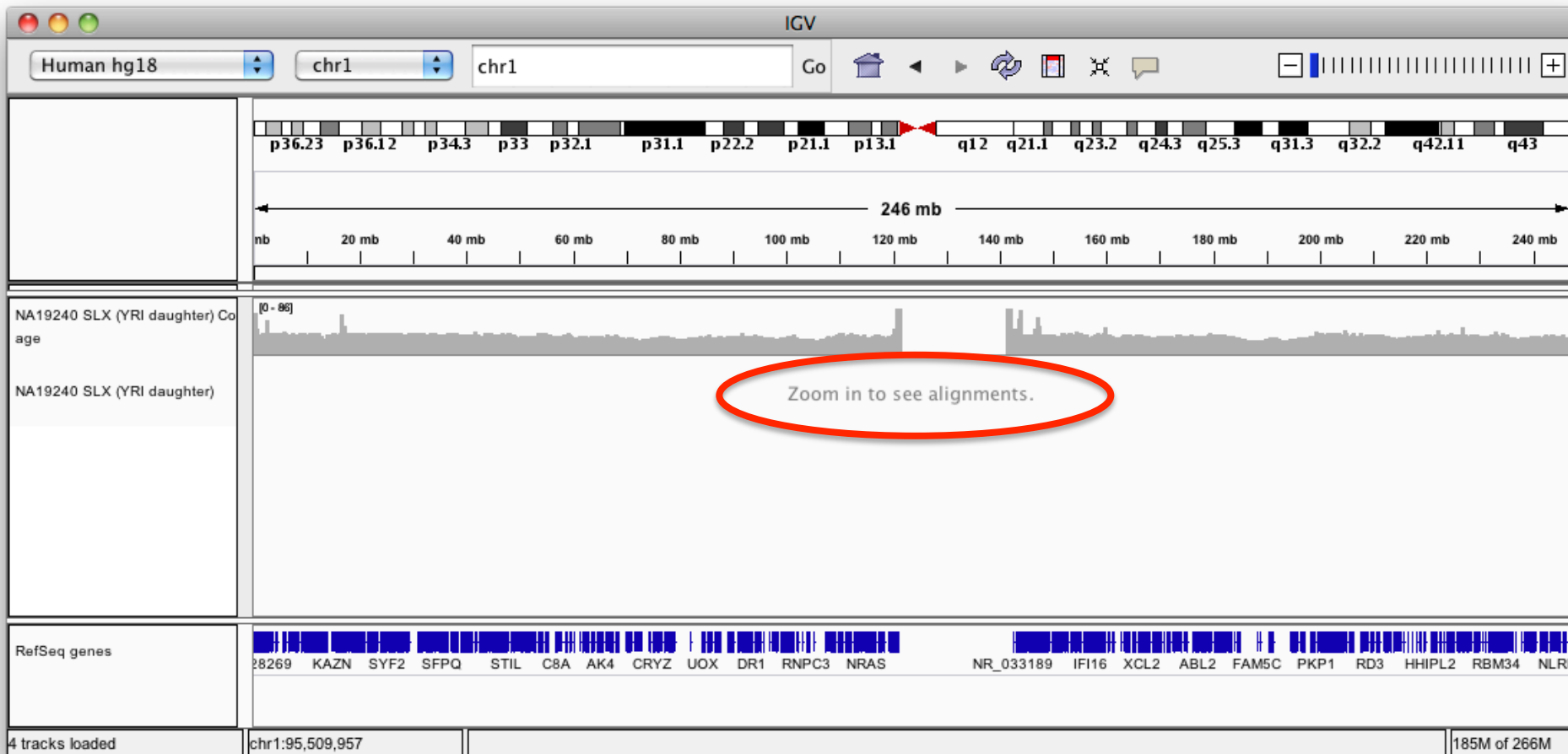


# File formats and track types

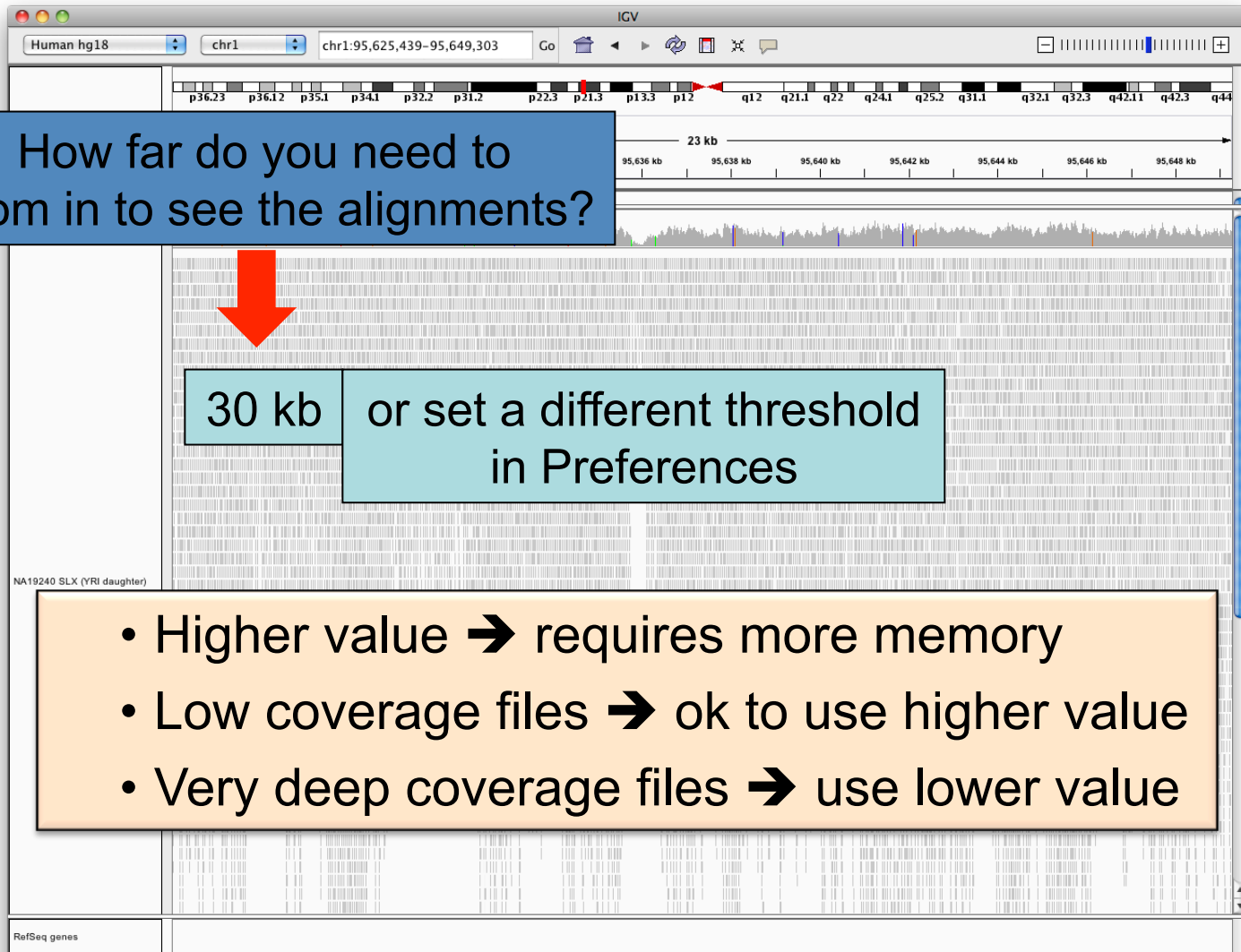
- The **file format** defines the track type.
- The **track type** determines the display options
  - [BAM](#)
  - [BED](#)
  - [BedGraph](#)
  - [bigBed](#)
  - [bigWig](#)
  - [Birdsuite Files](#)
  - [CBS](#)
  - [CN](#)
  - [Cufflinks Files](#)
  - [Custom File Formats](#)
  - [Cytoband](#)
  - [FASTA](#)
  - [GCT](#)
  - [genePred](#)
  - [GFF](#)
  - [GISTIC](#)
  - [Goby](#)
  - [GWAS](#)
  - [IGV](#)
  - [LOH](#)
  - [MAF](#)
  - [Merged BAM File \(.bam.list\)](#)
  - [MUT](#)
  - [PSL](#)
  - [RES](#)
  - [SAM](#)
  - [Sample Information](#)
  - [SEG](#)
  - [SNP](#)
  - [TAB](#)
  - [TDF](#)
  - [Track Line](#)
  - [Type Line](#)
  - [VCF](#)
  - [WIG](#)
- For current list see: [www.broadinstitute.org/igv/FileFormats](http://www.broadinstitute.org/igv/FileFormats)

# Viewing alignments

## Whole chromosome view



# Viewing alignments – Zoom in



How far do you need to zoom in to see the alignments?

