bioinformatics.ca

# Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

# CC BY-SA 4.0 DEED

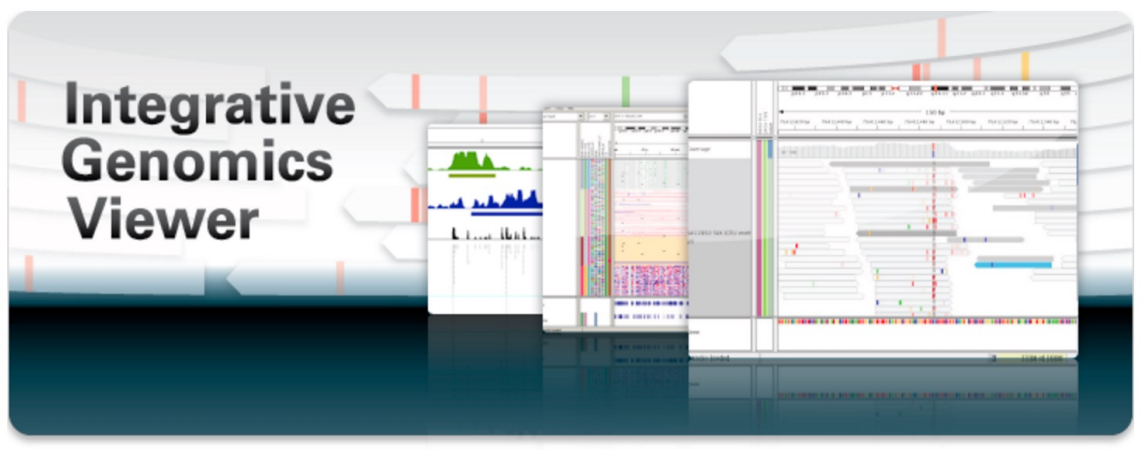
**Attribution-ShareAlike 4.0 International**

