

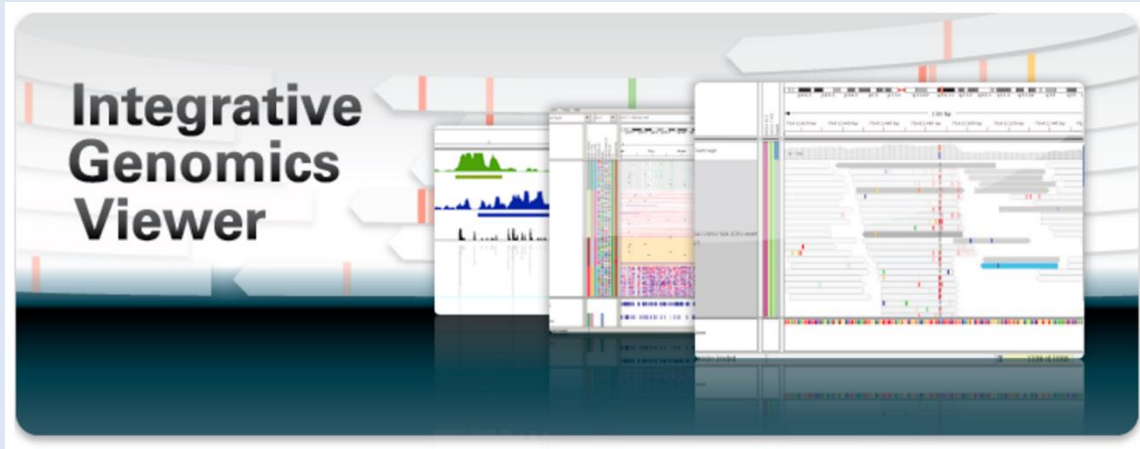


Cold  
Spring  
Harbor  
Laboratory

Module 2:  
Introduction to IGV  
**The Integrative Genomics Viewer**

Arpad Danos, Felicia Gomez, Obi Griffith, Malachi Griffith,  
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal

Advanced Sequencing Technologies & Bioinformatics Analysis November 10-23, 2024



 **Washington University in St. Louis**  
SCHOOL OF MEDICINE

# 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)

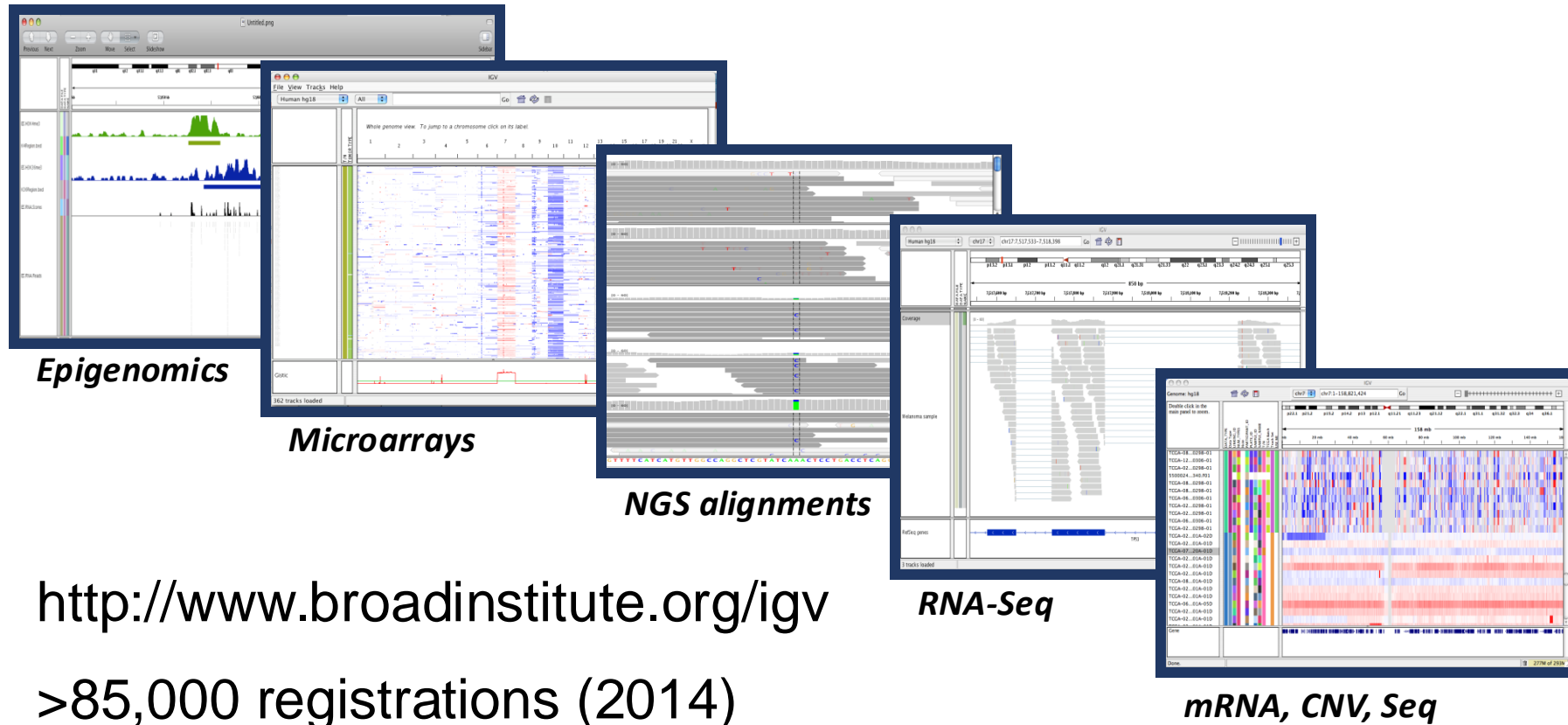


Savant  
Genome  
Browser

- 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

# Integrative Genomics Viewer (IGV)

*Desktop application for the interactive visual exploration of integrated genomic datasets*

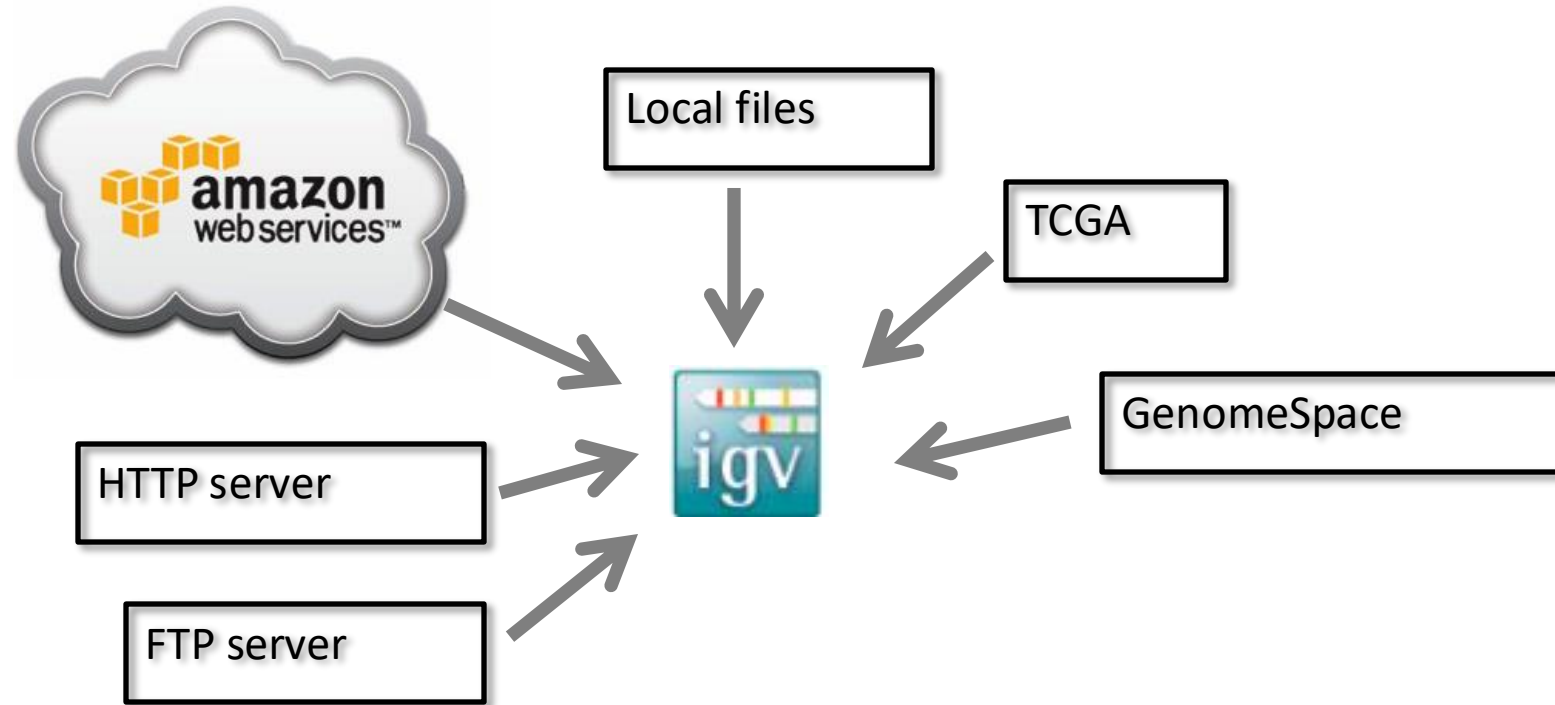


# Features

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

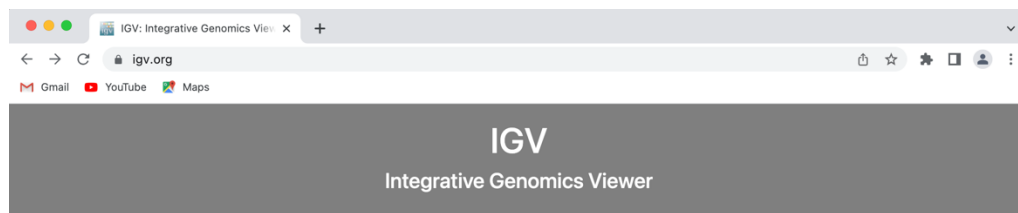
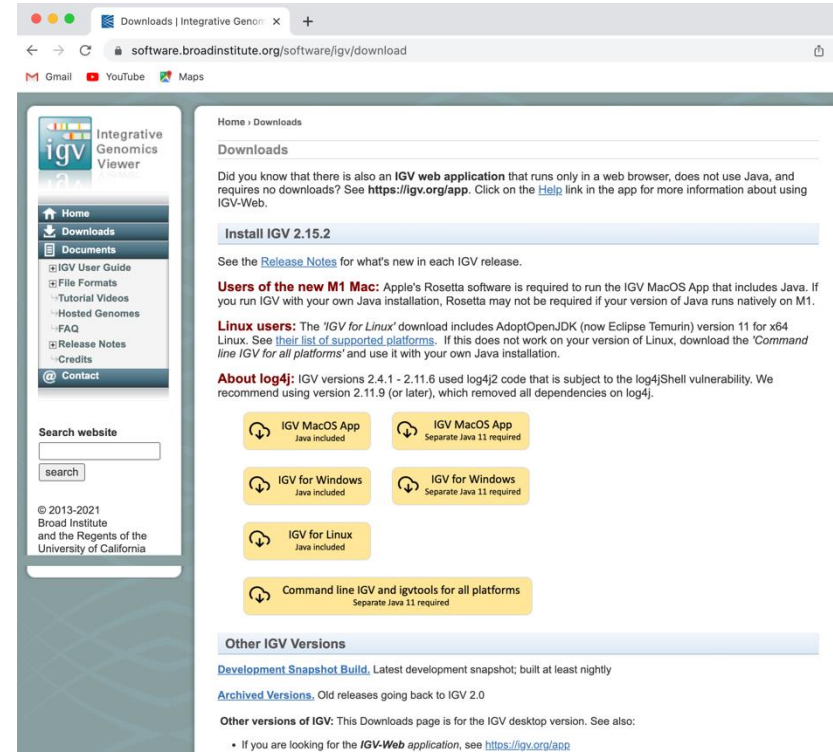
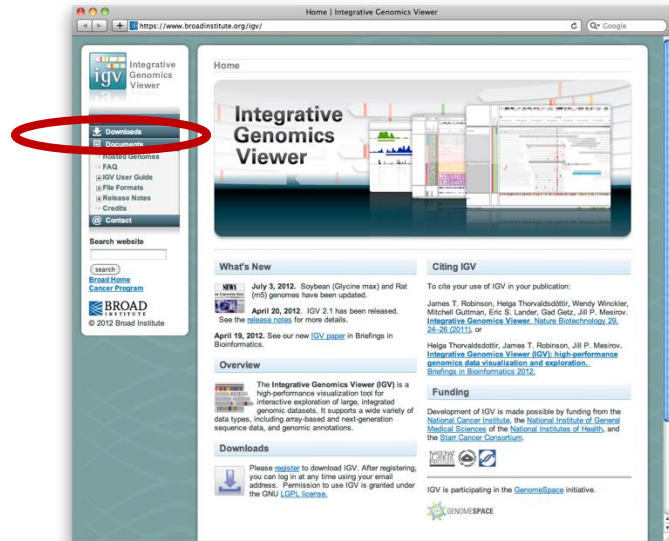


- View **local** files without uploading.
- View **remote** files without downloading the whole dataset.