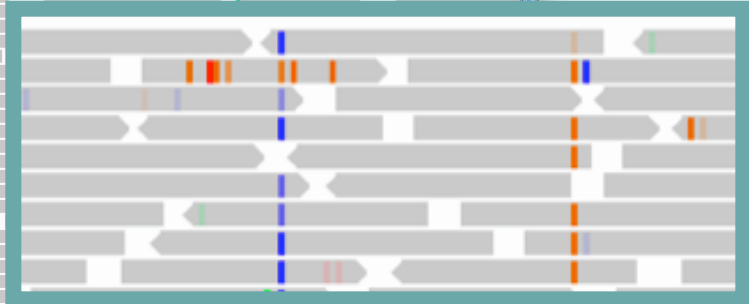
30 kb or set a different threshold in Preferences

- Higher value → requires more memory
- Low coverage files → ok to use higher value
- Very deep coverage files → use lower value

# Viewing alignments – Zoom in



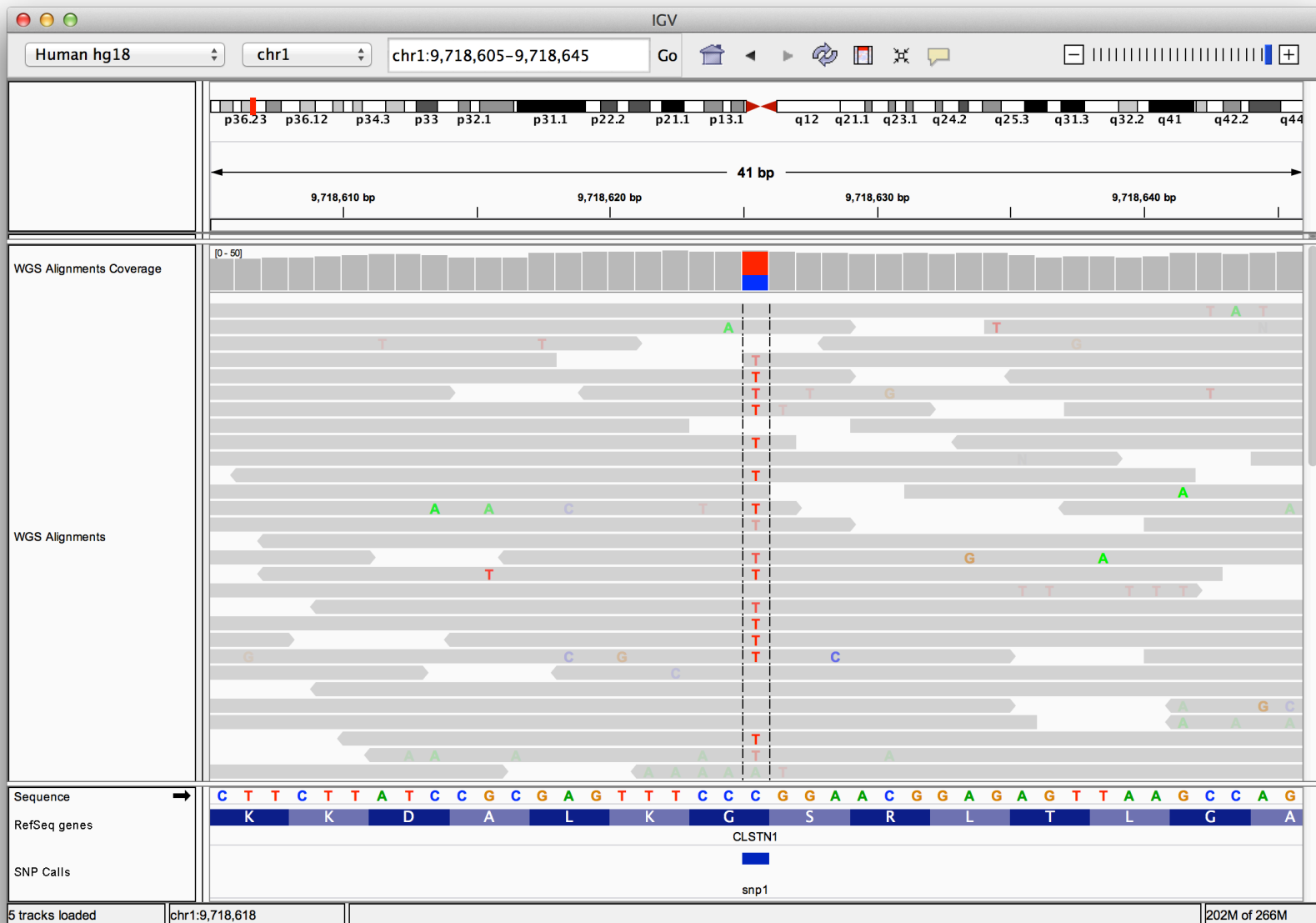
Bases that do not match the reference sequence are highlighted by color



# SNVs and Structural variations

- Important metrics for evaluating the validity of SNVs:
  - Coverage
  - Amount of support
  - Strand bias / PCR artifacts
  - Mapping qualities
  - Base qualities
- Important metrics for evaluating SVs:
  - Coverage
  - Insert size
  - Read pair orientation

# Viewing SNPs and SNVs



# Viewing SNPs and SNVs





# Viewing Structural Events

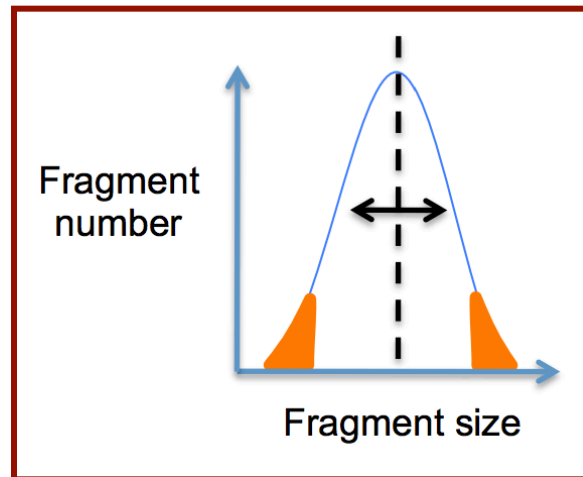
- Paired reads can yield evidence for genomic “structural events”, such as deletions, translocations, and inversions.
- Alignment coloring options help highlight these events based on:
  - Inferred insert size (template length)
  - Pair orientation (relative strand of pair)