# The Integrative Genomics Viewer

Malachi Griffith, Obi Griffith, Isabel Risch, Vida Talebian

RNA-seq Analysis 2024. June 17-19, 2024

bioinformatics.ca




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

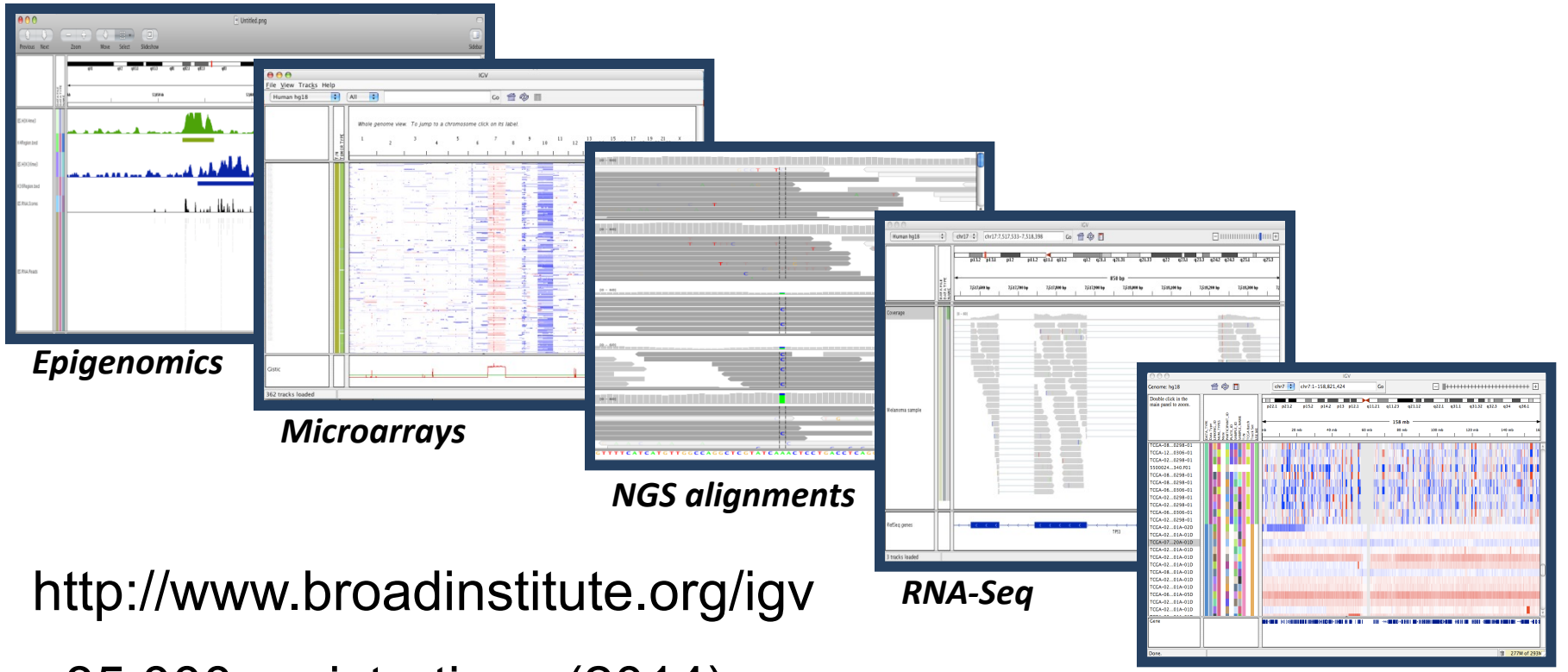

Trackster (part of Galaxy)

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*




http://www.broadinstitute.org/igv

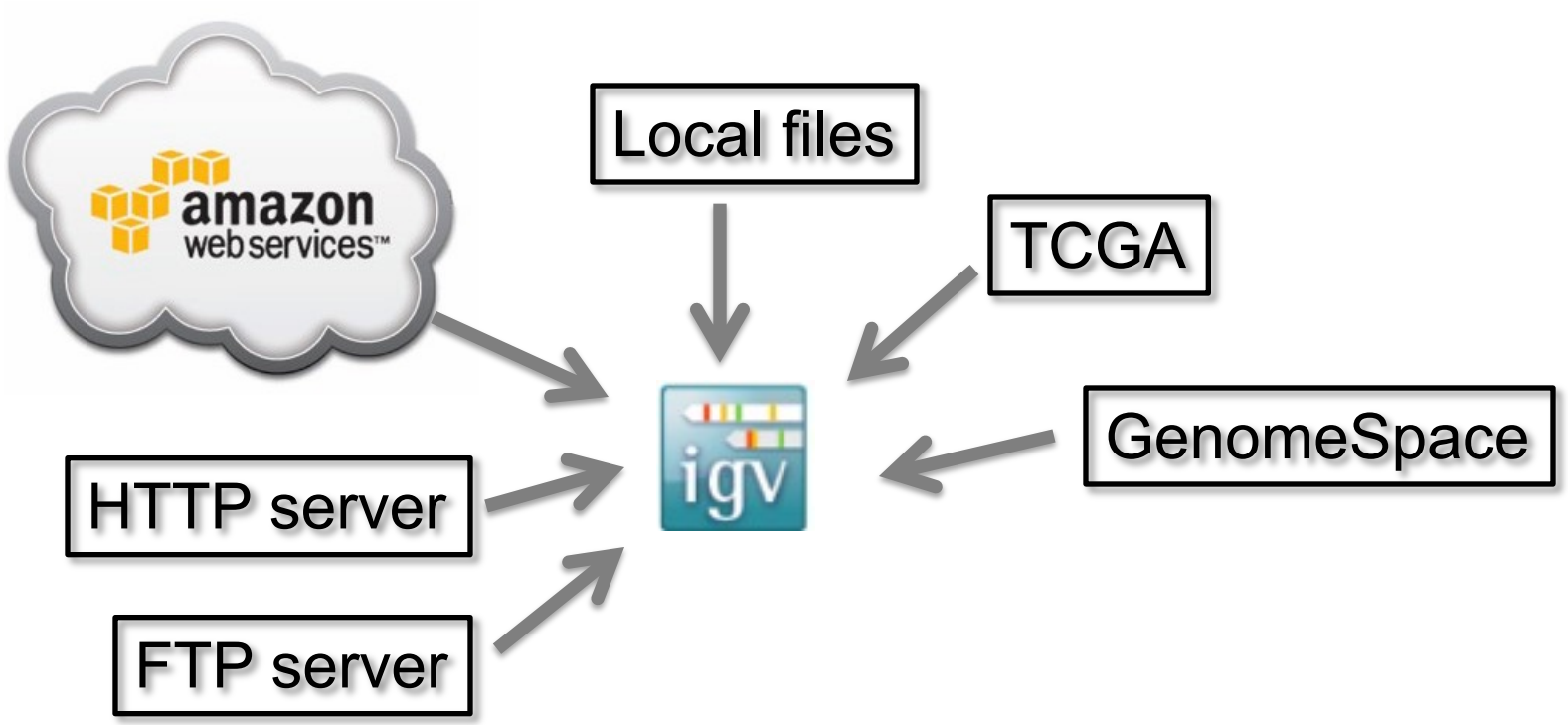
>85,000 registrations (2014)