# Using IGV: the basics

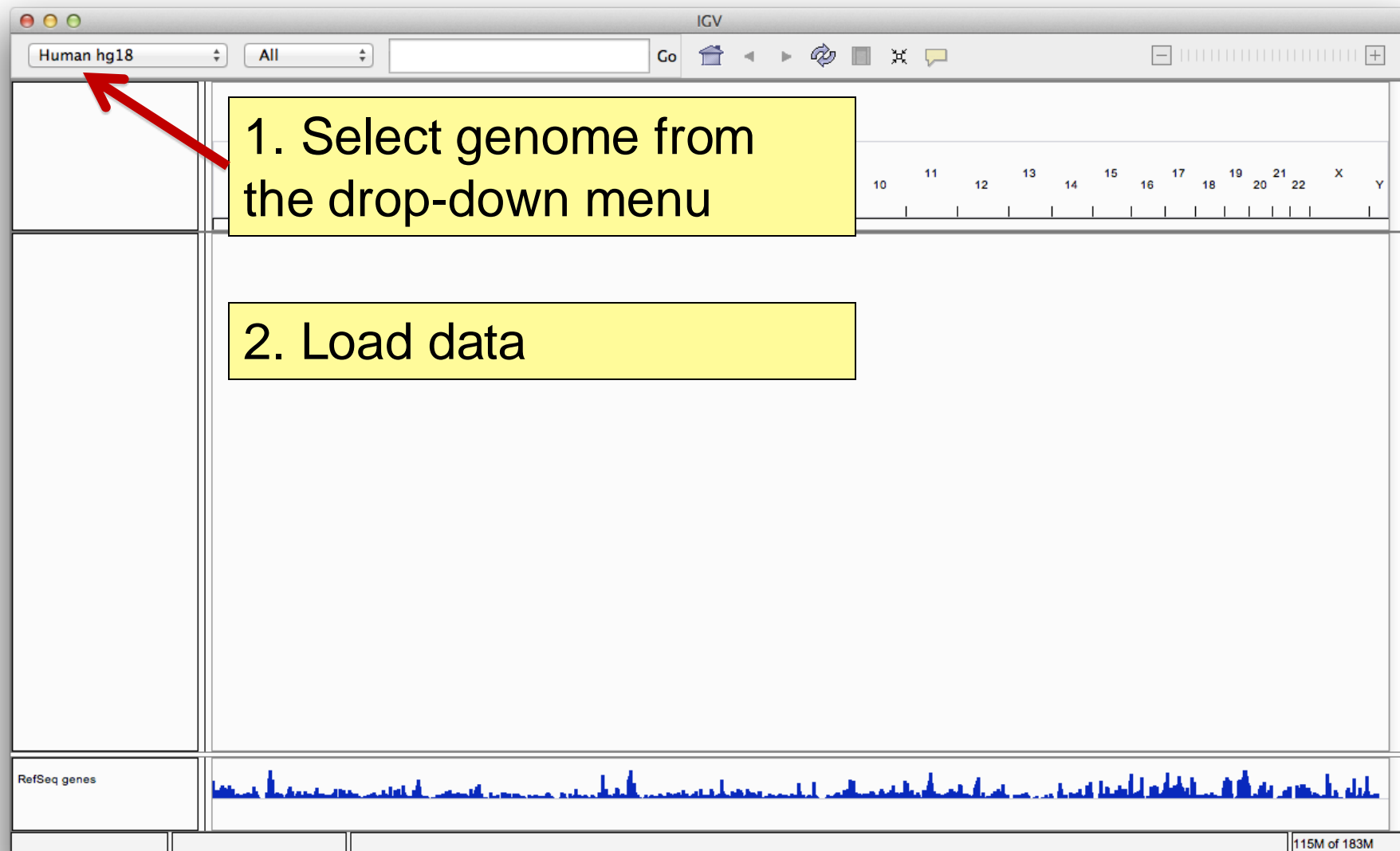
- Launch IGV
- Select a reference genome
- Load data
- Navigate through the data
  - WGS data
    - SNVs
    - structural variations

# Launch IGV



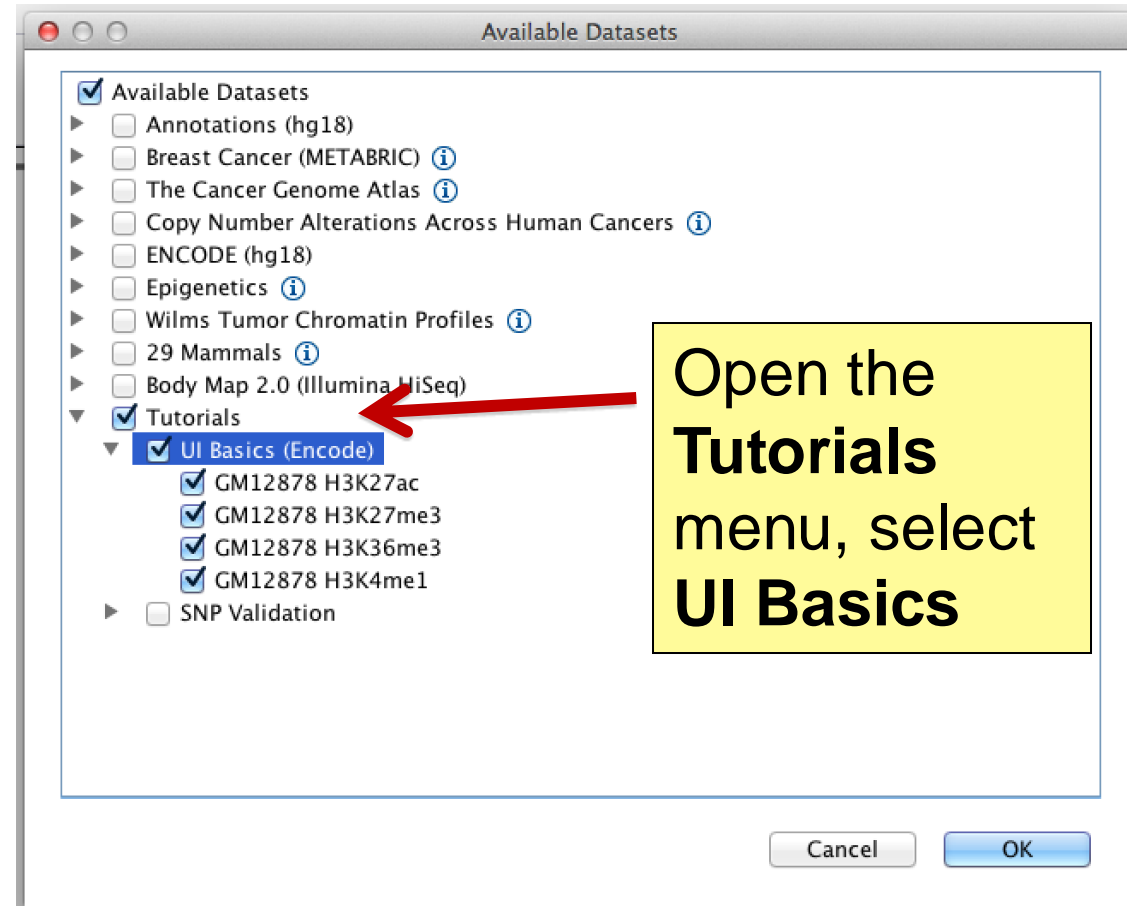
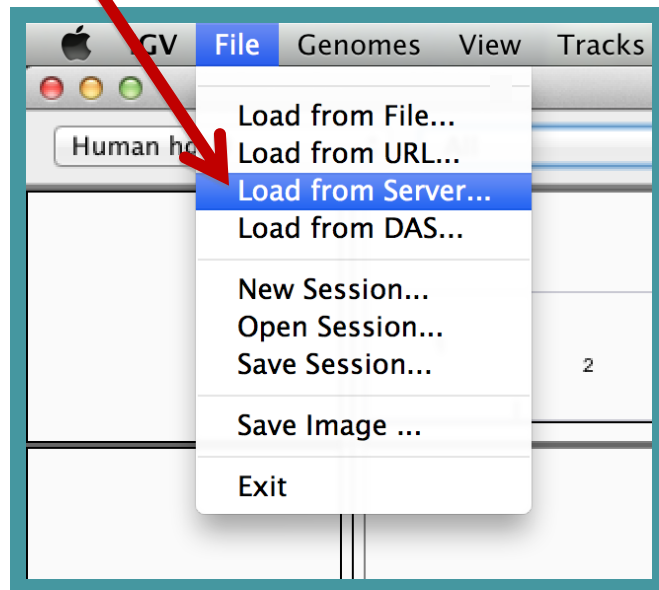


# Select reference genome



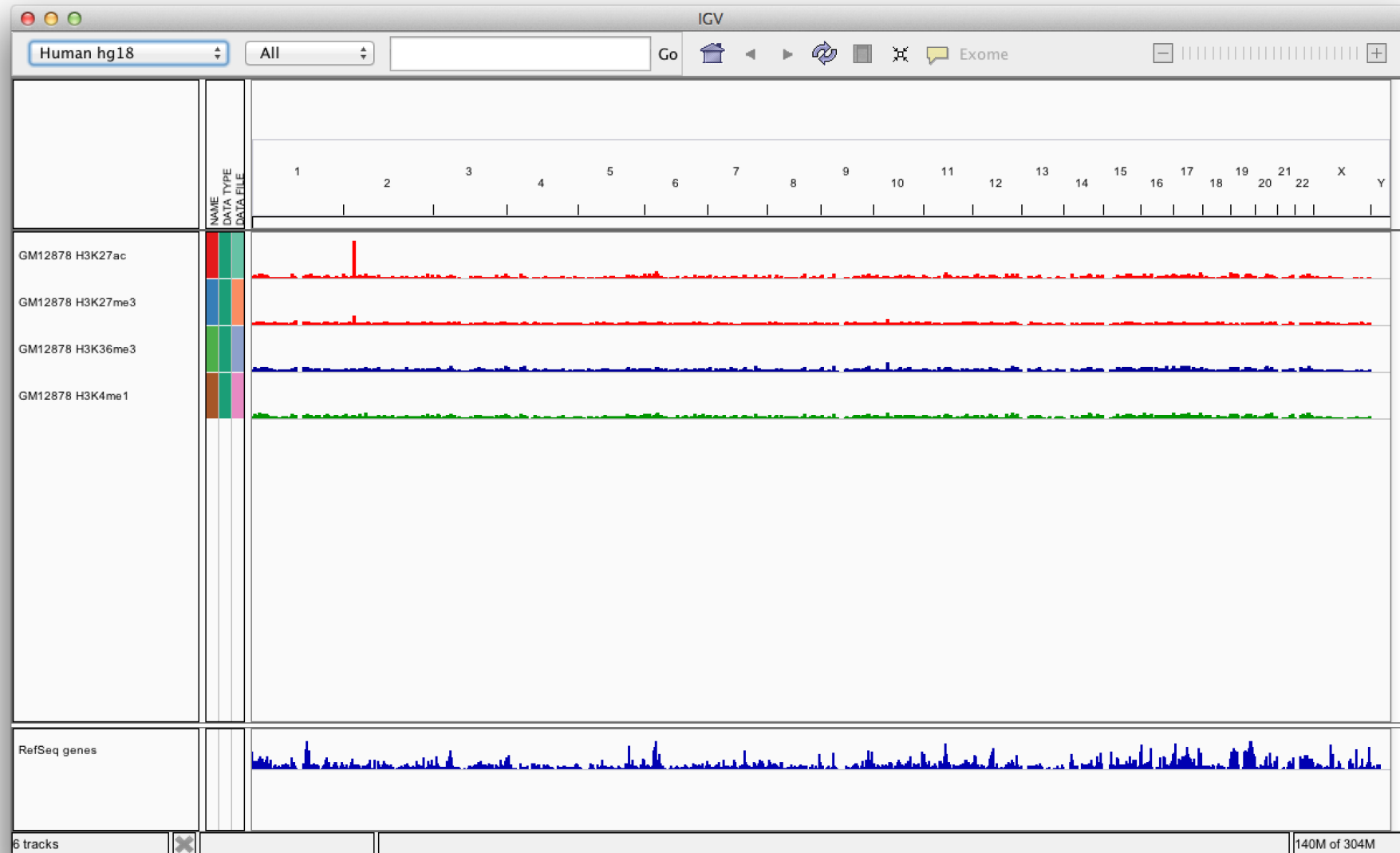
# Load data

Select **File > Load from Server...**

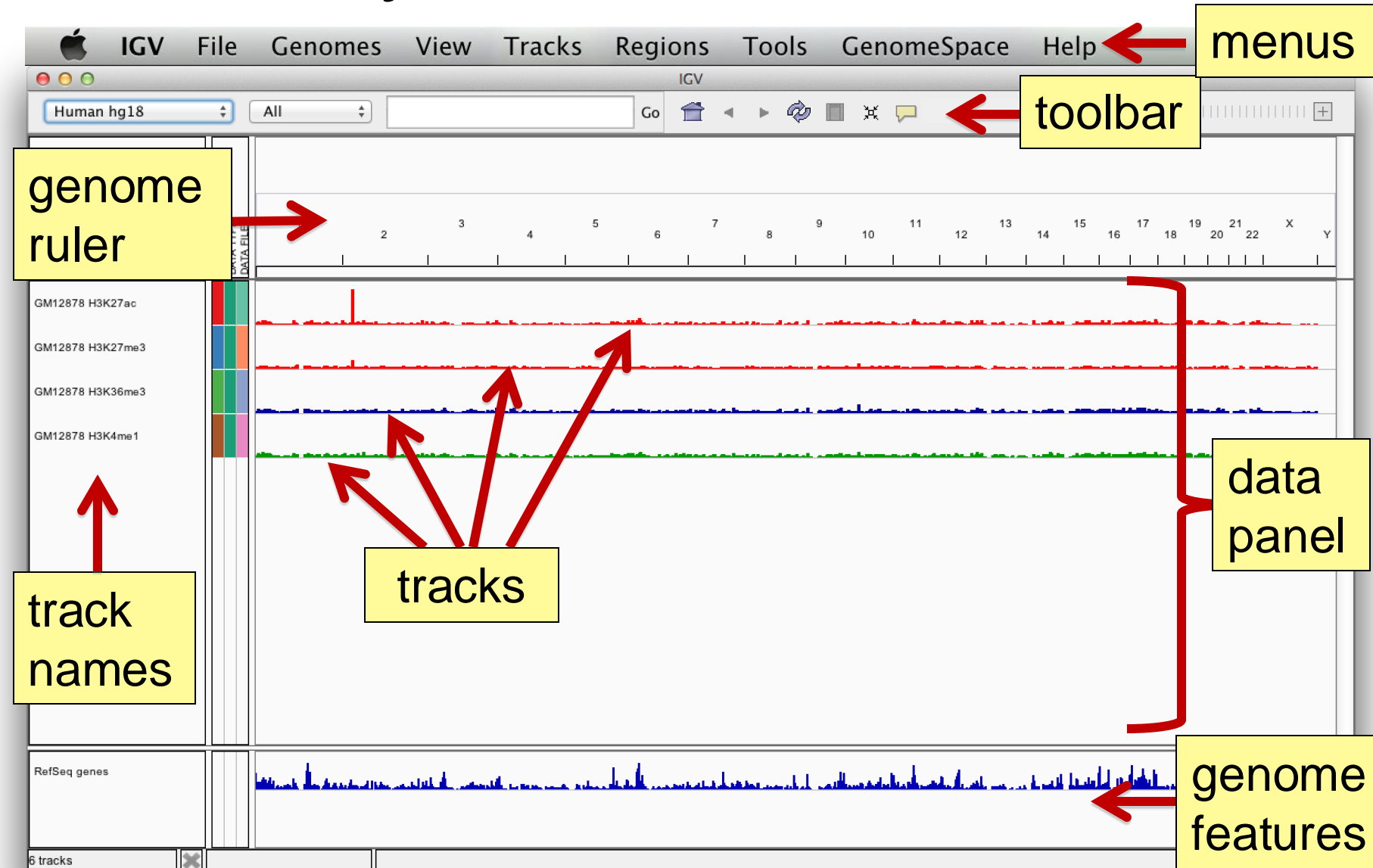


Open the **Tutorials** menu, select **UI Basics**

# 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](#)
- [BEDPE](#)
- [BedGraph](#)
- [bigBed](#)
- [bigWig](#)
- [Birdsuite Files](#)
- [broadPeak](#)
- [CBS](#)
- [Chemical Reactivity Probing Profiles](#)
- [chrom.sizes](#)
- [CN](#)
- [Custom File Formats](#)
- [Cytoband](#)
- [FASTA](#)
- [GCT](#)
- [CRAM](#)
- [genePred](#)
- [GFF/GTF](#)
- [GISTIC](#)
- [Goby](#)
- [GWAS](#)
- [IGV](#)
- [LOH](#)
- [MAF \(Multiple Alignment Format\)](#)
- [MAF \(Mutation Annotation Format\)](#)
- [Merged BAM File](#)
- [narrowPeak](#)
- [PSL](#)
- [RES](#)
- [RNA Secondary Structure Formats](#)
- [SAM](#)
- [Sample Info \(Attributes\) file](#)
- [SEG](#)
- [TDF](#)
- [Track Line](#)
- [Type Line](#)
- [VCF](#)
- [WIG](#)