# Paired-end sequencing

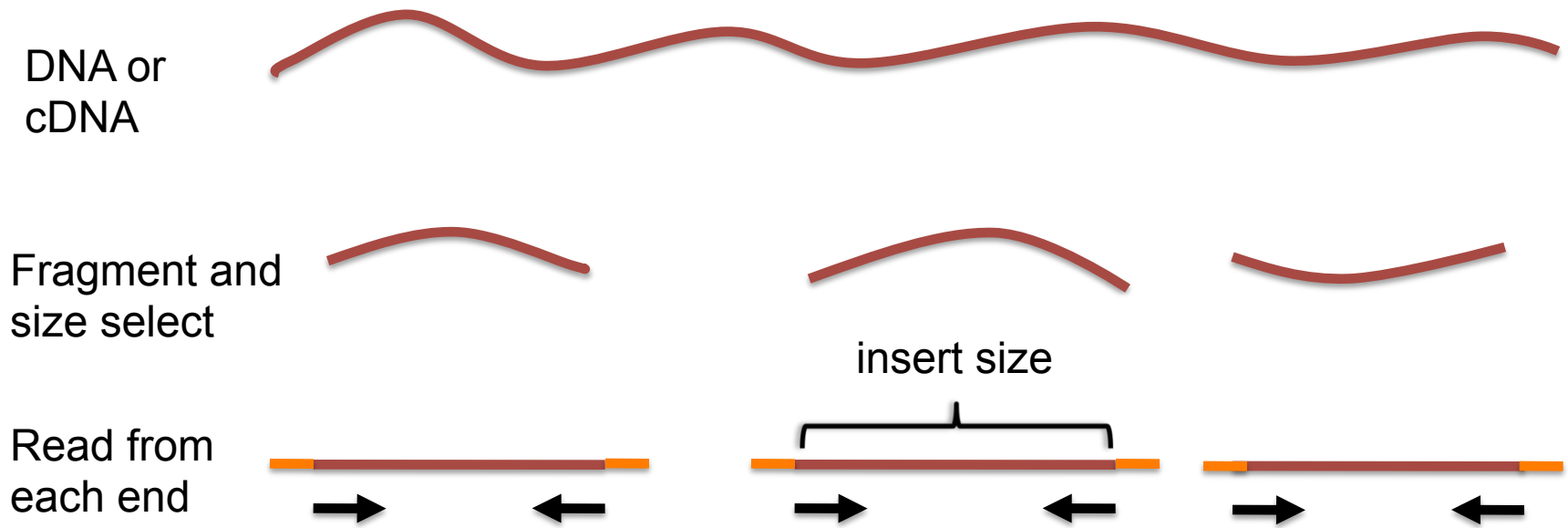
DNA or cDNA



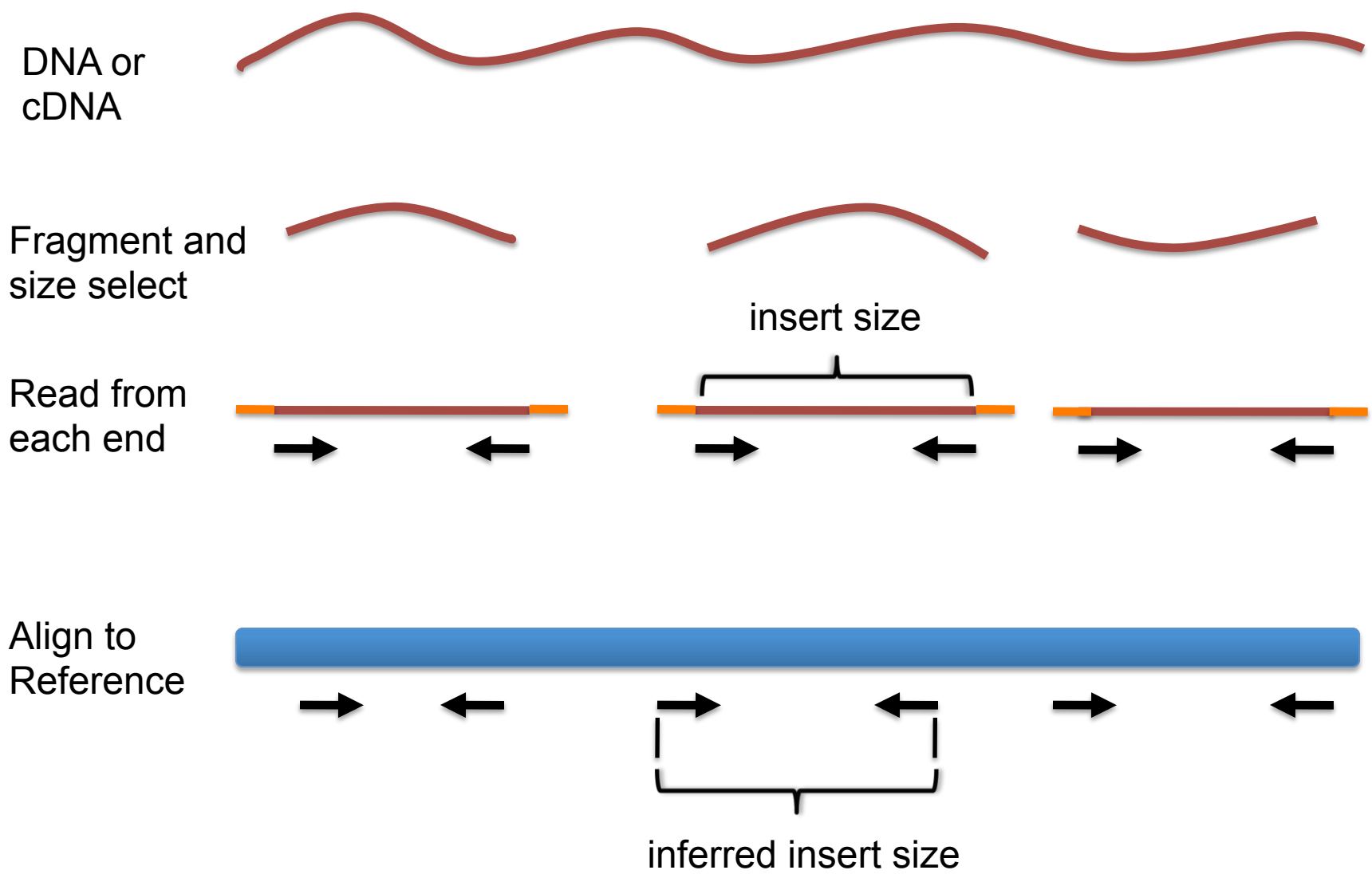
Fragment and size select



# Paired-end sequencing



# Paired-end sequencing



# Interpreting inferred insert size

The “inferred insert size” can be used to detect structural variants including

- Deletions
- Insertions
- Inter-chromosomal rearrangements: (Undefined insert size)

# Deletion

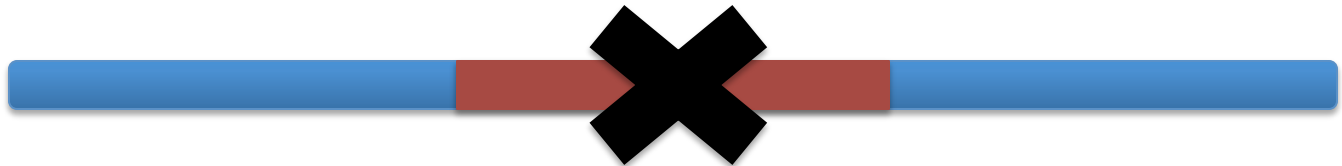
What is the effect of a deletion on inferred insert size?

# Deletion

Reference  
Genome



Subject



# Deletion

Reference  
Genome



Subject



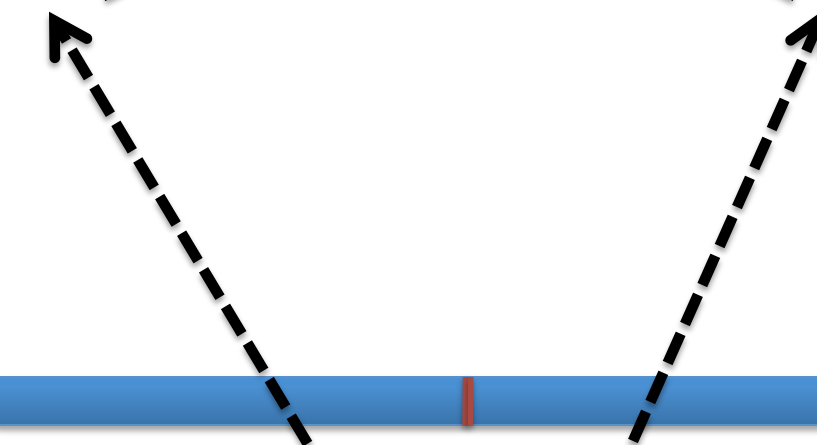


# Deletion

Reference  
Genome



Subject



# Deletion

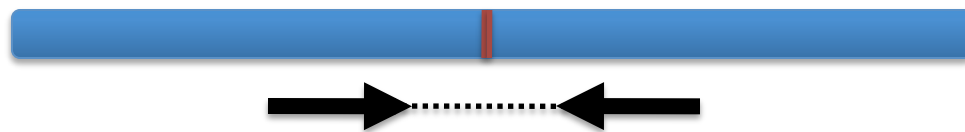
Inferred insert size is  $>$  expected value