# 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



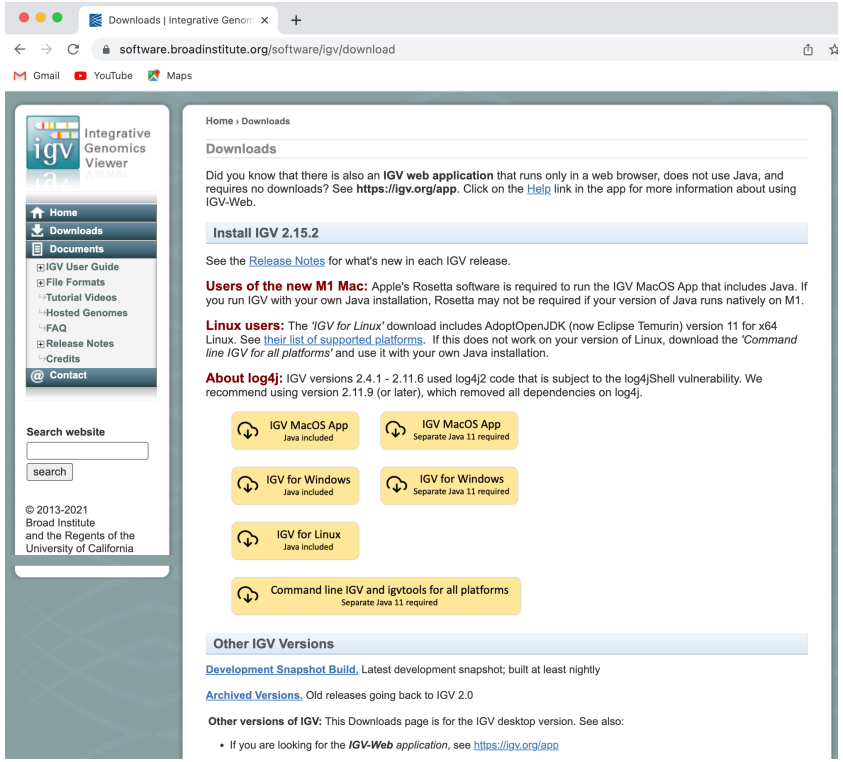

- View **local** files without uploading.
- View **remote** files without downloading the whole dataset.