- For

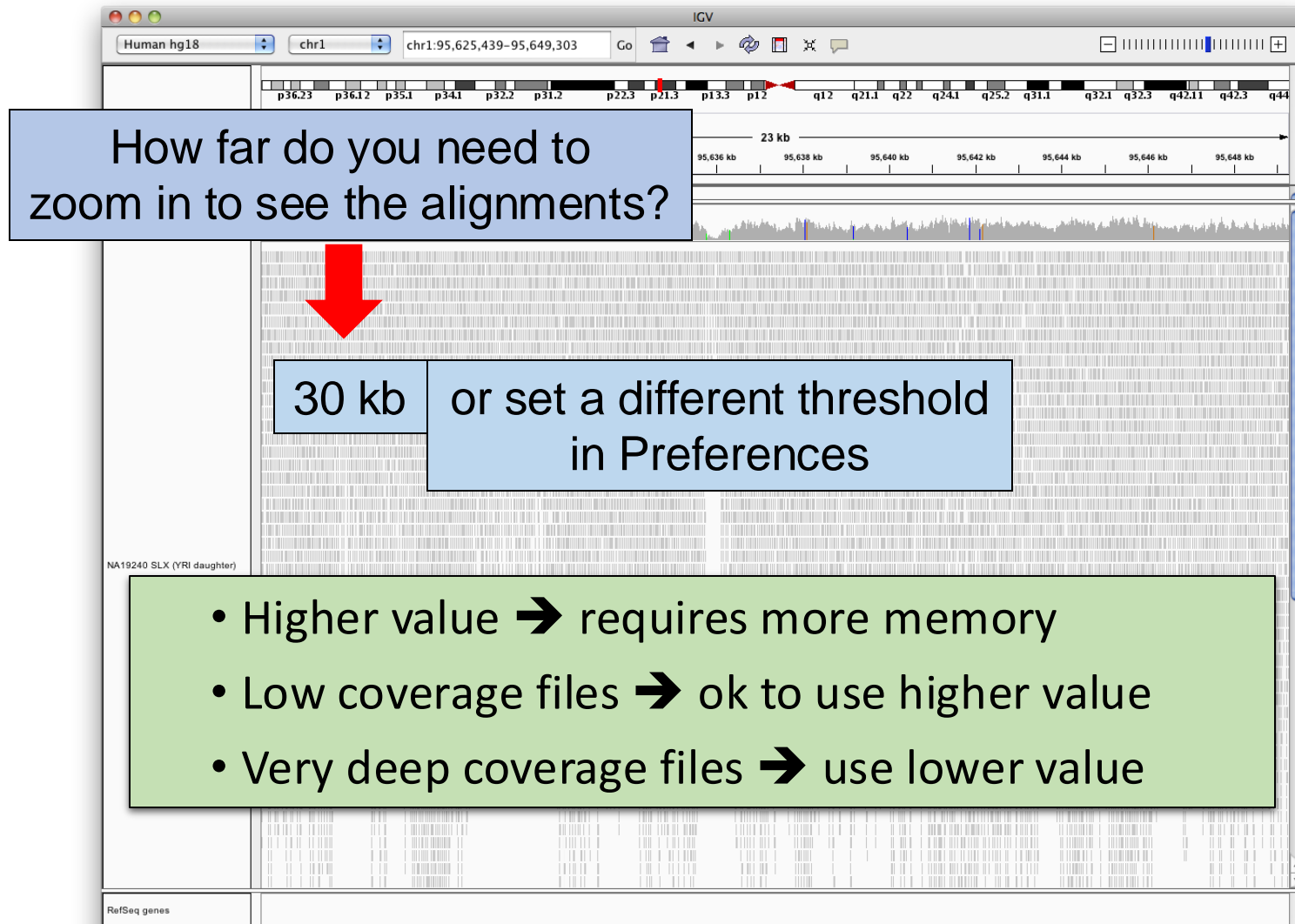
[ftware.broadinstitute.org/software/igv/FileFormats](https://software.broadinstitute.org/software/igv/FileFormats)

# Viewing alignments

## Whole chromosome view



# Viewing alignments – Zoom in



# Viewing alignments – Zoom in

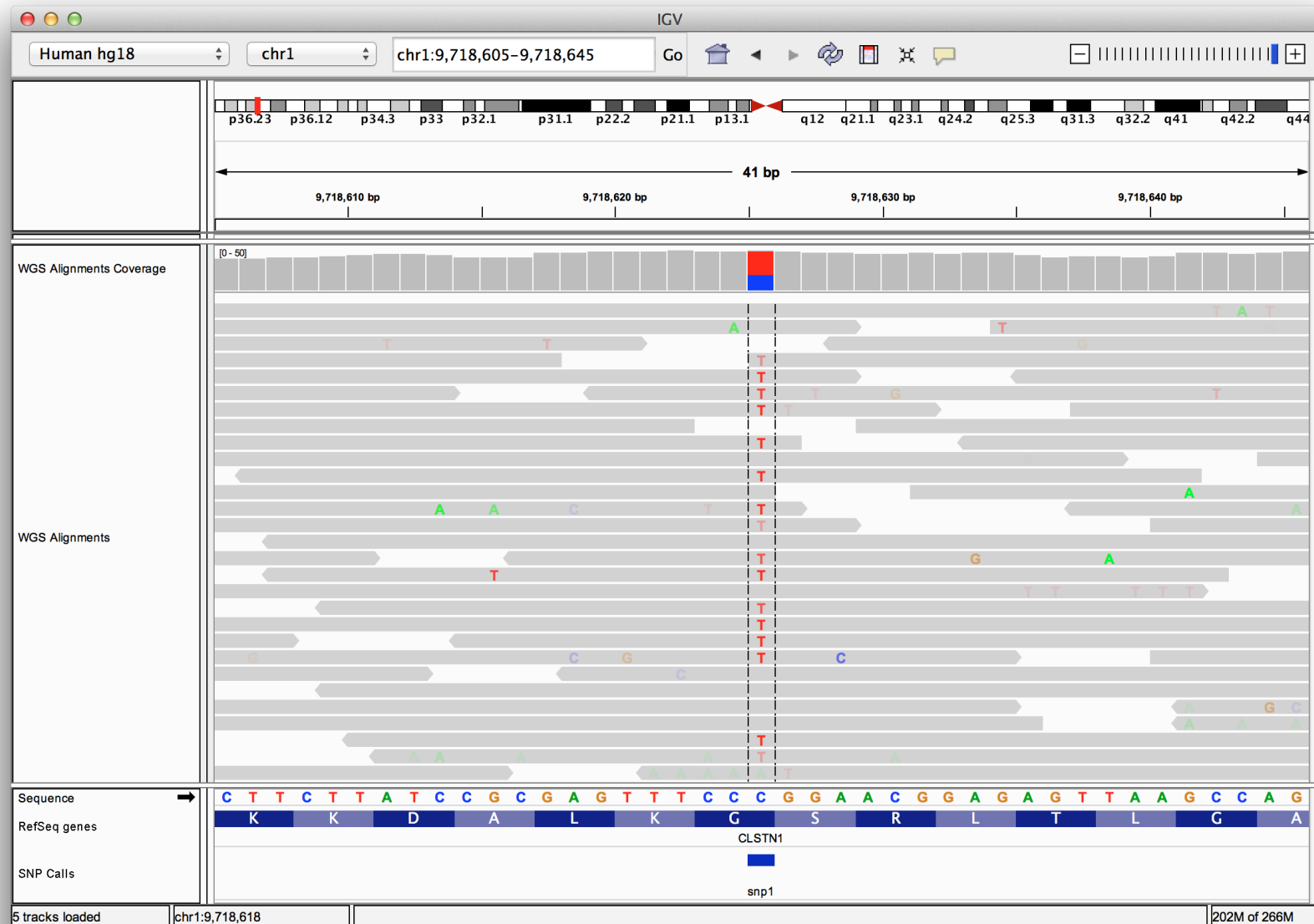




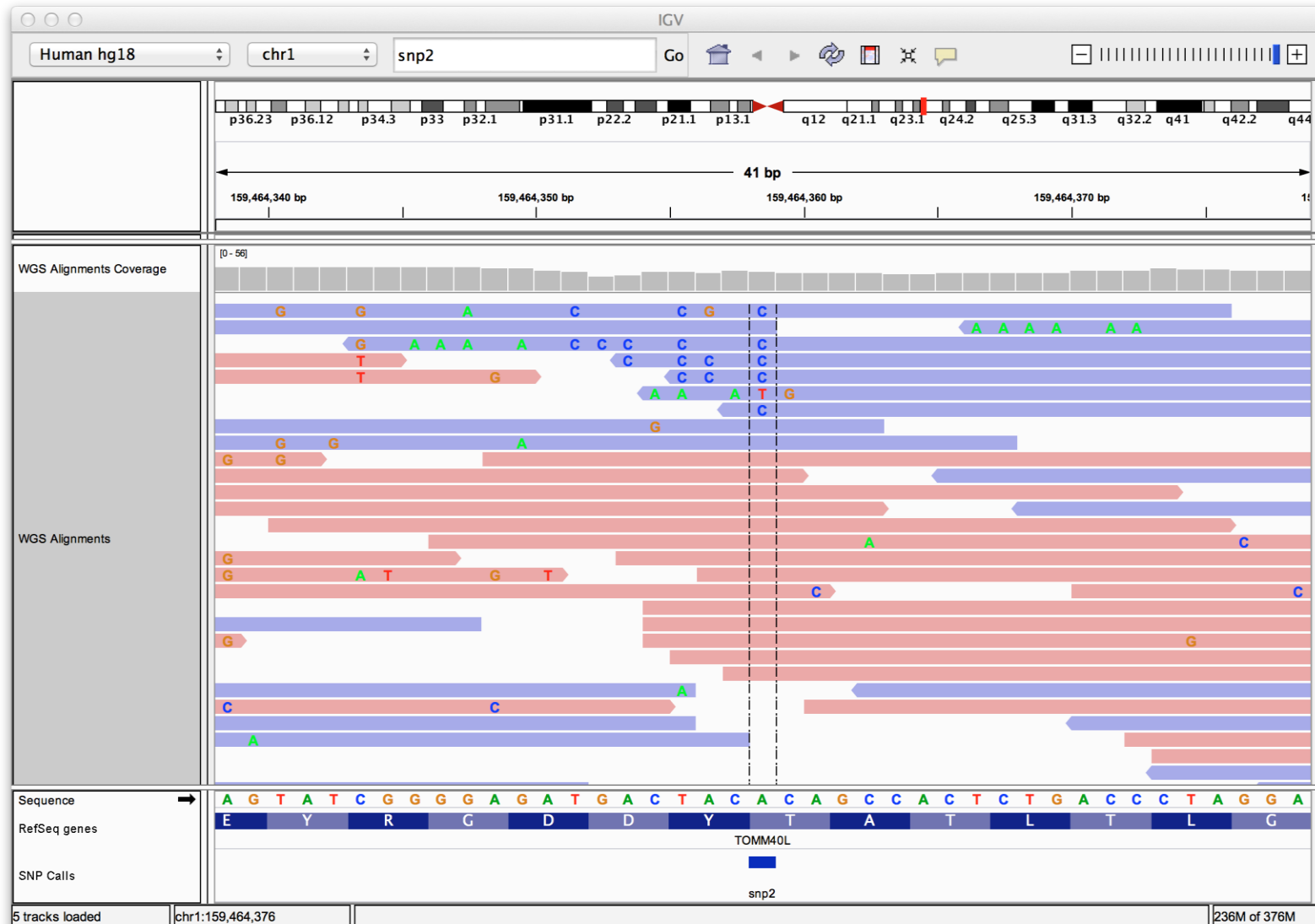
# SNVs and Structural variations

- 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

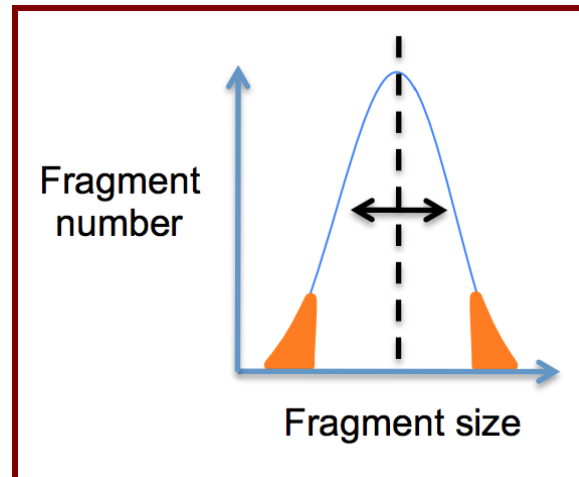
- Paired reads can yield evidence for genomic “structural events”, such as deletions, translocations, and inversions.
- Alignment coloring options help highlight these events based on:
  - Inferred insert size (template length)
  - Pair orientation (relative strand of pair)