Reference  
Genome



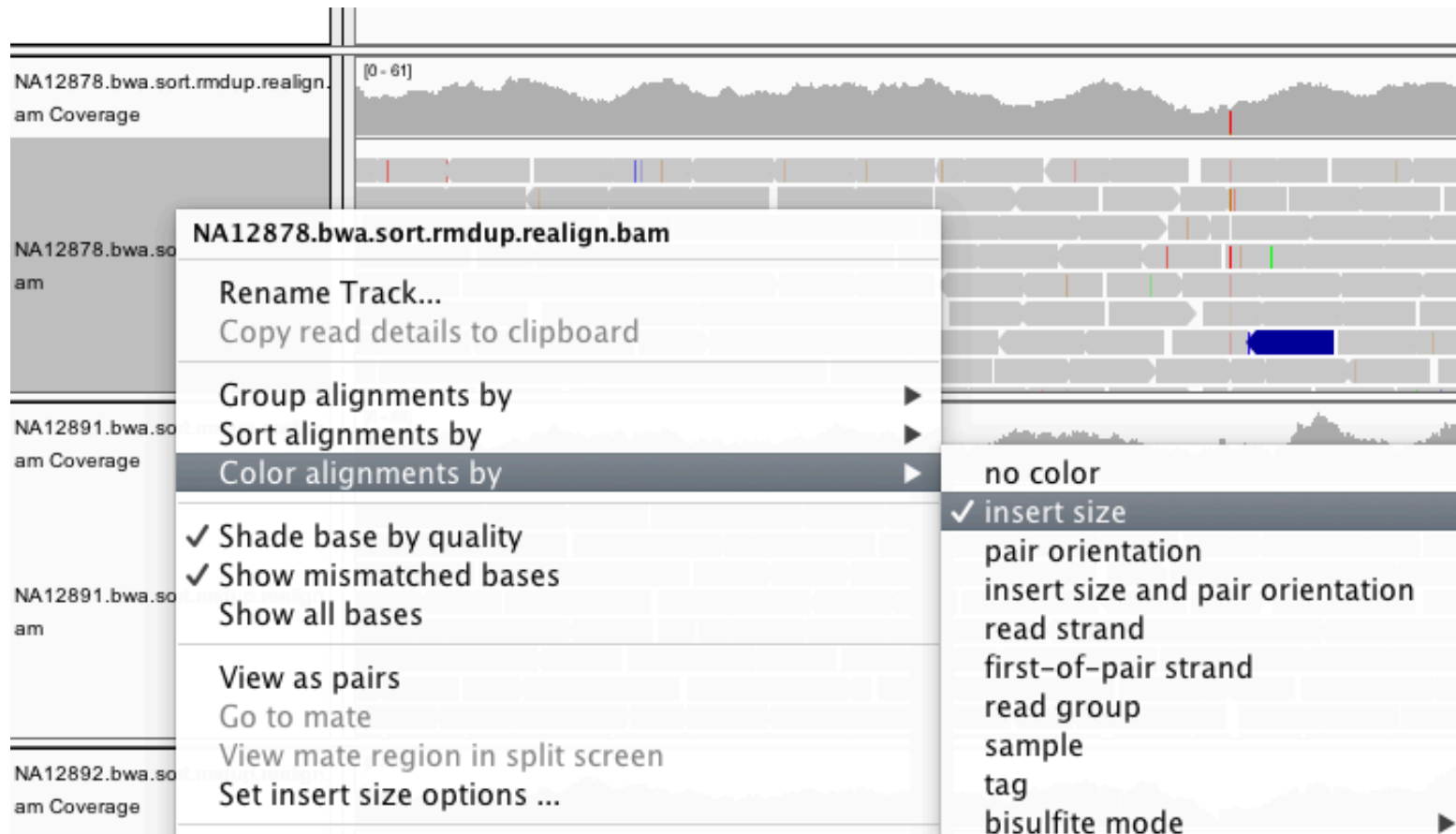
inferred insert size

Subject



expected insert size

# Color by insert size

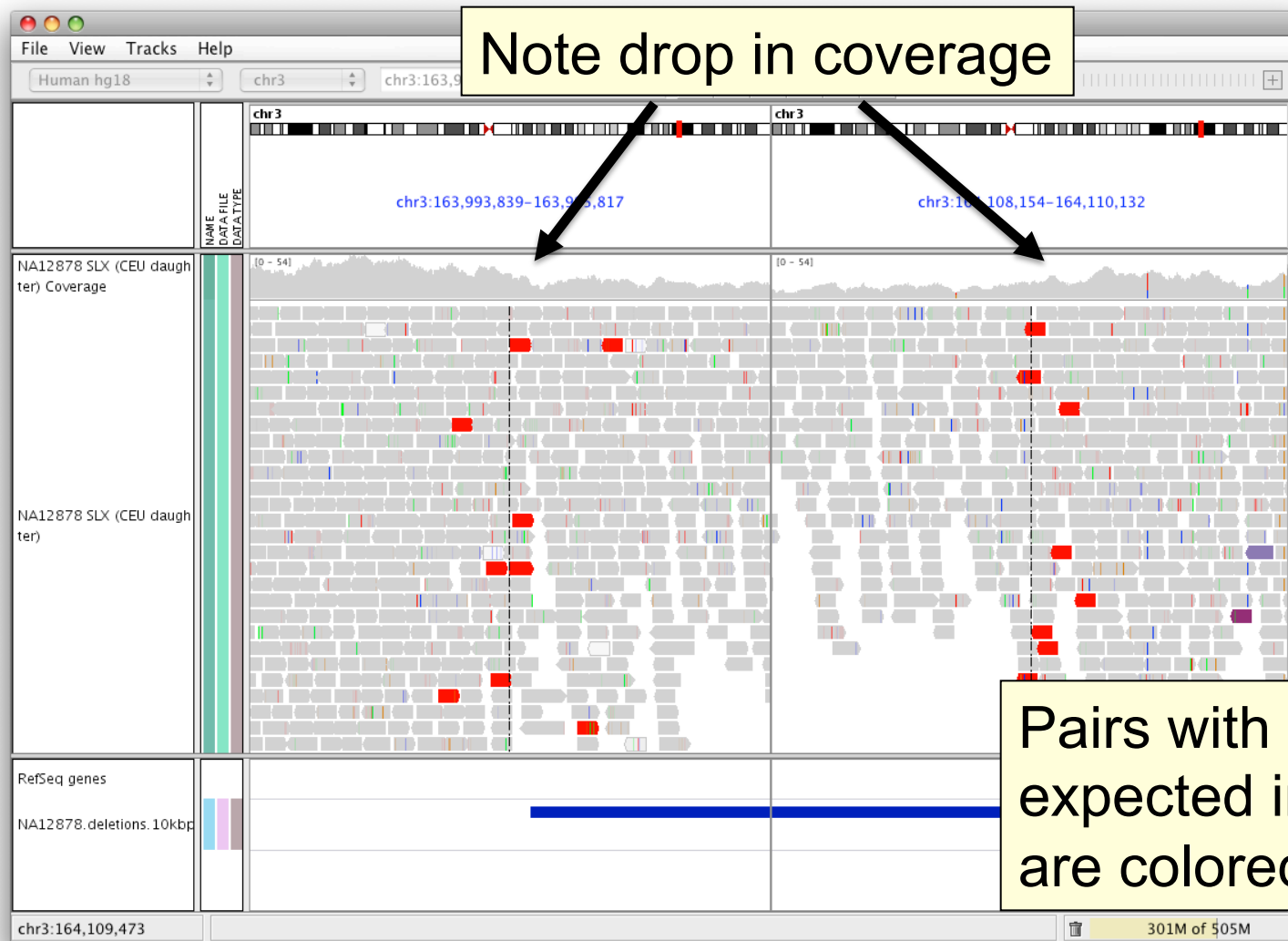


The screenshot shows a genome browser interface with a track titled "NA12878.bwa.sort.rmdup.realign.bam" selected. A context menu is open over the track, listing various options. The "Color alignments by" option is highlighted, and its sub-menu is also open, showing "insert size" as the selected option. The background shows a coverage plot and alignment tracks for several samples.



Context Menu Options:

- Rename Track...
- Copy read details to clipboard
- Group alignments by
- Sort alignments by
- Color alignments by**
  - no color
  - insert size**
  - pair orientation
  - insert size and pair orientation
  - read strand
  - first-of-pair strand
  - read group
  - sample
  - tag
  - bisulfite mode
- ✓ Shade base by quality
- ✓ Show mismatched bases
- Show all bases
- View as pairs
- Go to mate
- View mate region in split screen
- Set insert size options ...

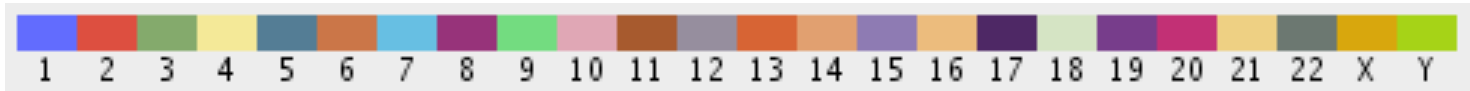
# Deletion



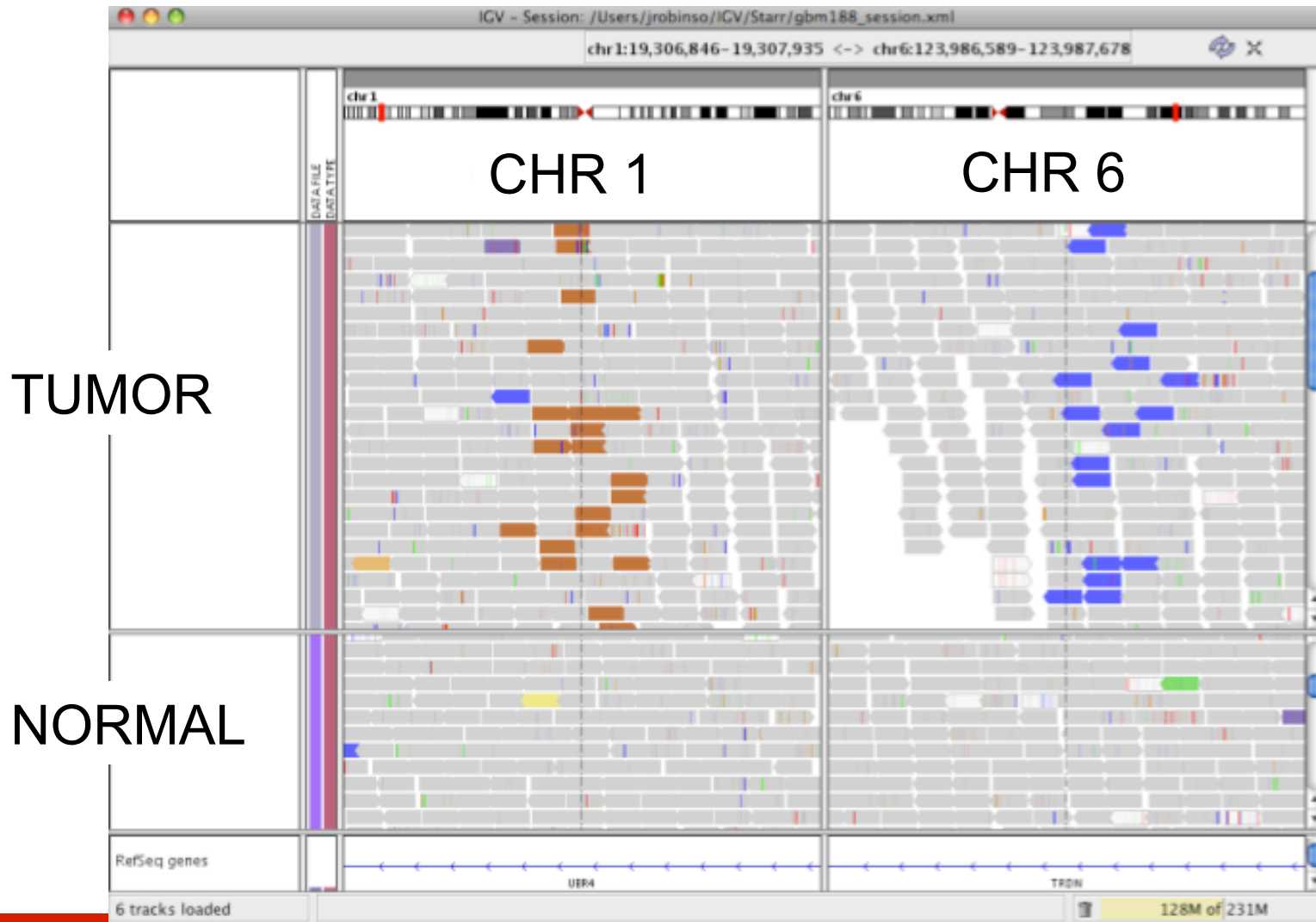
# Insert size color scheme

- Smaller than expected insert size: 
- Larger than expected insert size: 
- Pairs on different chromosomes