# Using IGV: the basics

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

# Launch IGV

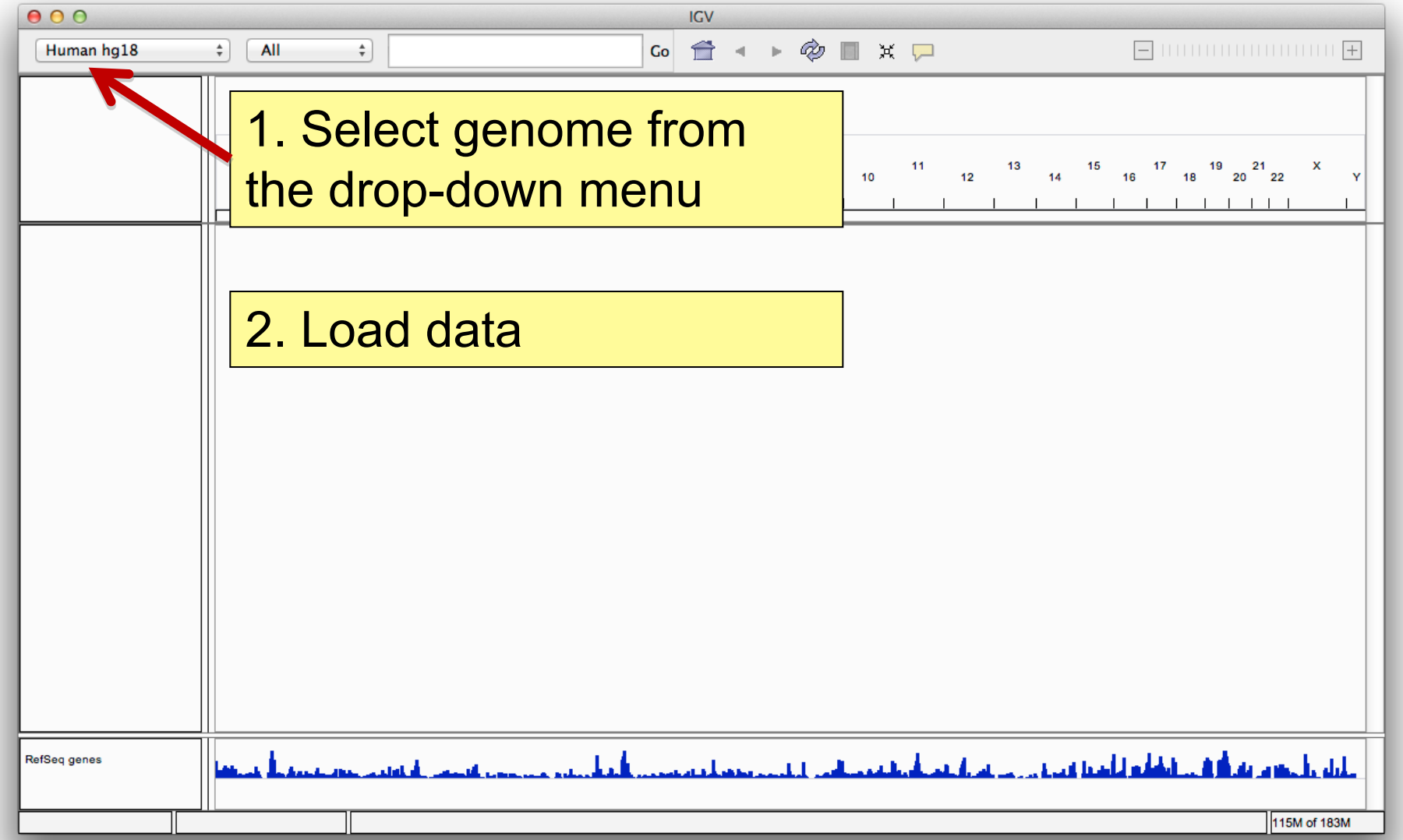

Home | Integrative Genomics Viewer
https://www.broadinstitute.org/igv/
Integrative Genomics Viewer
Downloads
Documents
FAQ
IGV User Guide
File Formats
Release Notes
Credits
Contact
Search website
search
Broad Home
Cancer Program
© 2012 Broad Institute
Home
Integrative Genomics Viewer
What's New
July 3, 2012. Soybean (Glycine max) and Rat (rn5) genomes have been updated.
April 20, 2012. IGV 2.1 has been released. See the release notes for more details.
April 19, 2012. See our new IGV paper in Briefings in Bioinformatics.
Overview
The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.
Downloads
Please register to download IGV. After registering, you can log in at any time using your email address. Permission to use IGV is granted under the GNU LGPL license.
Citing IGV
To cite your use of IGV in your publication:
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011), or
Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 2012.
Funding
Development of IGV is made possible by funding from the National Cancer Institute, the National Institute of General Medical Sciences of the National Institutes of Health, and the Starr Cancer Consortium.
IGV is participating in the GenomeSpace initiative.
GENOMESPACE


Downloads | Integrative Genom
software.broadinstitute.org/software/igv/download
Gmail
YouTube
Maps
Integrative Genomics Viewer
Home
Downloads
Documents
IGV User Guide
File Formats
Tutorial Videos
Hosted Genomes
FAQ
Release Notes
Credits
Contact
Search website
search
© 2013-2021
Broad Institute
and the Regents of the
University of California
Home › Downloads
Downloads
Did you know that there is also an IGV web application that runs only in a web browser, does not use Java, and requires no downloads? See https://igv.org/app. Click on the Help link in the app for more information about using IGV-Web.
Install IGV 2.15.2
See the Release Notes for what's new in each IGV release.
Users of the new M1 Mac: Apple's Rosetta software is required to run the IGV MacOS App that includes Java. If you run IGV with your own Java installation, Rosetta may not be required if your version of Java runs natively on M1.
Linux users: The 'IGV for Linux' download includes AdoptOpenJDK (now Eclipse Temurin) version 11 for x64 Linux. See their list of supported platforms. If this does not work on your version of Linux, download the 'Command line IGV for all platforms' and use it with your own Java installation.
About log4j: IGV versions 2.4.1 - 2.11.6 used log4j2 code that is subject to the log4jShell vulnerability. We recommend using version 2.11.9 (or later), which removed all dependencies on log4j.
IGV MacOS App
Java included
IGV MacOS App
Separate Java 11 required
IGV for Windows
Java included
IGV for Windows
Separate Java 11 required
IGV for Linux
Java included
Command line IGV and igvtools for all platforms
Separate Java 11 required
Other IGV Versions
Development Snapshot Build. Latest development snapshot; built at least nightly
Archived Versions. Old releases going back to IGV 2.0
Other versions of IGV: This Downloads page is for the IGV desktop version. See also:
• If you are looking for the IGV-Web application, see https://igv.org/app


IGV: Integrative Genomics Viev
igv.org
Gmail
YouTube
Maps
IGV
Integrative Genomics Viewer
IGV
IGV desktop application
IGV Web App
IGV in a web browser
Juicebox Web
Hi-C contact map viewer