# Paired-end sequencing

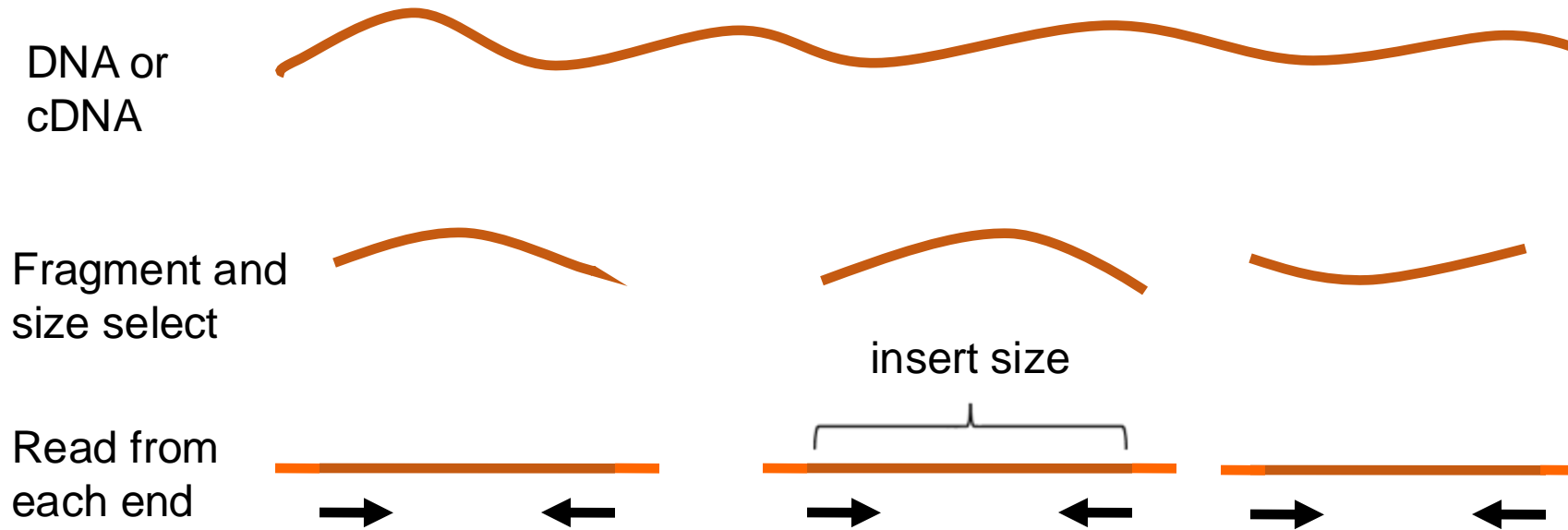
DNA or  
cDNA



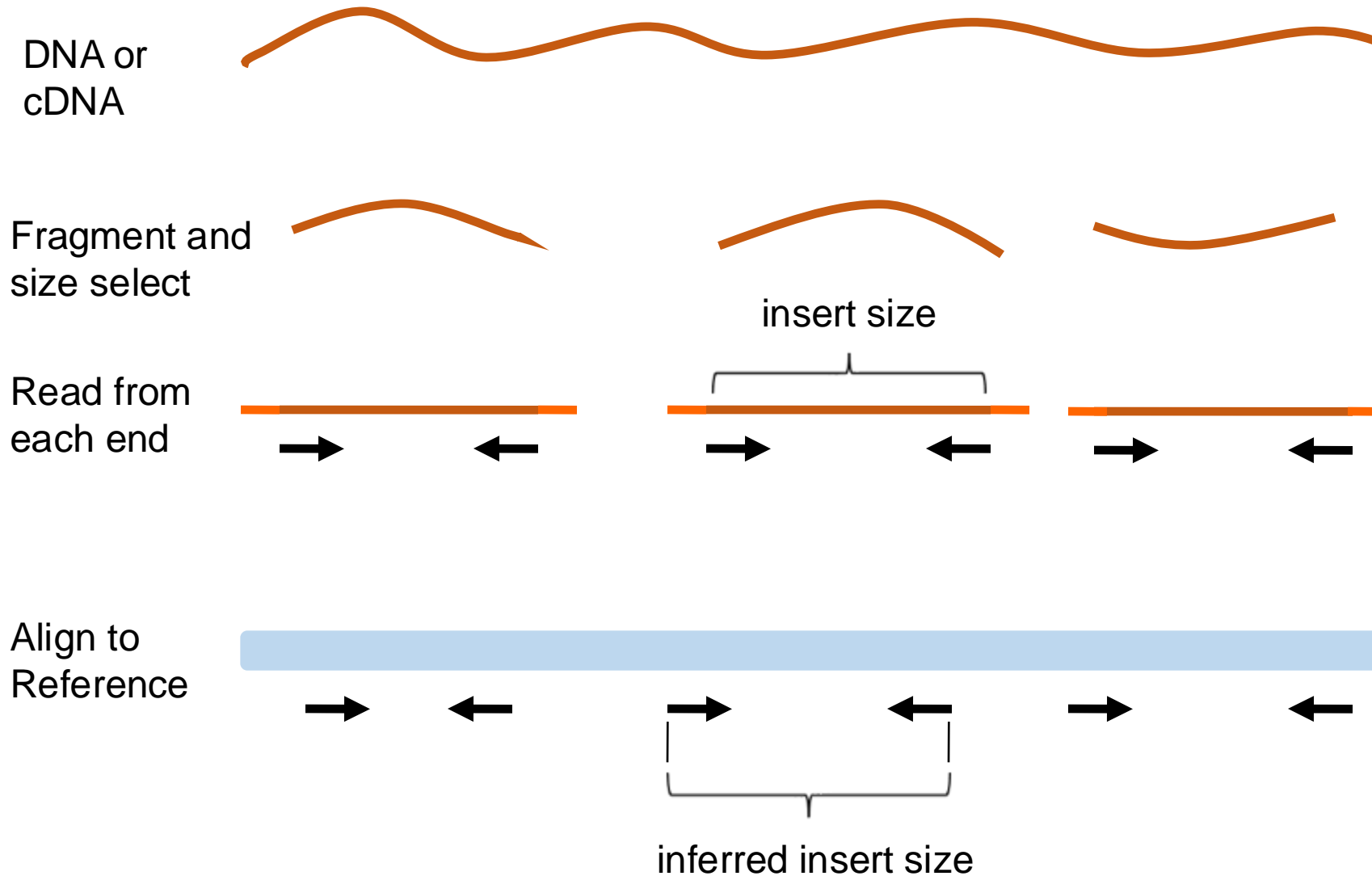
Fragment and  
size select



# 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)

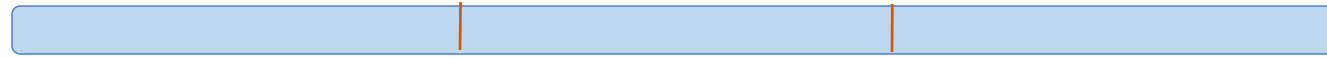


# Deletion

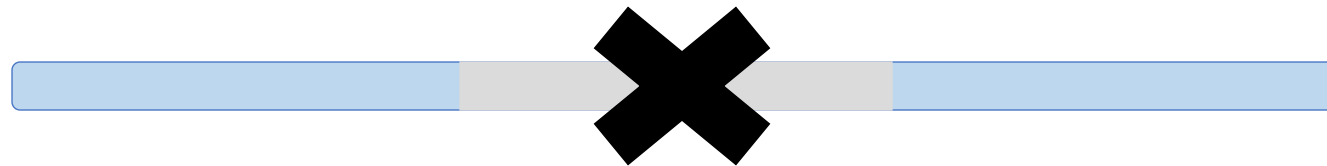
What is the effect of a deletion on inferred insert size?