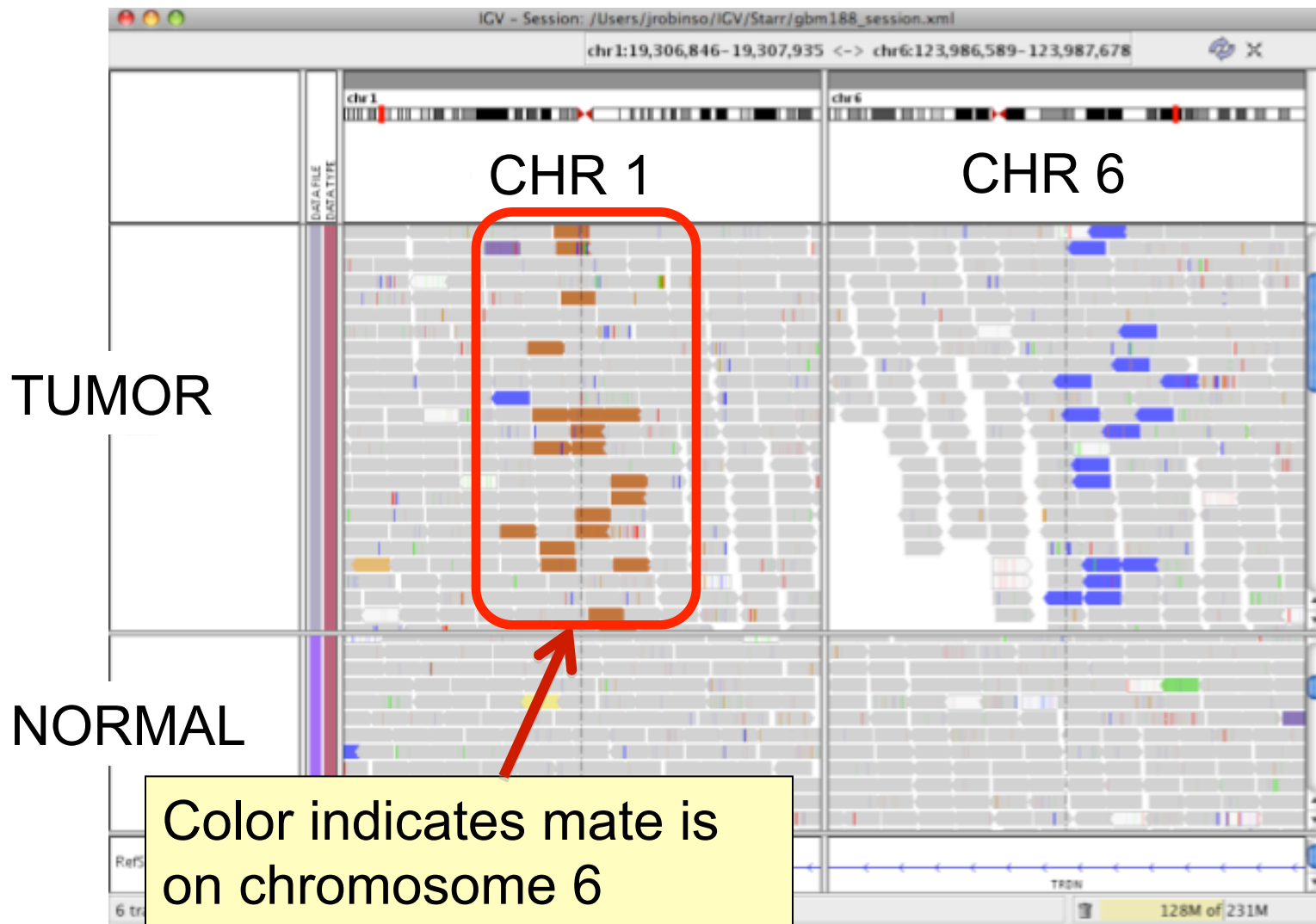
*Each end colored by chromosome of its mate*



# Rearrangement



# Rearrangement



# Interpreting Read-Pair Orientations

Orientation of paired reads can reveal structural events:

- Inversions
- Duplications
- Translocations
- Complex rearrangements

Orientation is defined in terms of

- read strand, left *vs* right, *and*
- read order, first *vs* second



# Inversion

Reference  
genome



# Inversion

Reference  
genome



# Inversion

Reference  
Genome



Subject



# Inversion

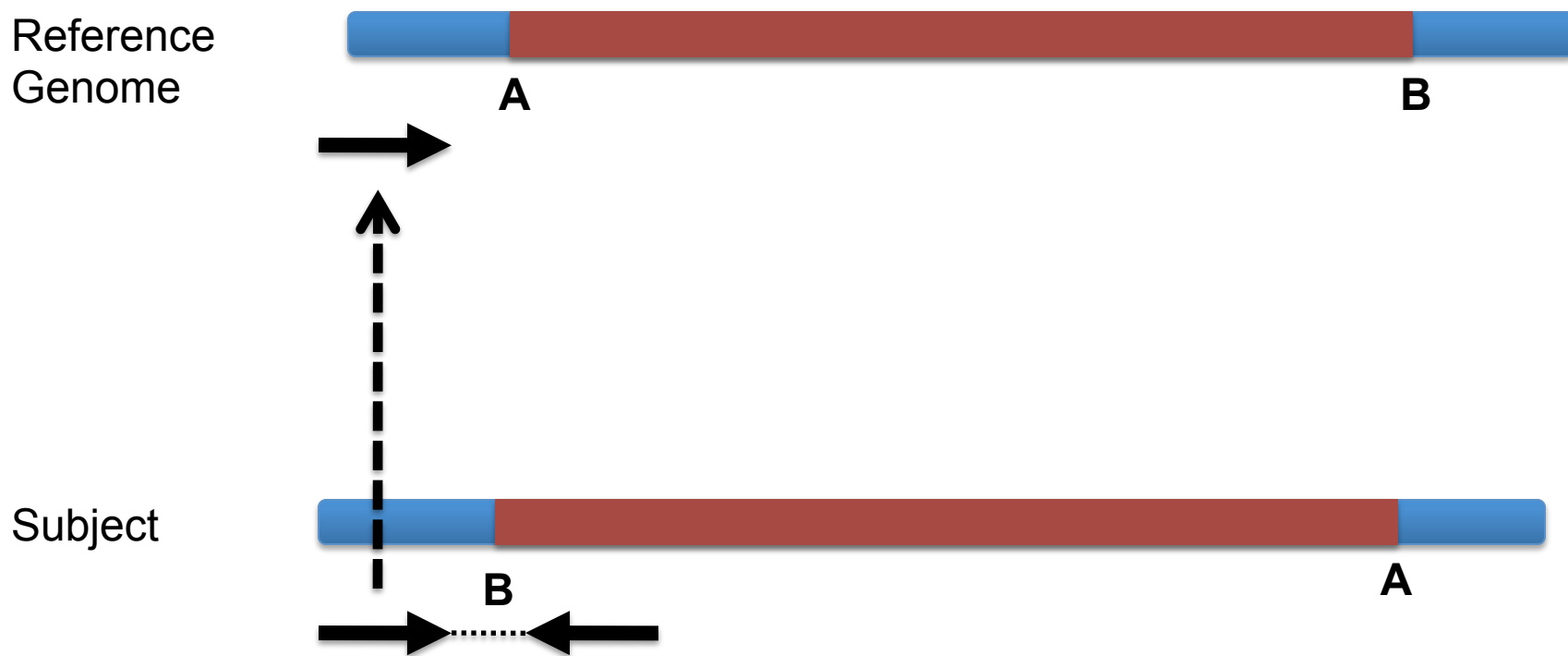
Reference  
Genome



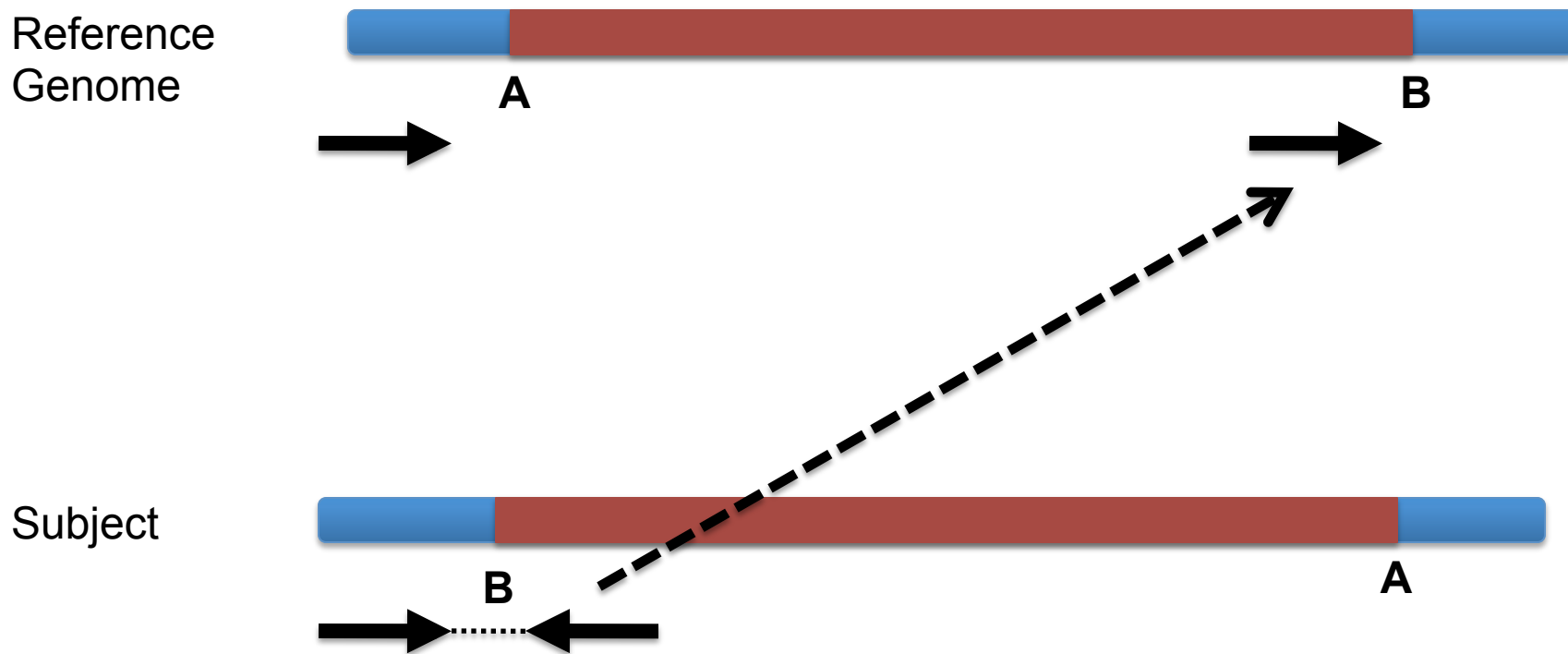
Subject



# Inversion



# Inversion



# Inversion

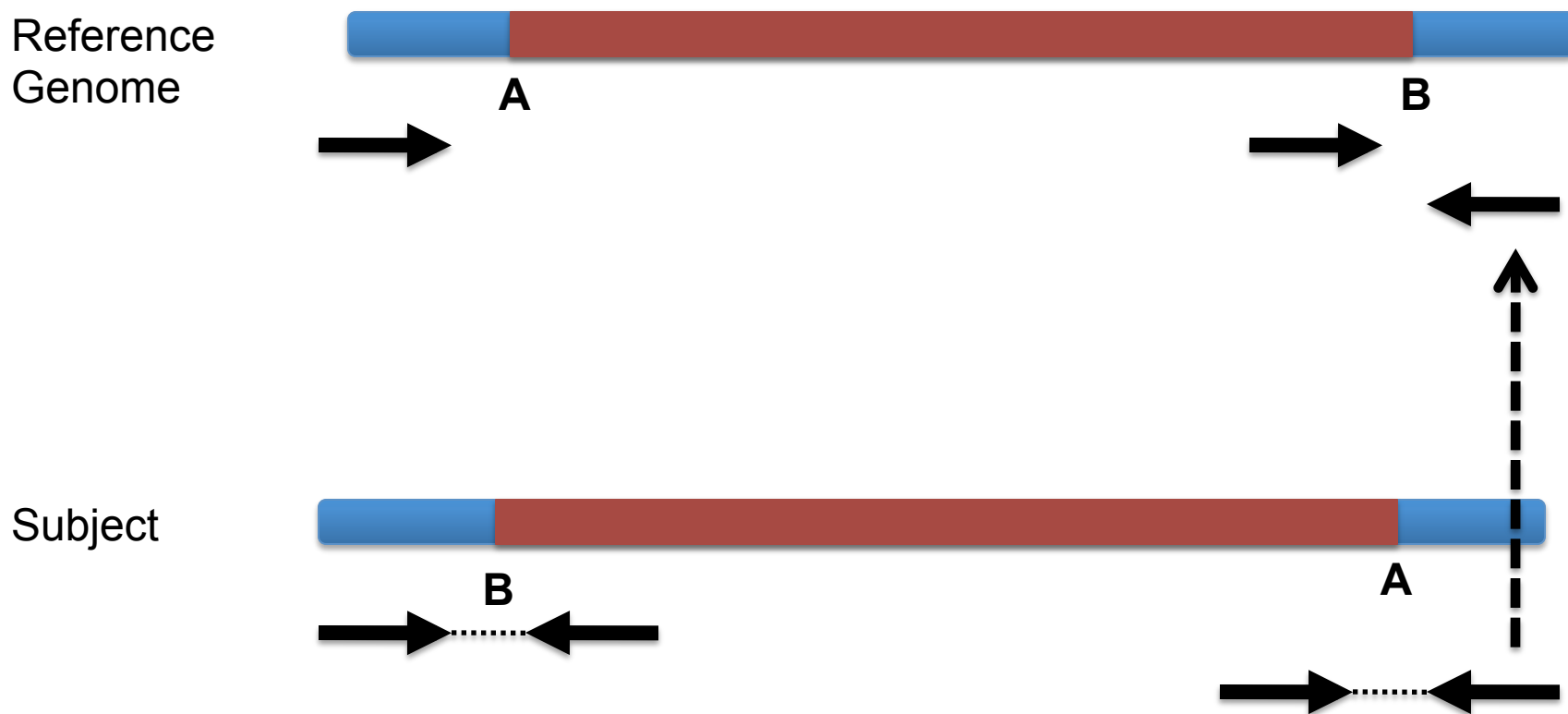
Reference  
Genome



Subject

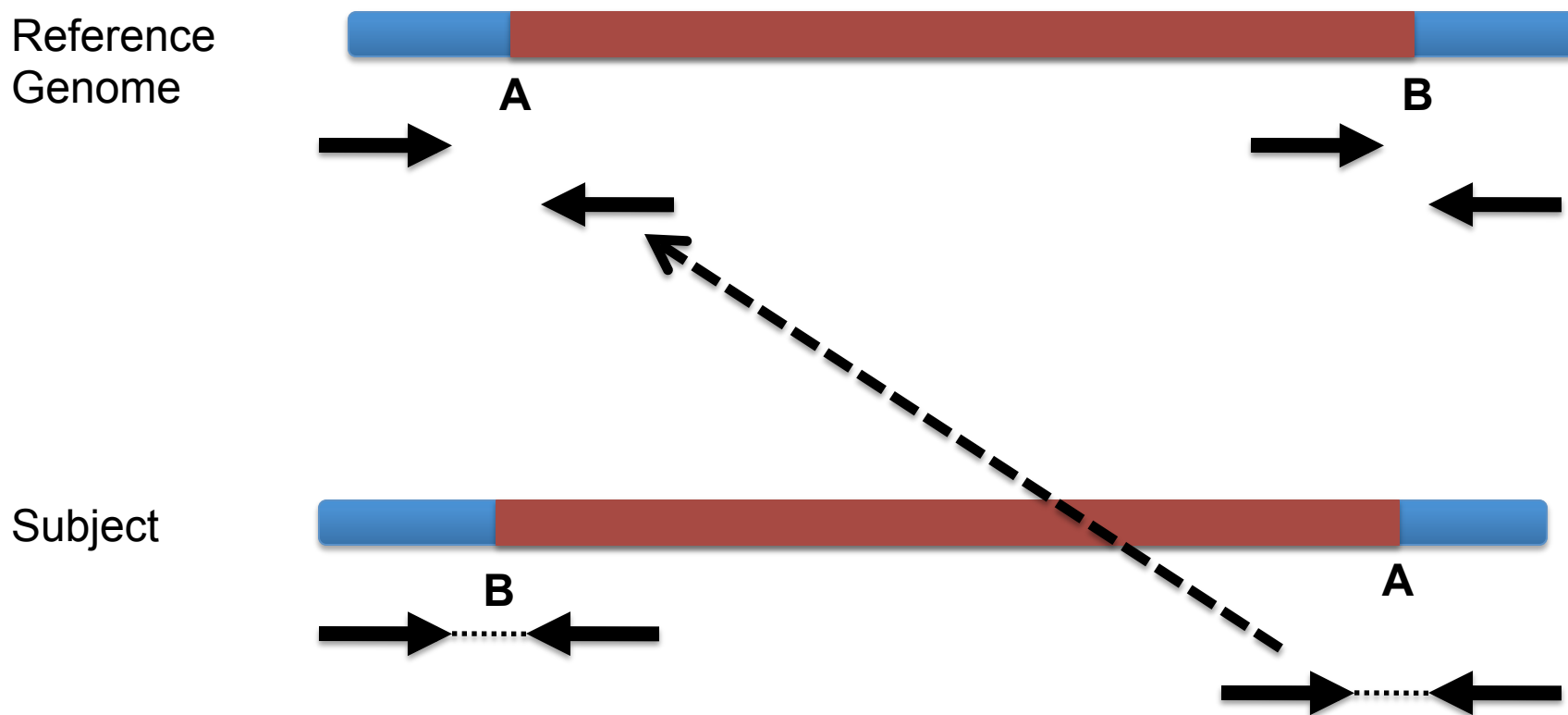


# Inversion





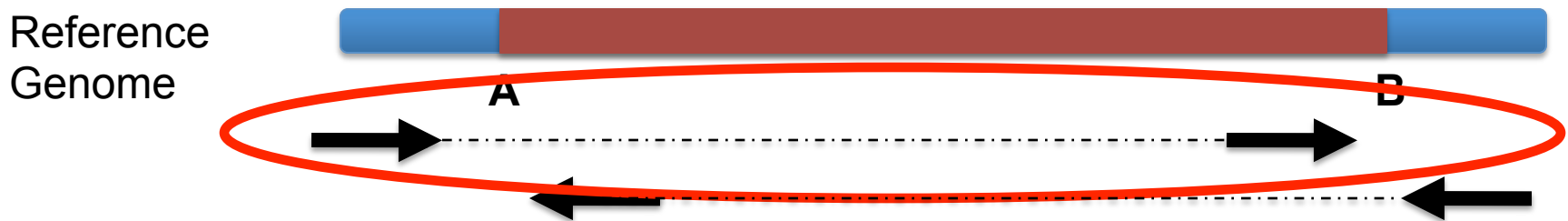
# Inversion



# Inversion



# Inversion



Anomaly: expected orientation of pair is inward facing (  $\longrightarrow$   $\longleftarrow$  )