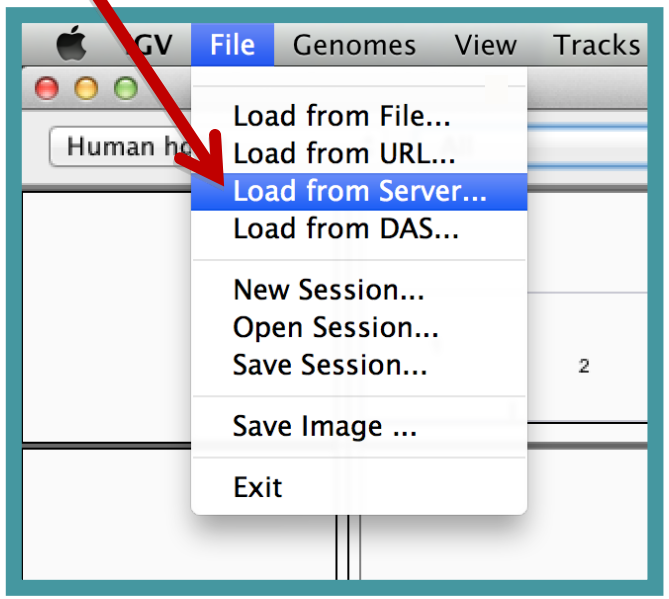

igv

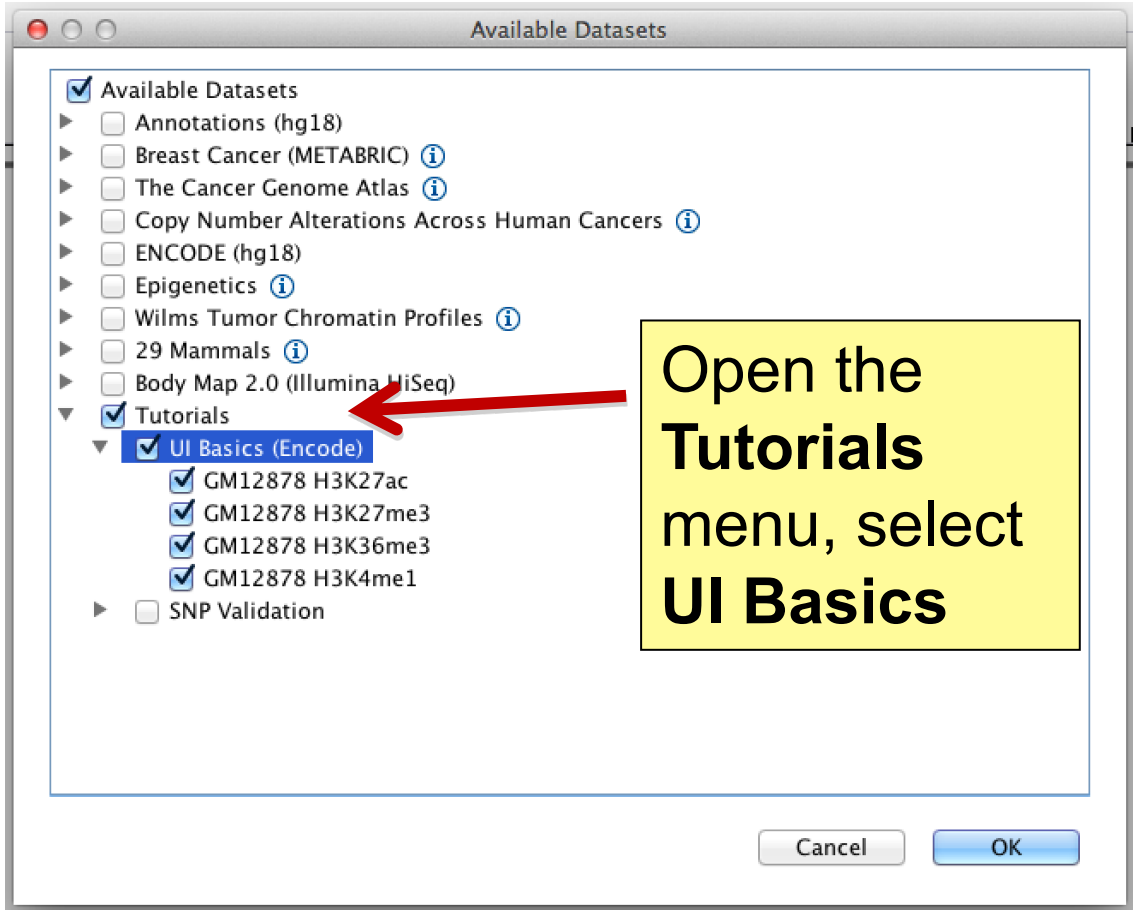
# Select reference