# Deletion

Reference  
Genome

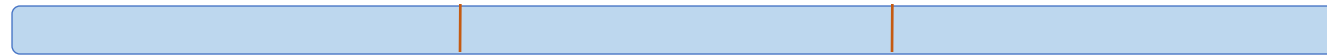


Subject



# Deletion

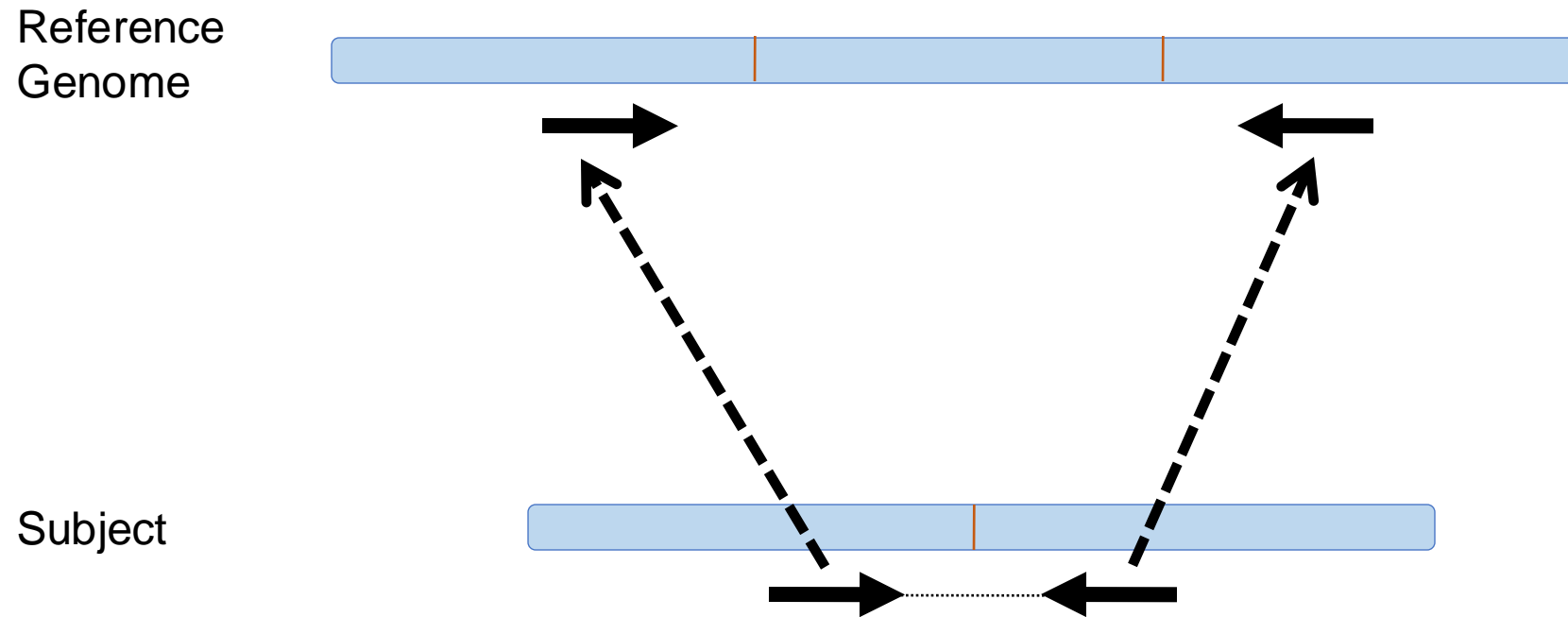
Reference  
Genome



Subject



# Deletion



# Deletion

Inferred insert size is  $>$  expected value

Reference  
Genome



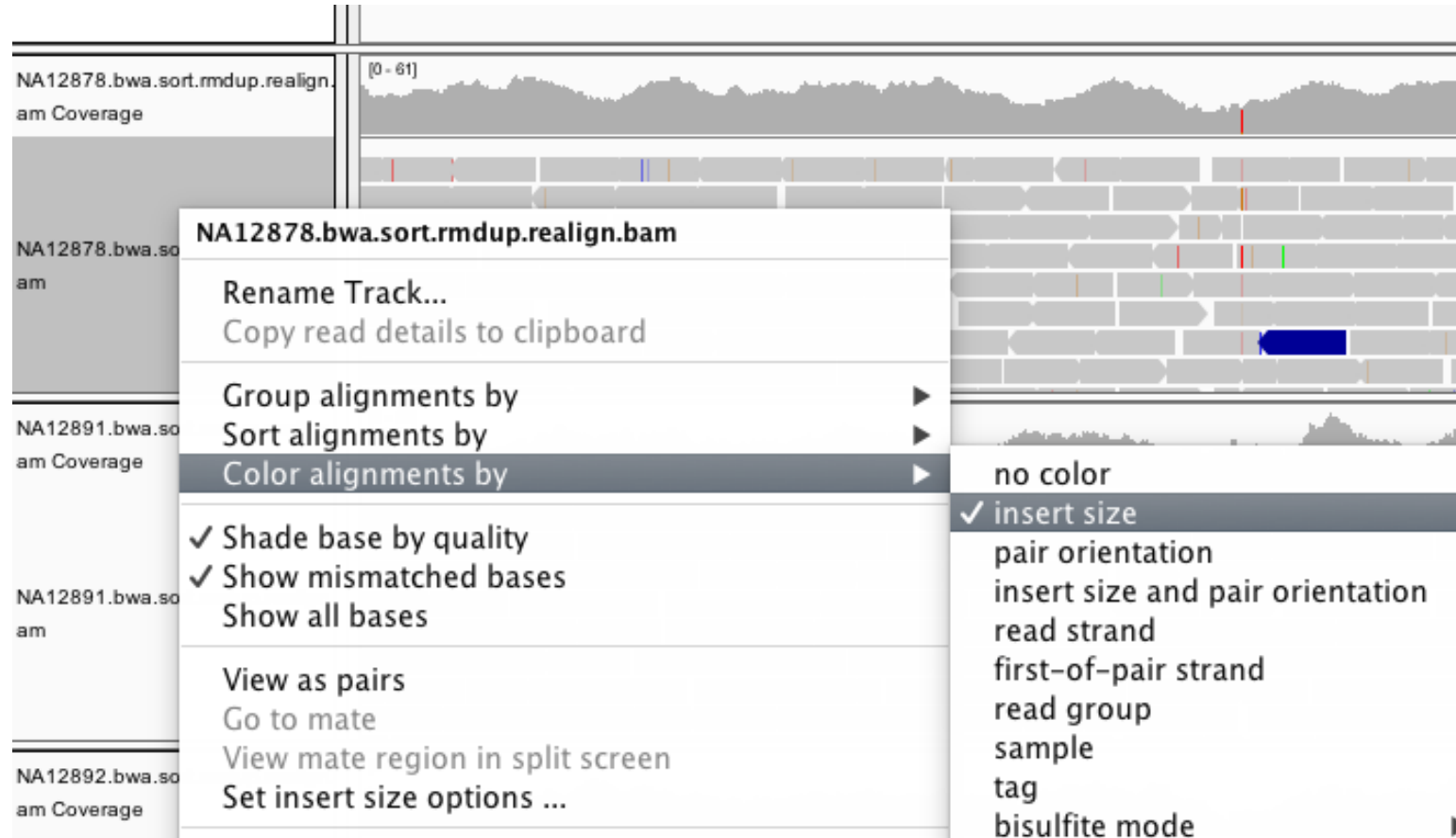
inferred insert size

Subject

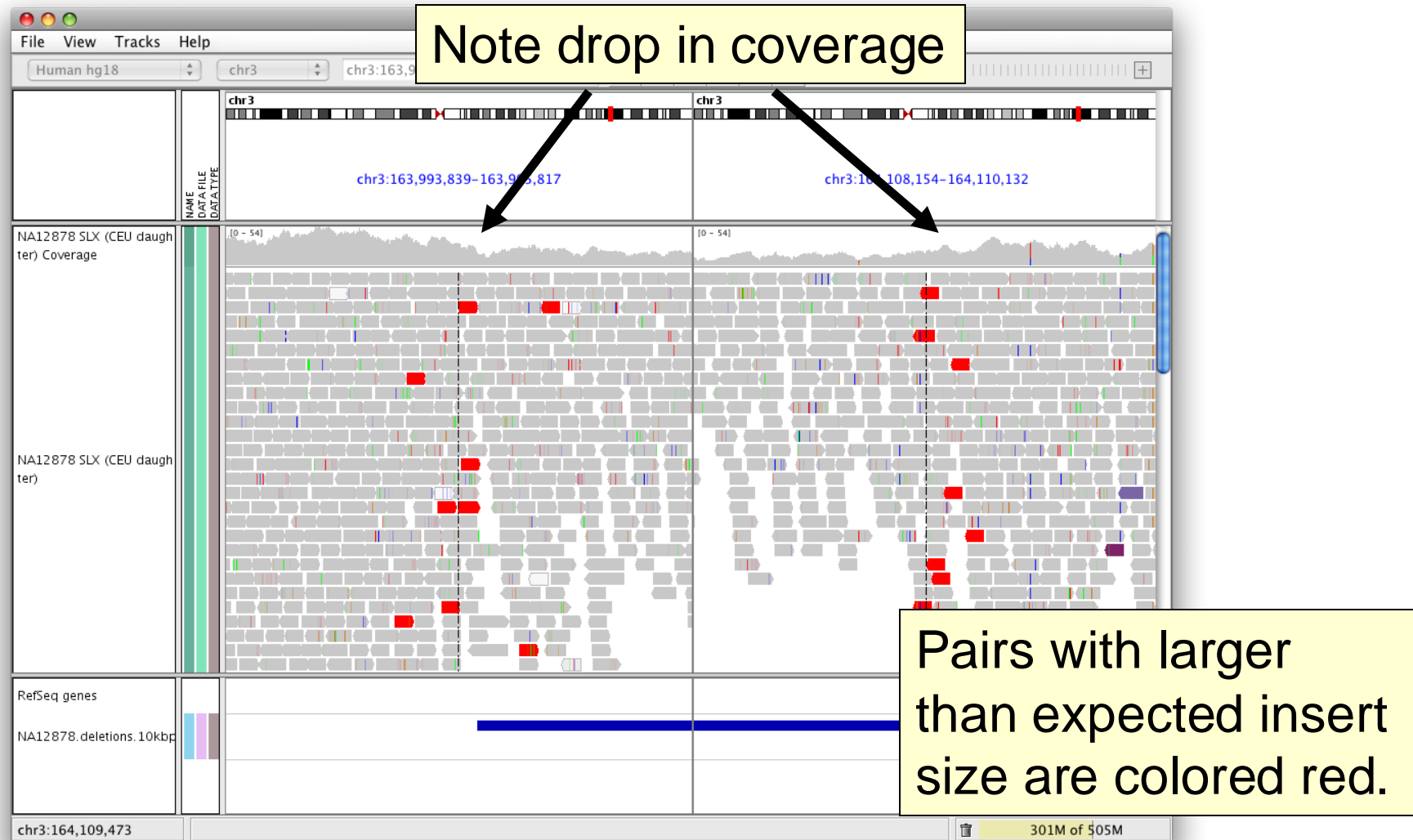


expected insert size



# Color by insert size



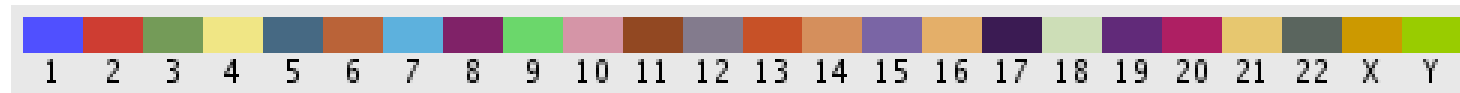
# Deletion



# Insert size color scheme

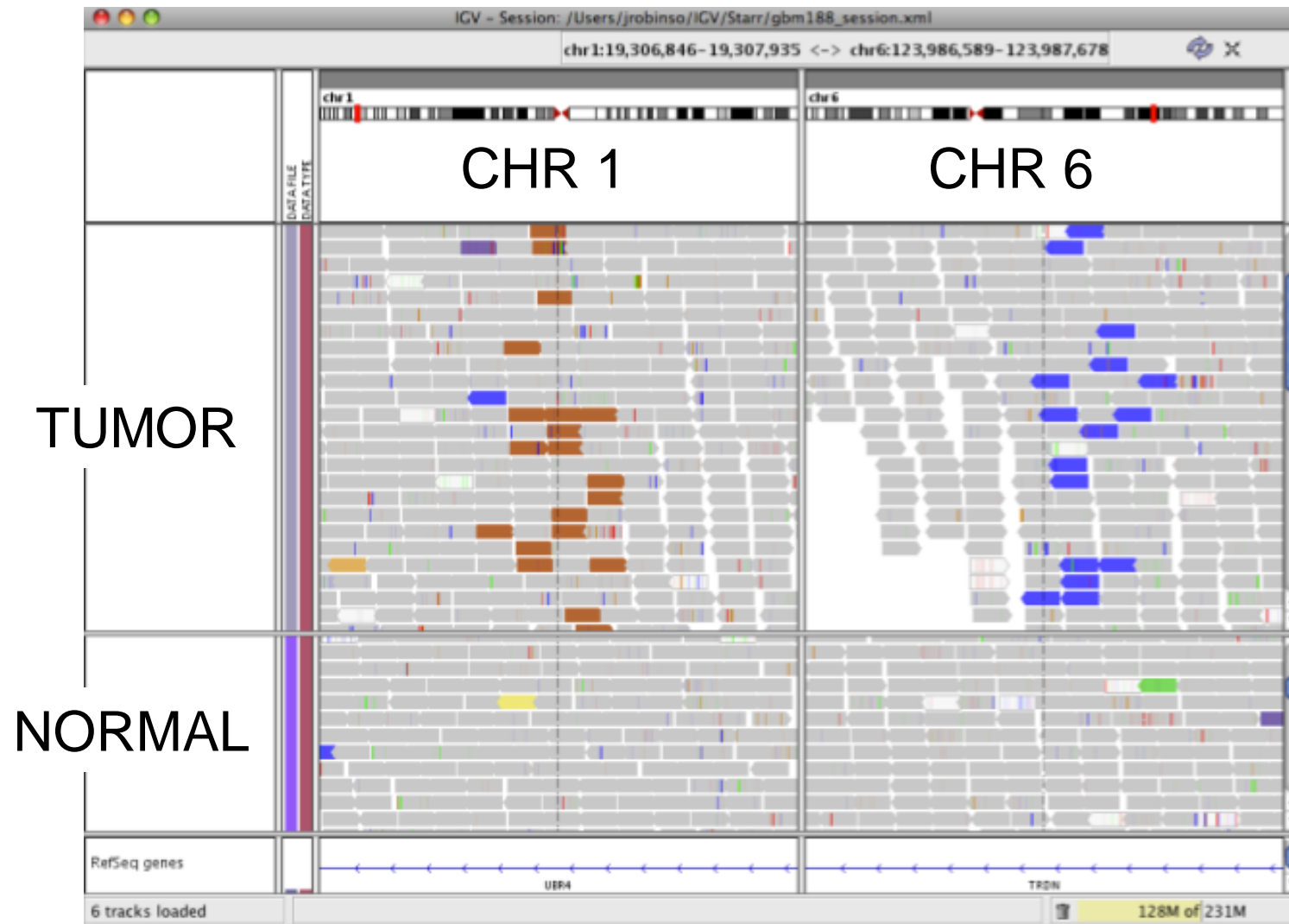
- Smaller than expected insert size: 
- Larger than expected insert size: 
- Pairs on different chromosomes