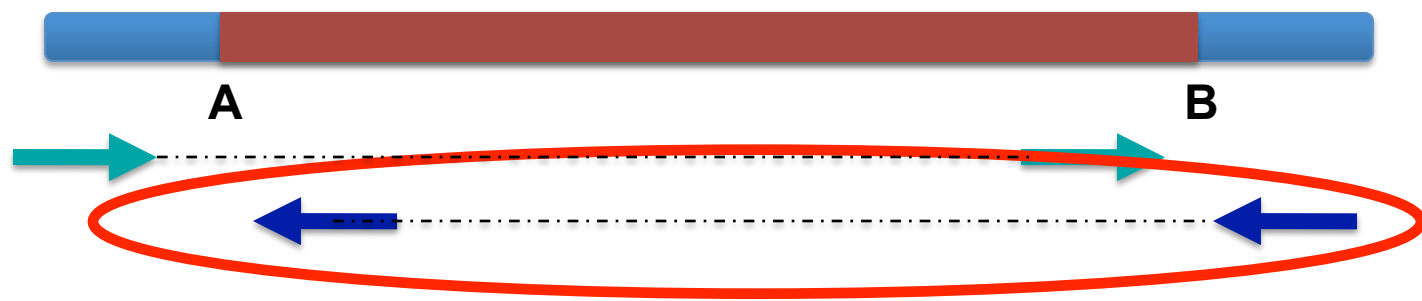
# Inversion



“Left” side pair

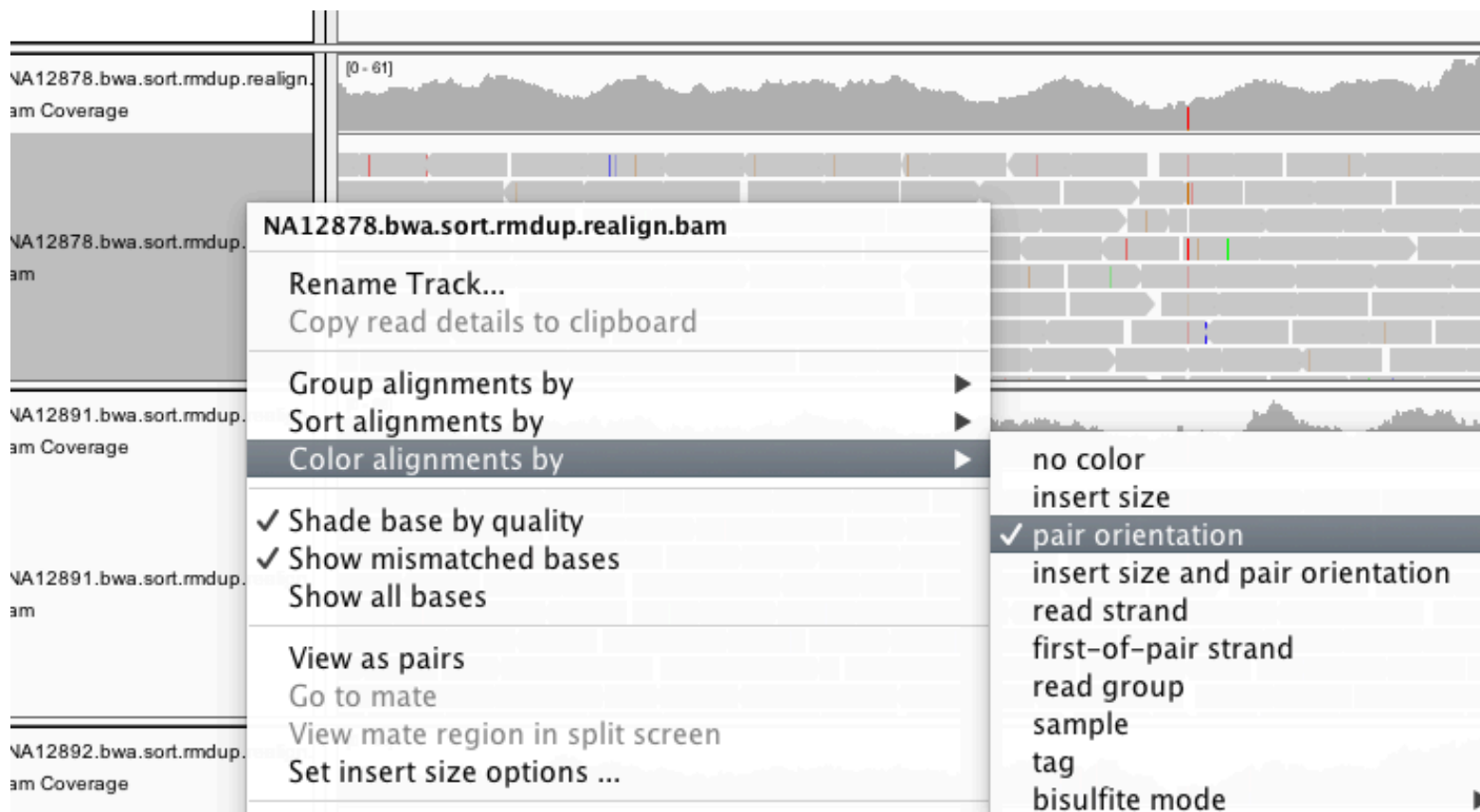
# Inversion

Reference  
Genome



“Right” side pair

# Color by pair orientation

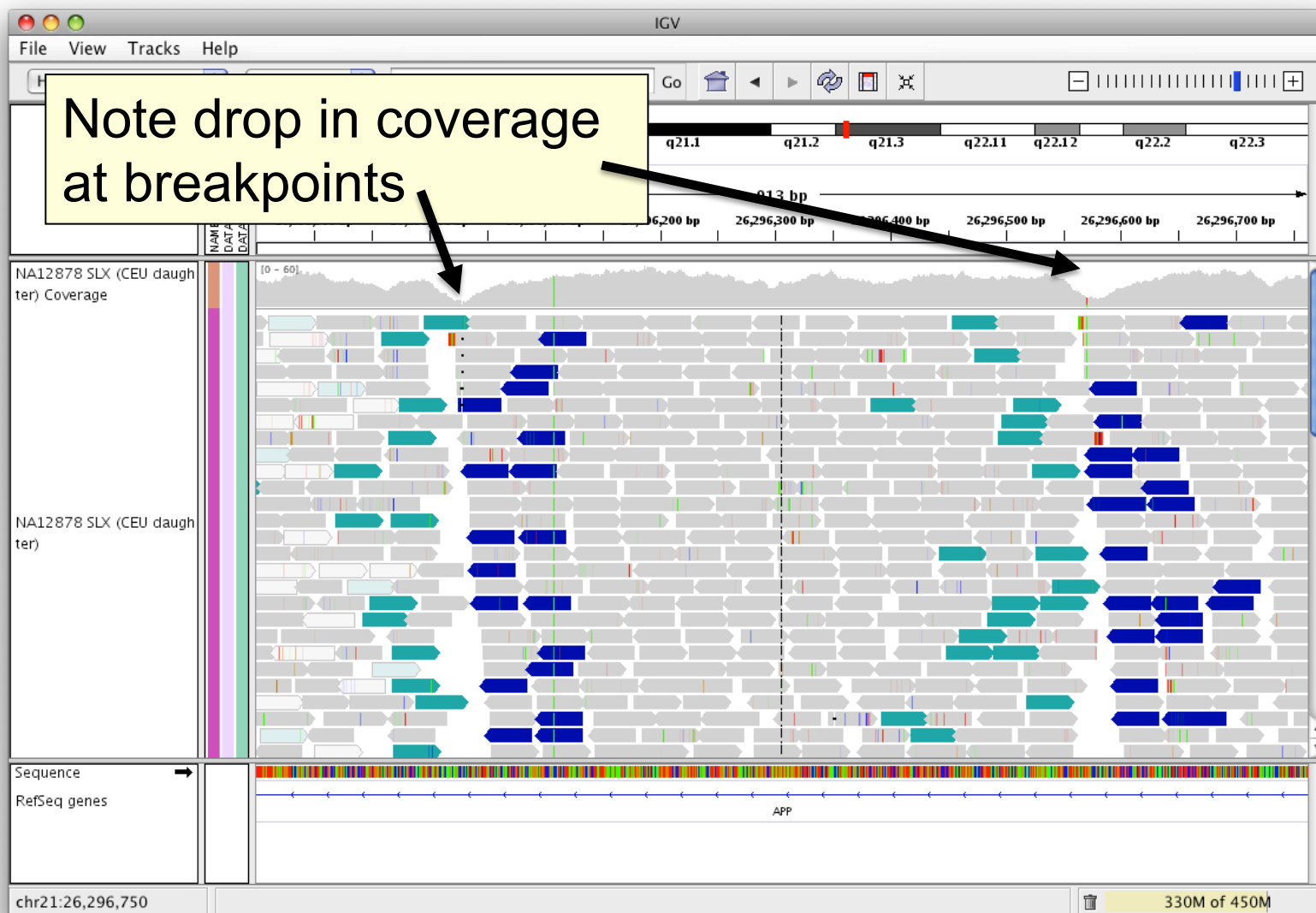


The screenshot shows a genomic browser interface with a track titled "NA12878.bwa.sort.rmdup.realign.bam". A context menu is open over the track, listing various actions. The "Color alignments by" option is selected, and its sub-menu is visible, showing "pair orientation" as the chosen option. The background shows a coverage plot and a read alignment track with various colored markers.

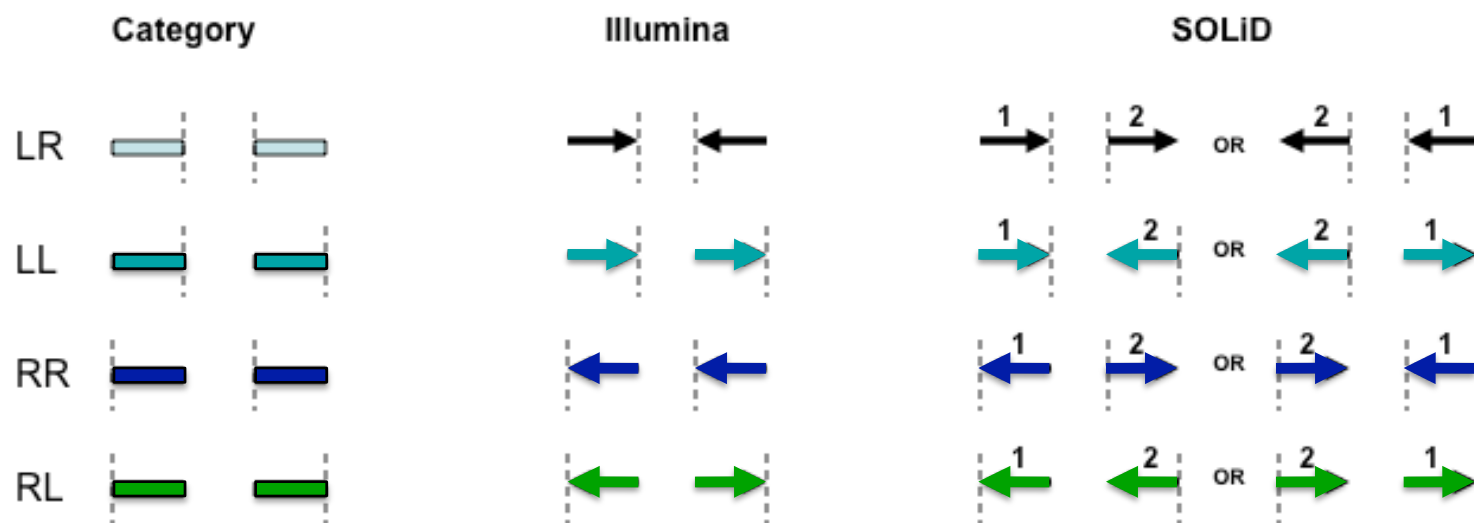
**NA12878.bwa.sort.rmdup.realign.bam**

- Rename Track...
- Copy read details to clipboard
- Group alignments by ▶
- Sort alignments by ▶
- Color alignments by ▶**
  - no color
  - insert size
  - pair orientation**
  - insert size and pair orientation
  - read strand
  - first-of-pair strand
  - read group
  - sample
  - tag
  - bisulfite mode ▶
- ✓ Shade base by quality
- ✓ Show mismatched bases
- Show all bases
- View as pairs
- Go to mate
- View mate region in split screen
- Set insert size options ...

# Inversion



## Interpretation of read pair orientations



- LR Normal reads.  
The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome.
- LL,RR Implies inversion in sequenced DNA with respect to reference.
- RL Implies duplication or translocation with respect to reference.

These categories only apply to reads where both mates map to the same chromosome.

*Figure courtesy of Bob Handsaker*



# IGV hands-on tutorial

[https://github.com/griffithlab/  
rnaseq\\_tutorial/wiki/IGV-Tutorial](https://github.com/griffithlab/rnaseq_tutorial/wiki/IGV-Tutorial)

# Manual Review Standard Operating Procedure (SOP) paper

© American College of Medical Genetics and Genomics

ARTICLE | Genetics  
in Medicine

Open

## Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples

Erica K. Barnell, BS<sup>1</sup>, Peter Ronning, BS<sup>1</sup>, Katie M. Campbell, BS<sup>1</sup>, Kilannin Krysiak, PhD<sup>1,2</sup>, Benjamin J. Ainscough, PhD<sup>1,3</sup>, Lana M. Sheta<sup>1</sup>, Shahil P. Pema<sup>1</sup>, Alina D. Schmidt, BS<sup>1</sup>, Megan Richters, BS<sup>1</sup>, Kelsy C. Cotto, BS<sup>1</sup>, Arpad M. Danos, PhD<sup>1</sup>, Cody Ramirez, BS<sup>1</sup>, Zachary L. Skidmore, MEng<sup>1</sup>, Nicholas C. Spies, BS<sup>1</sup>, Jasreet Hundal, MS<sup>1</sup>, Malik S. Sediqzad<sup>1</sup>, Jason Kunisaki, BS<sup>1</sup>, Felicia Gomez, PhD<sup>1</sup>, Lee Trani, BS<sup>1</sup>, Matthew Matlock, BS<sup>1</sup>, Alex H. Wagner, PhD<sup>1</sup>, S. Joshua Swamidass, MD/PhD<sup>4,5</sup>, Malachi Griffith, PhD<sup>1,2,3,6</sup> and Obi L. Griffith, PhD<sup>1,2,3,6</sup>