*Each end colored by chromosome of its mate*

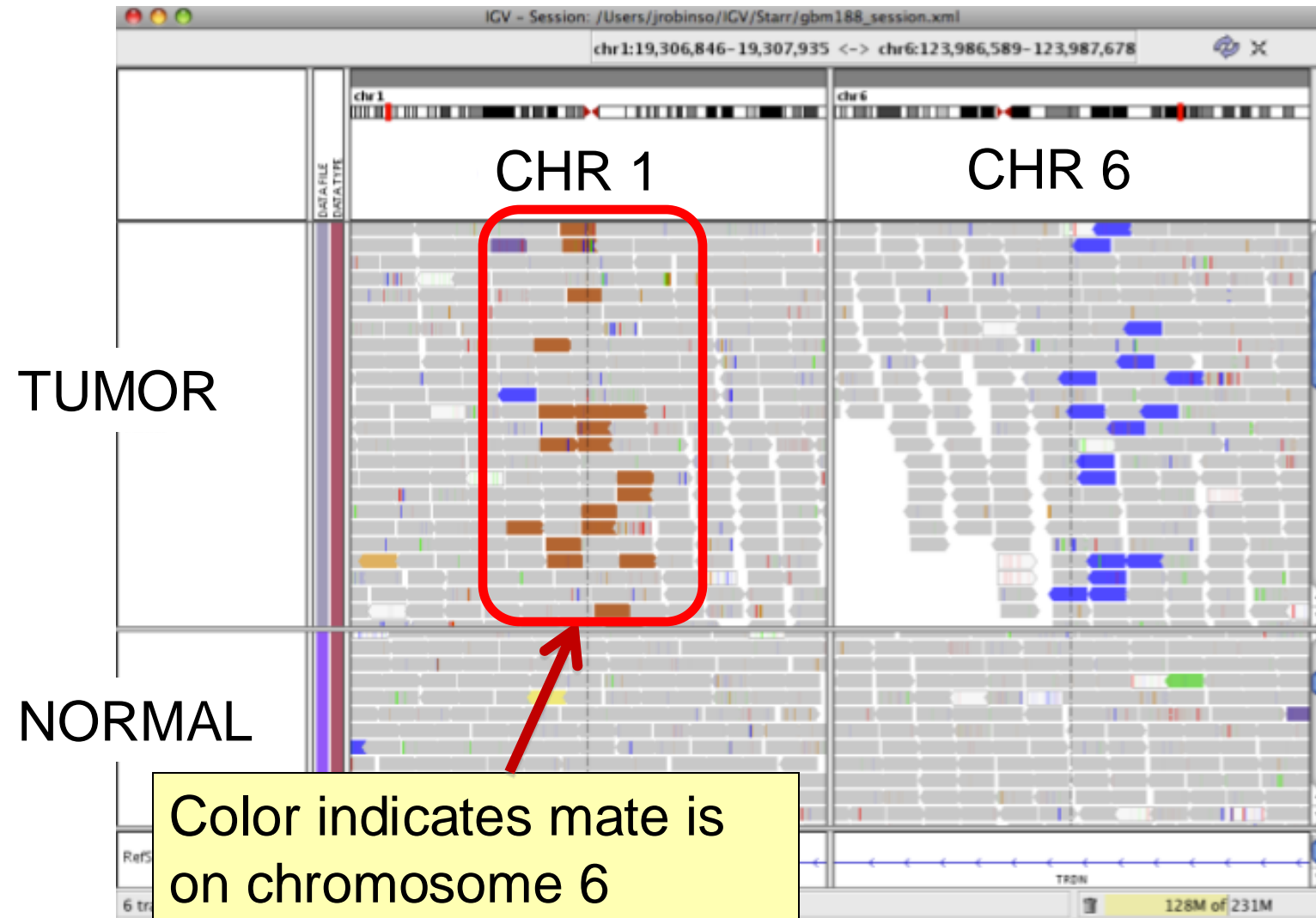




# Rearrangement



# 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

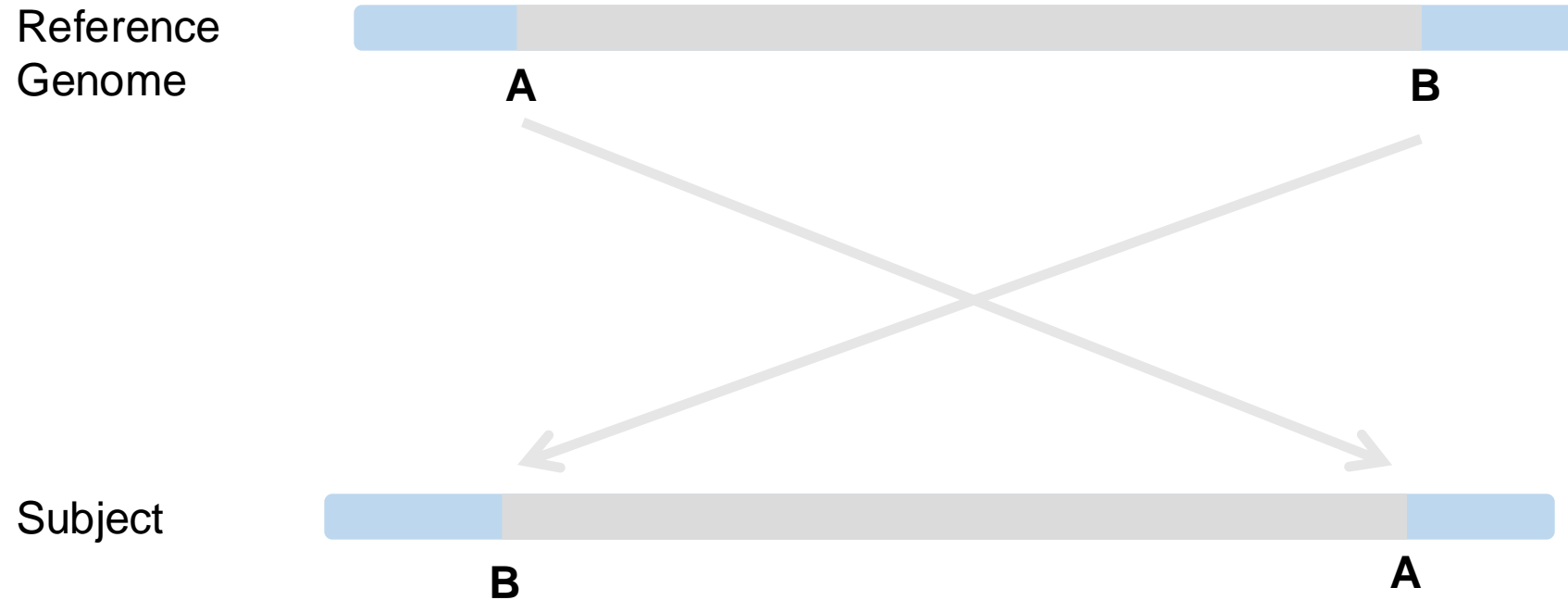


# Inversion

Reference  
genome



# Inversion

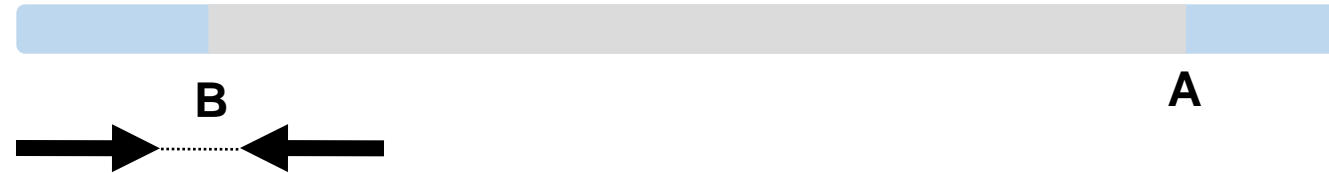


# Inversion

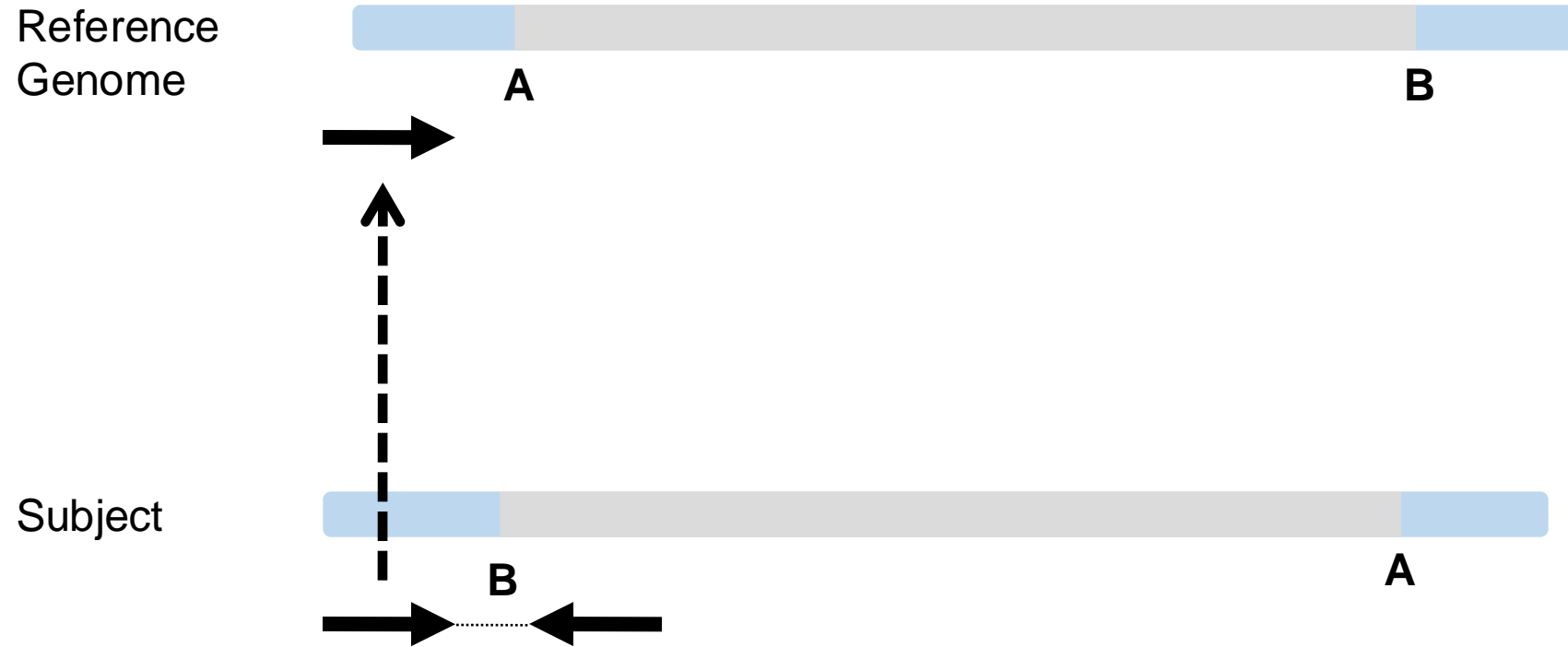
Reference  
Genome



Subject

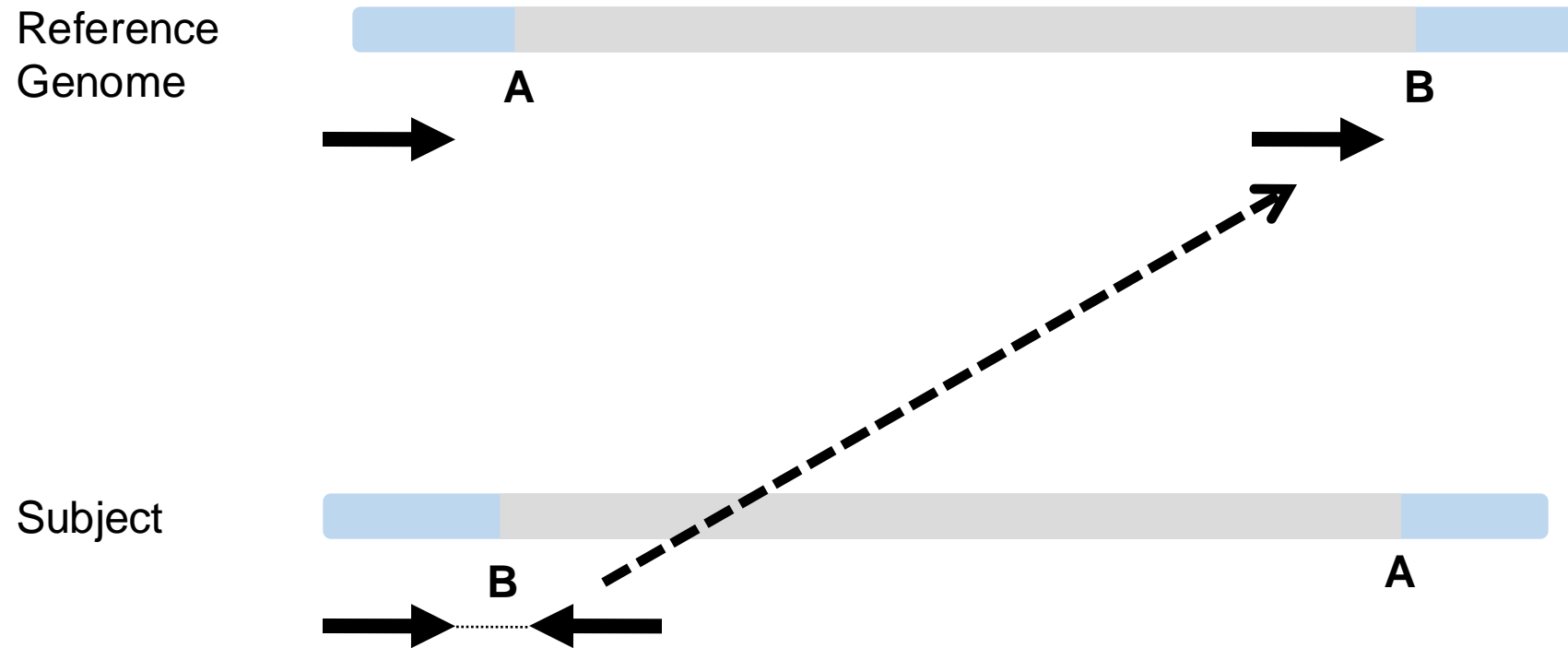


# Inversion



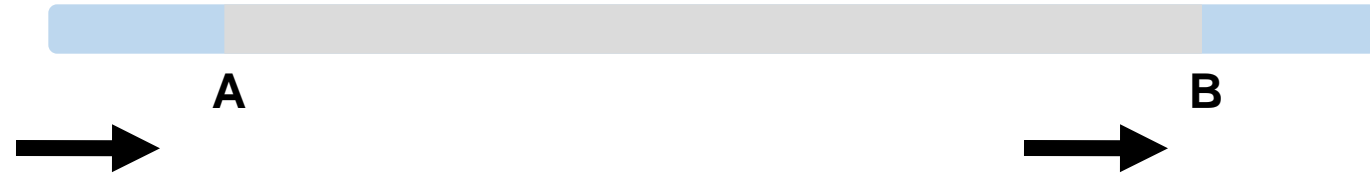


# Inversion

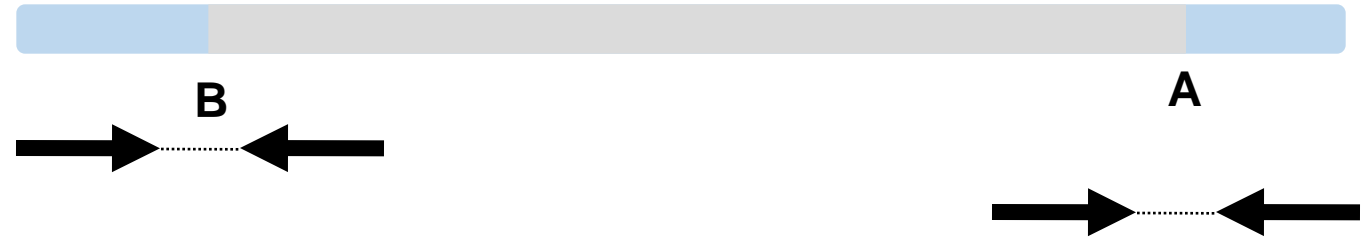


# Inversion

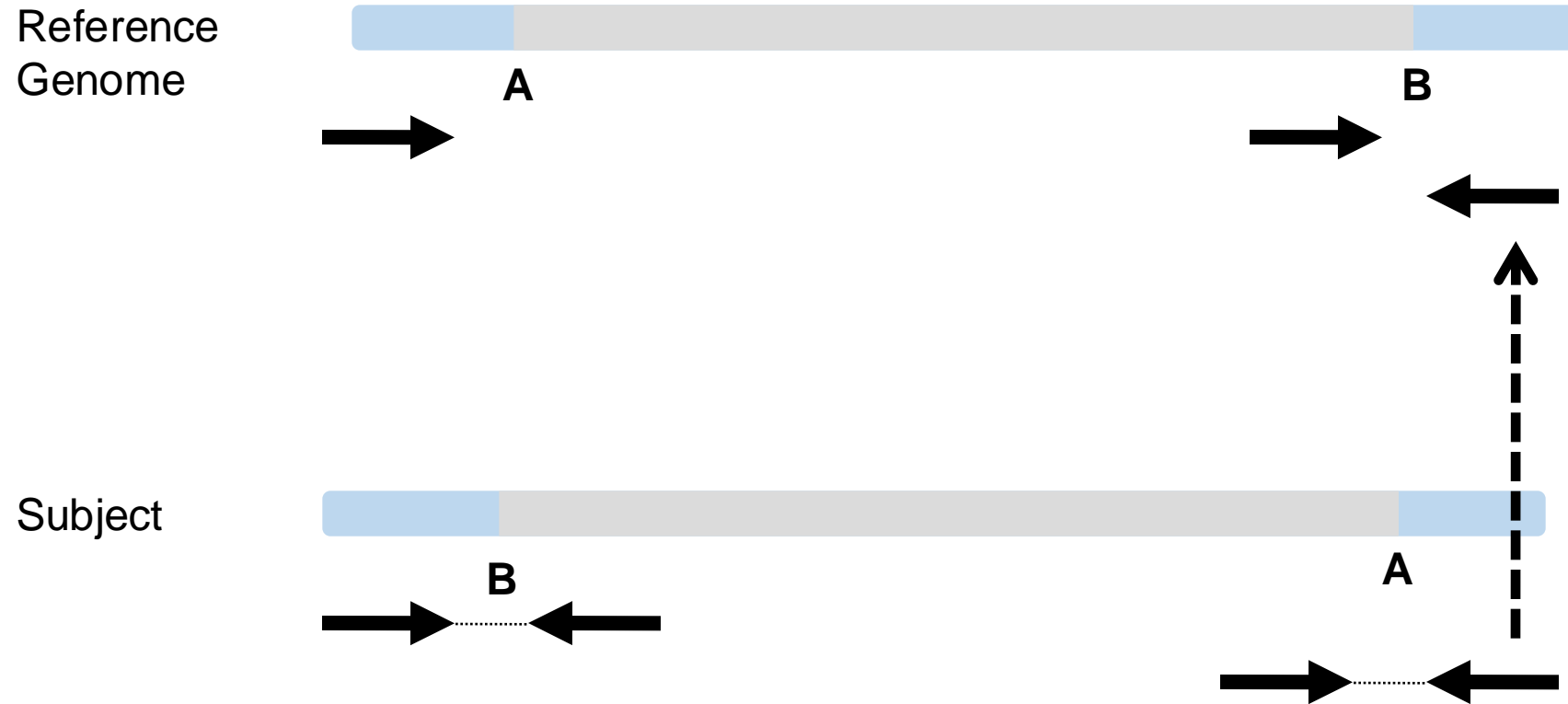
Reference  
Genome



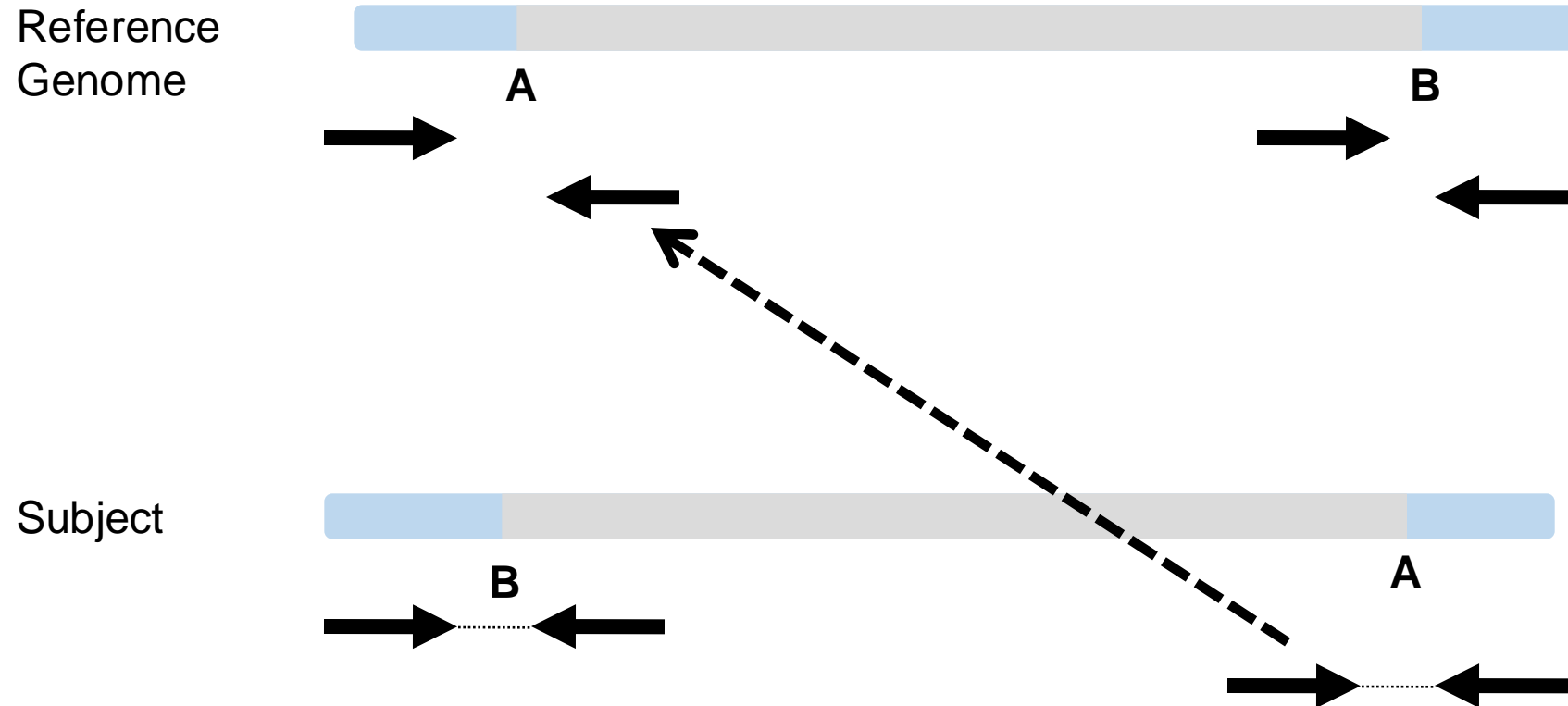
Subject



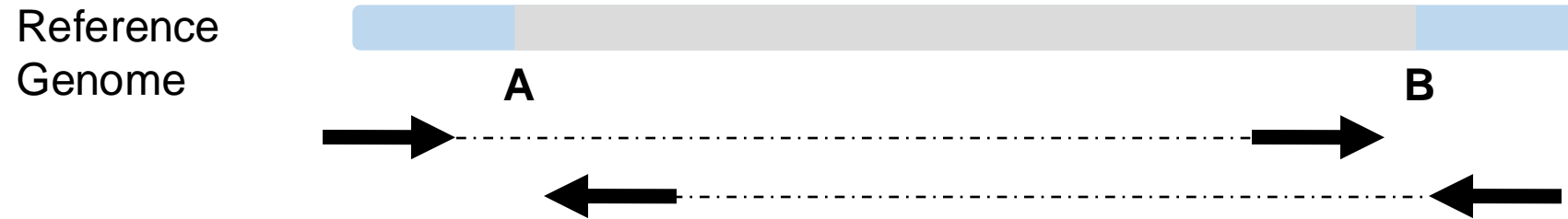
# Inversion



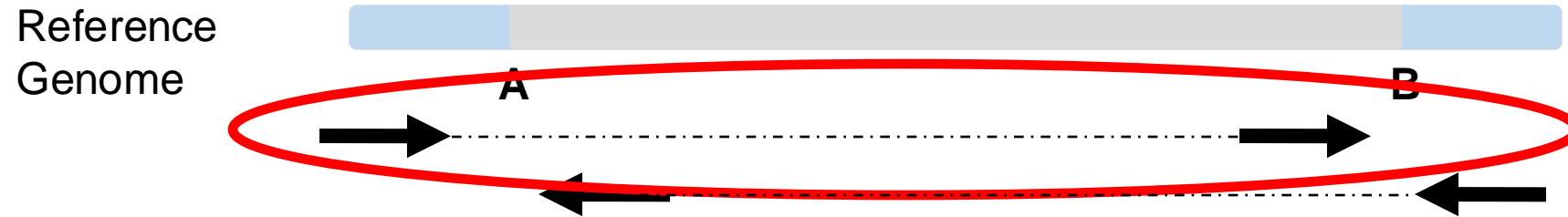
# Inversion



# Inversion

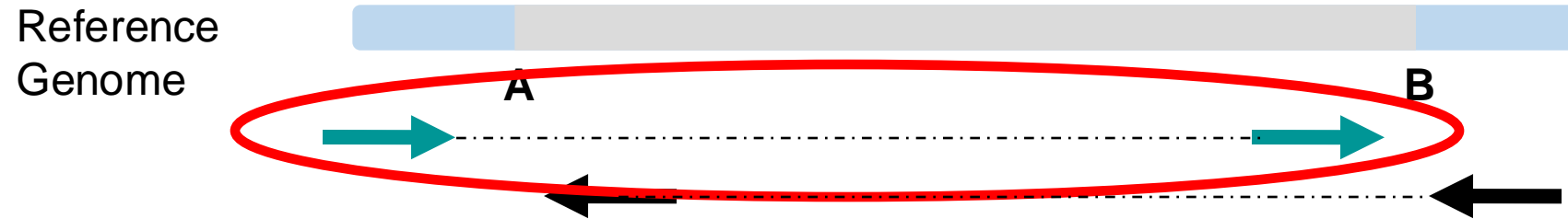


# Inversion



Anomaly: expected orientation of pair is  
inward facing (  $\rightarrow \leftarrow$  )