**Purpose:** Following automated variant calling, manual review of aligned read sequences is required to identify a high-quality list of somatic variants. Despite widespread use in analyzing sequence data, methods to standardize manual review have not been described, resulting in high inter- and intralab variability.

**Methods:** This manual review standard operating procedure (SOP) consists of methods to annotate variants with four different calls and 19 tags. The calls indicate a reviewer's confidence in each variant and the tags indicate commonly observed sequencing patterns and artifacts that inform the manual review call. Four individuals were asked to classify variants prior to, and after, reading the SOP and accuracy was assessed by comparing reviewer calls with orthogonal validation sequencing.

**Results:** After reading the SOP, average accuracy in somatic variant identification increased by 16.7% ( $p$  value = 0.0298) and average interreviewer agreement increased by 12.7% ( $p$  value < 0.001). Manual review conducted after reading the SOP did not significantly increase reviewer time.

**Conclusion:** This SOP supports and enhances manual somatic variant detection by improving reviewer accuracy while reducing the interreviewer variability for variant calling and annotation.

*Genetics in Medicine* (2018) <https://doi.org/10.1038/s41436-018-0278-z>

**Keywords:** somatic variant refinement; manual review

# DeepSVR Paper

nature  
genetics

TECHNICAL REPORT

<https://doi.org/10.1038/s41588-018-0257-y>

## A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data

Benjamin J. Ainscough <sup>1,2,12</sup>, Erica K. Barnell <sup>1,12</sup>, Peter Ronning<sup>1</sup>, Katie M. Campbell <sup>1</sup>, Alex H. Wagner <sup>1</sup>, Todd A. Fehniger <sup>2,3</sup>, Gavin P. Dunn<sup>4</sup>, Ravindra Uppaluri<sup>5</sup>, Ramaswamy Govindan<sup>2,3</sup>, Thomas E. Rohan<sup>6</sup>, Malachi Griffith <sup>1,2,3,7</sup>, Elaine R. Mardis<sup>8,9</sup>, S. Joshua Swamidass<sup>10,11\*</sup> and Obi L. Griffith <sup>1,2,3,7\*</sup>

**Cancer genomic analysis requires accurate identification of somatic variants in sequencing data. Manual review to refine somatic variant calls is required as a final step after automated processing. However, manual variant refinement is time-consuming, costly, poorly standardized, and non-reproducible. Here, we systematized and standardized somatic variant refinement using a machine learning approach. The final model incorporates 41,000 variants from 440 sequencing cases. This model accurately recapitulated manual refinement labels for three independent testing sets (13,579 variants) and accurately predicted somatic variants confirmed by orthogonal validation sequencing data (212,158 variants). The model improves on manual somatic refinement by reducing bias on calls otherwise subject to high inter-reviewer variability.**

Break