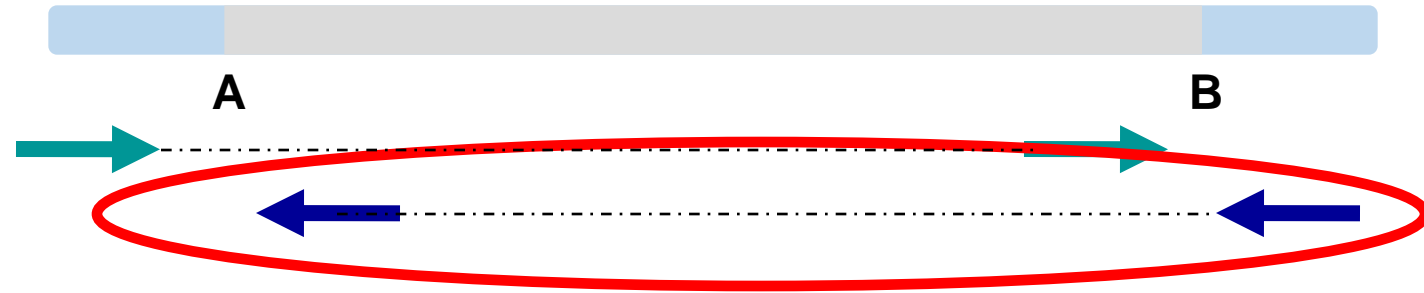
# Inversion



“Left” side pair

# Inversion

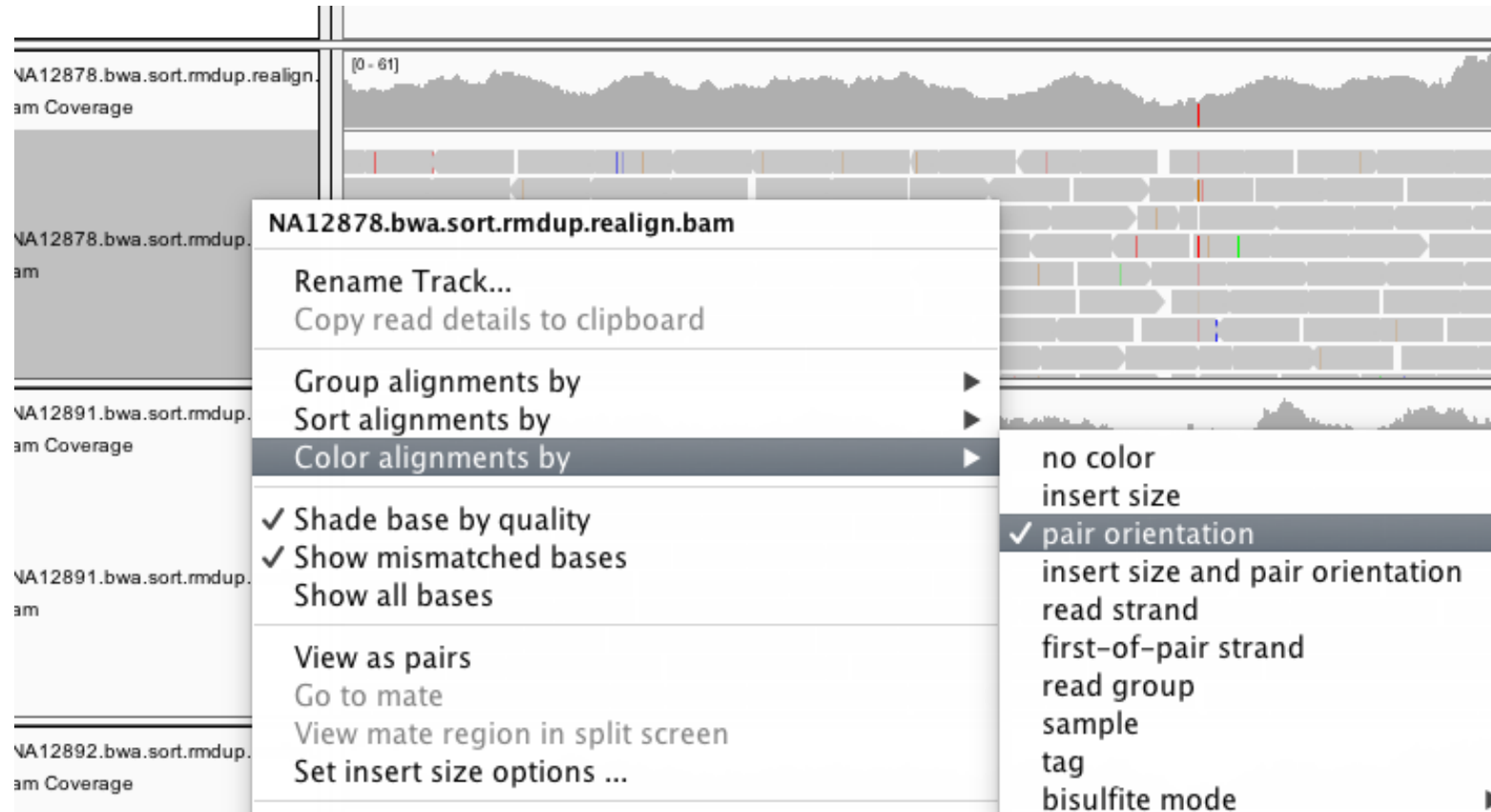
Reference  
Genome



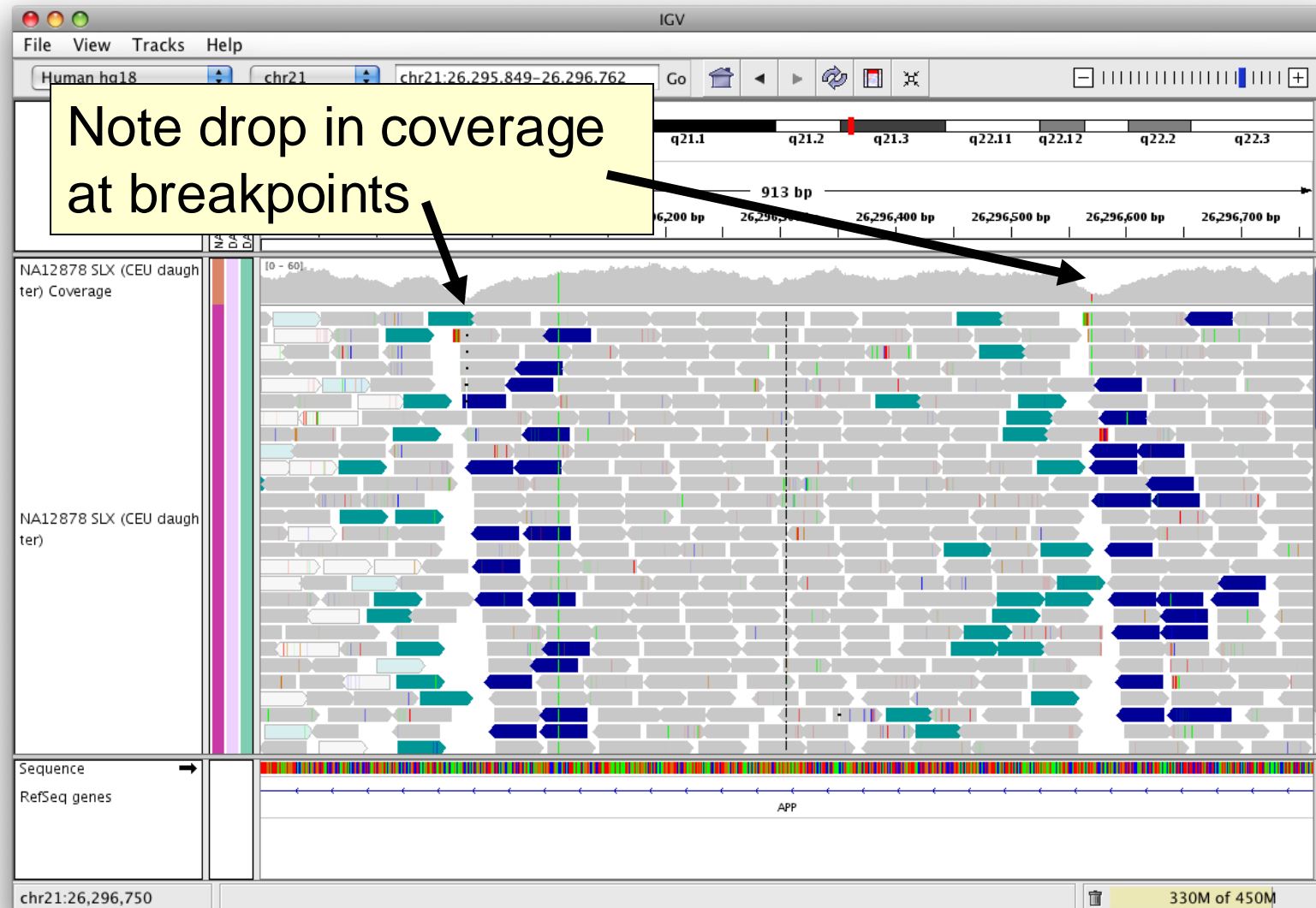
“Right” side pair



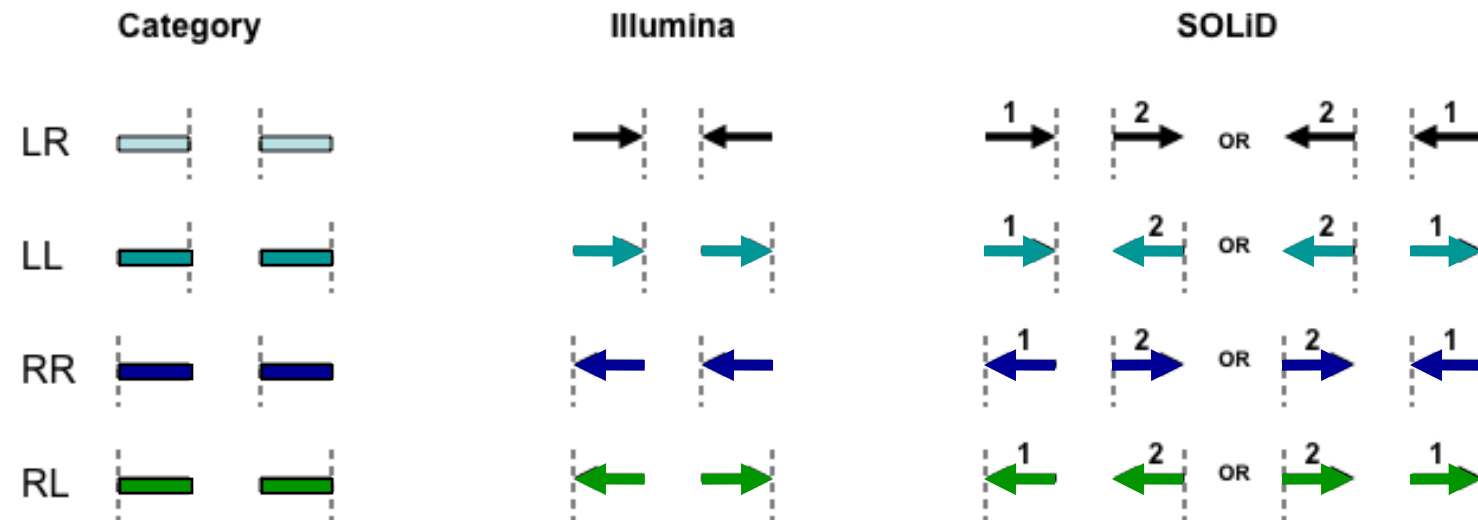
# Color by pair orientation



# Inversion



## Interpretation of read pair orientations

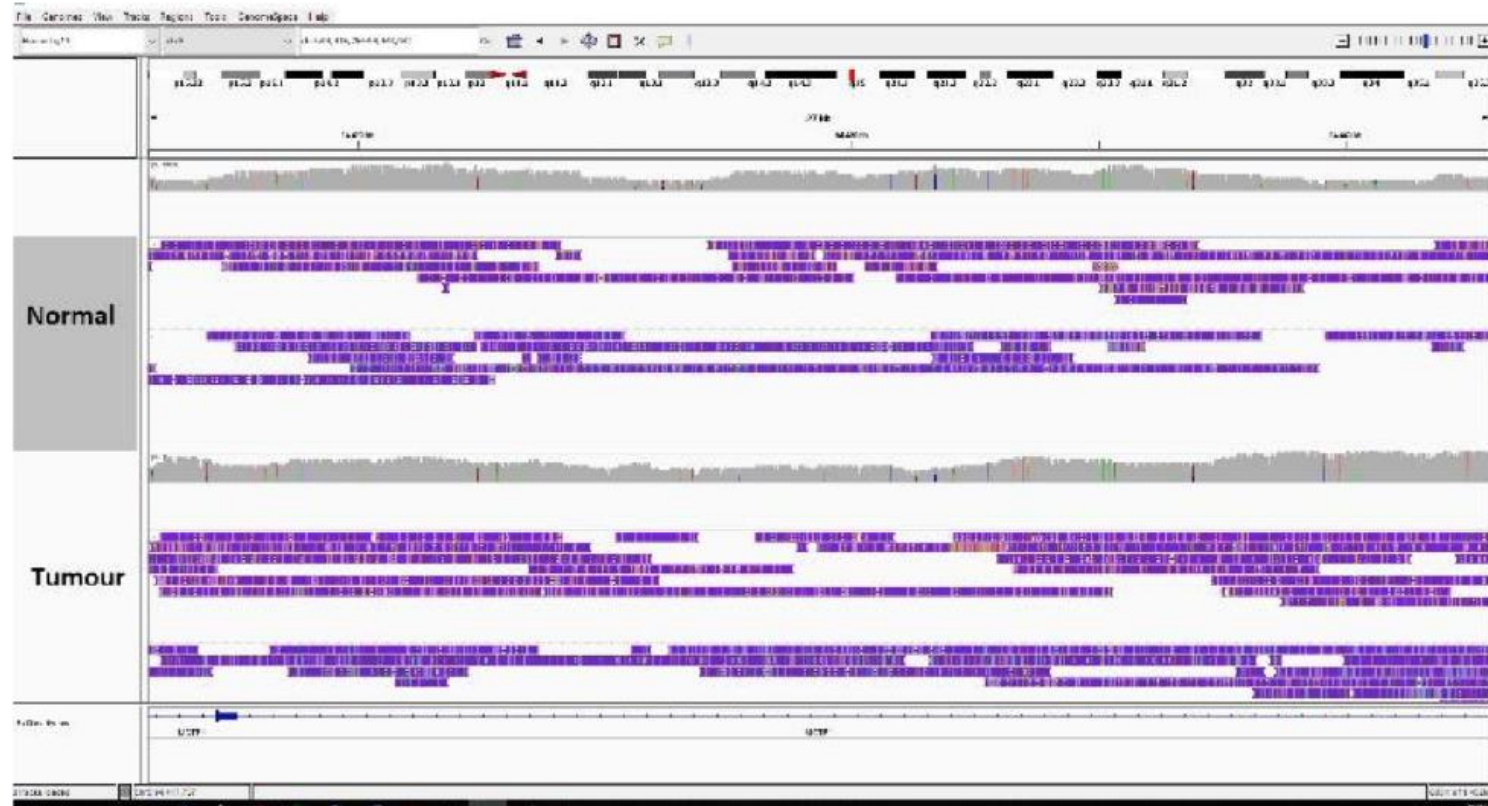


- LR      Normal reads.  
The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome.
- LL,RR      Implies inversion in sequenced DNA with respect to reference.
- RL      Implies duplication or translocation with respect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

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

*Figure courtesy of Bob Handsaker*

# Long read considerations



- Commonly see lots of small indels and single base errors that are simply noise
- Can be removed to be able to view the data more cleanly

# Long read considerations

The screenshot shows the IGV interface with the 'Alignments' track selected. The 'View' menu is open, showing 'Preferences...' and 'Color Legends ...'. The 'Alignments' track options are displayed, including 'Track Display Options', 'Alignment Track Options', 'Coverage Track Options', 'Splice Junction Track Options', and 'Insert Size Options'. A yellow box on the left contains the text 'Setting an indel threshold hides noise from small indels' with an arrow pointing to the 'Hide indels < 20 bases' checkbox in the 'Alignment Track Options' section.

**Setting an indel threshold hides noise from small indels**

**Alignment Track Options**

- On initial load show: ☒ Alignment Track ☒ Coverage Track ☐ Splice Junction Track
- Visibility range threshold (kb): 1000 *Range at which alignments become visible*
- ☐ Downsample reads Max read count: 100 per window size (bases): 50
- ☒ Shade mismatched bases by quality: 5 to 20
- Mapping quality threshold: 0
- ☒ Label indels > 1 bases
- ☒ Flag clipping > 0 bases
- ☒ Hide indels < 20 bases
- ☒ Filter duplicate reads
- ☐ Filter vendor failed reads
- ☐ Filter secondary alignments
- ☐ Show center line
- ☐ Flag unmapped pairs
- ☐ Show soft-clipped bases
- ☐ Quick consensus mode
- ☐ Filter supplementary alignments
- Hidden SAM ta... SA,MD,XA,RG

**Coverage Track Options**

- Coverage allele-fraction threshold: 0.2 ☒ Quality weight allele fraction

**Splice Junction Track Options**

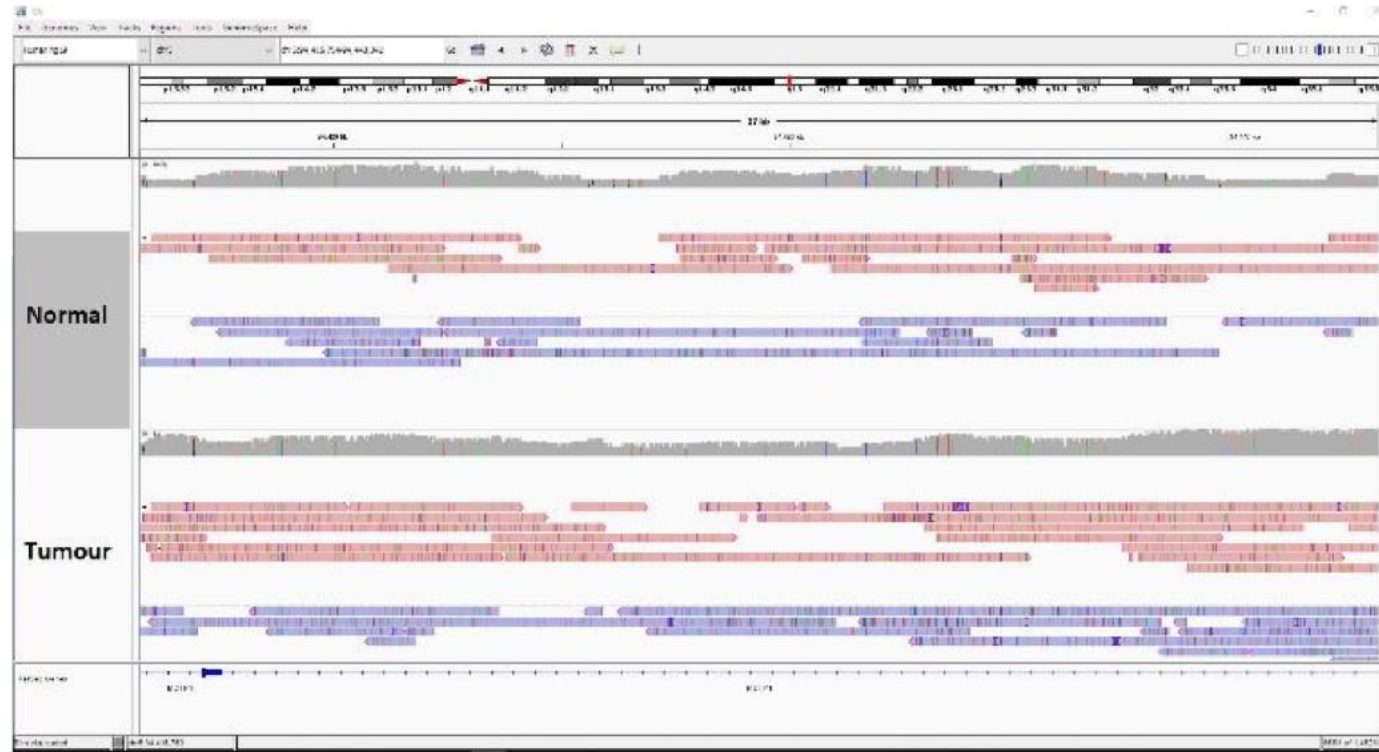
- ☐ Show flanking regions Min flanking width: 0 Min junction coverage: 1

**Insert Size Options**

- Defaults Minimum (bp): 50 Maximum (bp): 1000 ☒ Compute Minimum (percentile): 0.5 Maximum (percentile): 99.5

OK Cancel

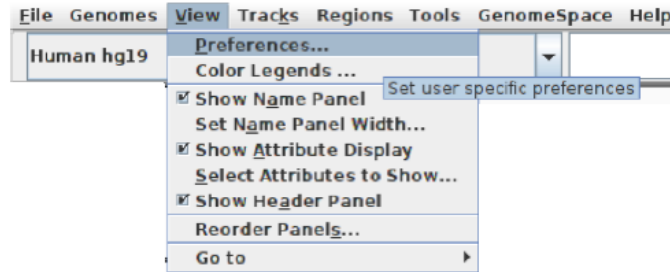
# Long read considerations



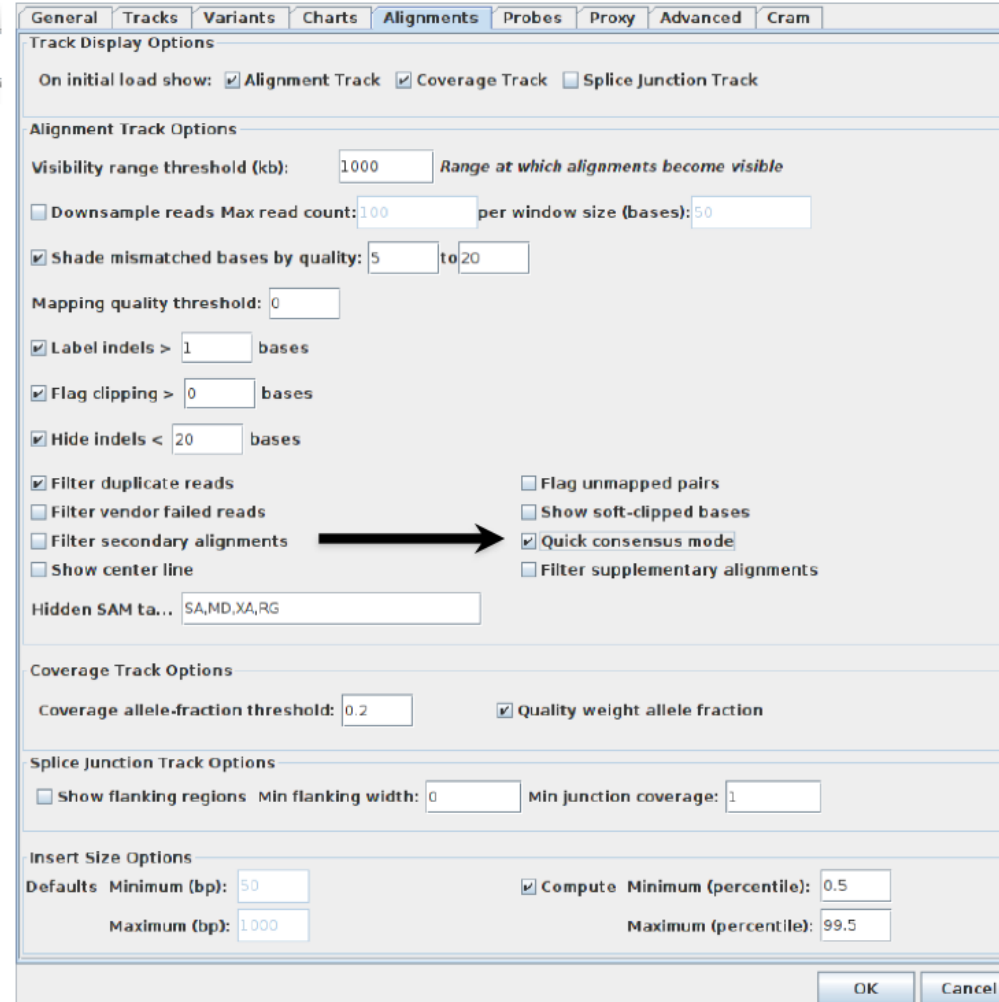
- Reads are not all purple dashes
- Next step would be to call a consensus at each position



# Long read considerations



Option for  
generating  
consensus  
sequences



# Long read considerations



- Much easier to parse through the genomic data
- Large insertions and deletions are also labelled now



# 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.

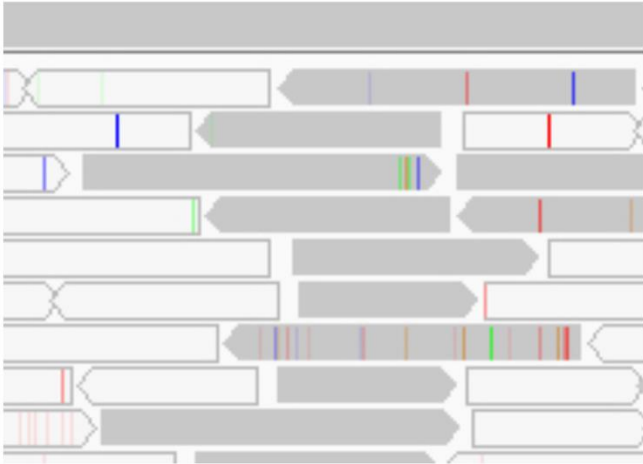
**Results:** After reading the SOP, average accuracy in somatic variant identification increased by 16.7% ( $p$  value = 0.0298) and average interreviewer agreement increased by 12.7% ( $p$  value < 0.001). Manual review conducted after reading the SOP did not significantly increase reviewer time.

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

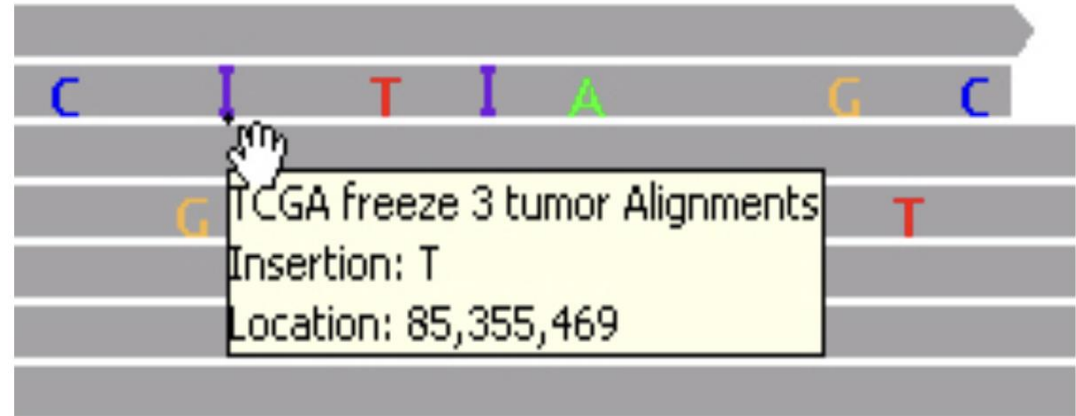
*Genetics in Medicine* (2018) <https://doi.org/10.1038/s41436-018-0278-z>

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

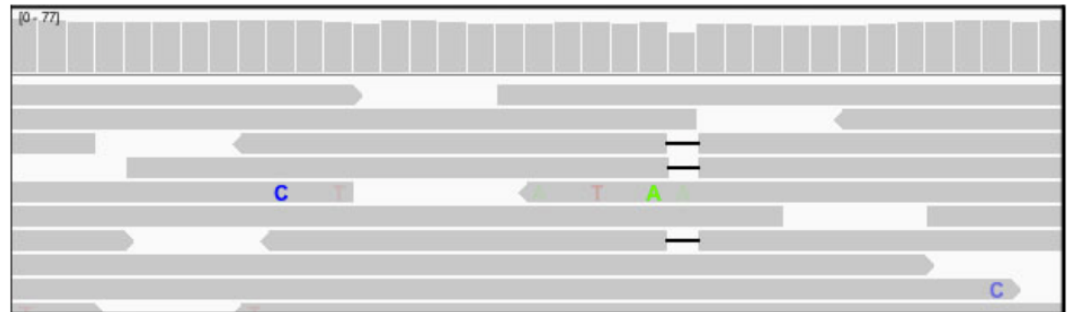
# Other notes



Transparent (White) reads:  
Low quality reads/  
mapping quality equal to zero



Purple **I** : Insertion



Gapped read/black bar: Deletion

We are on a Coffee Break & Networking  
Session