genome



# Load data

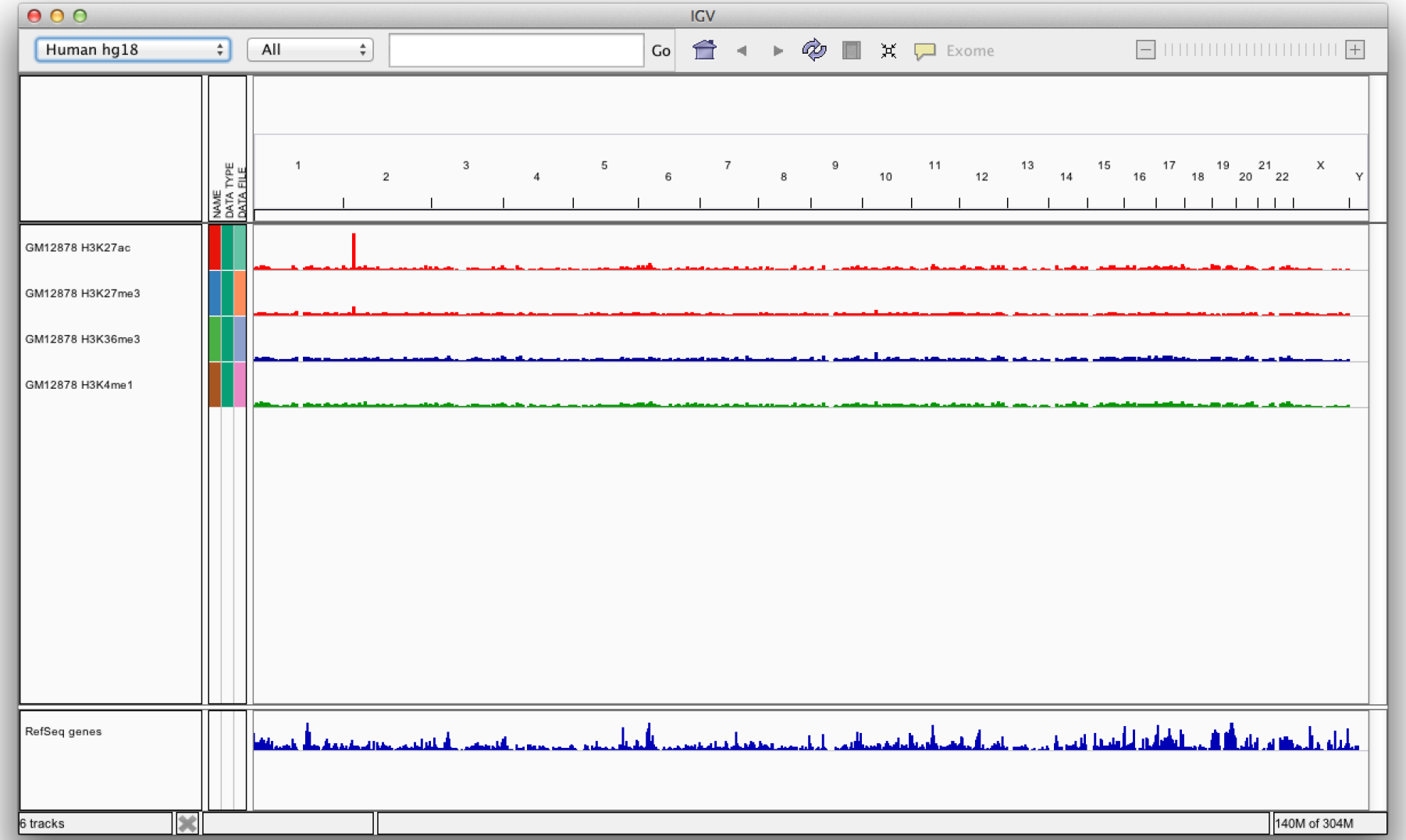
Select **File > Load from Server…**



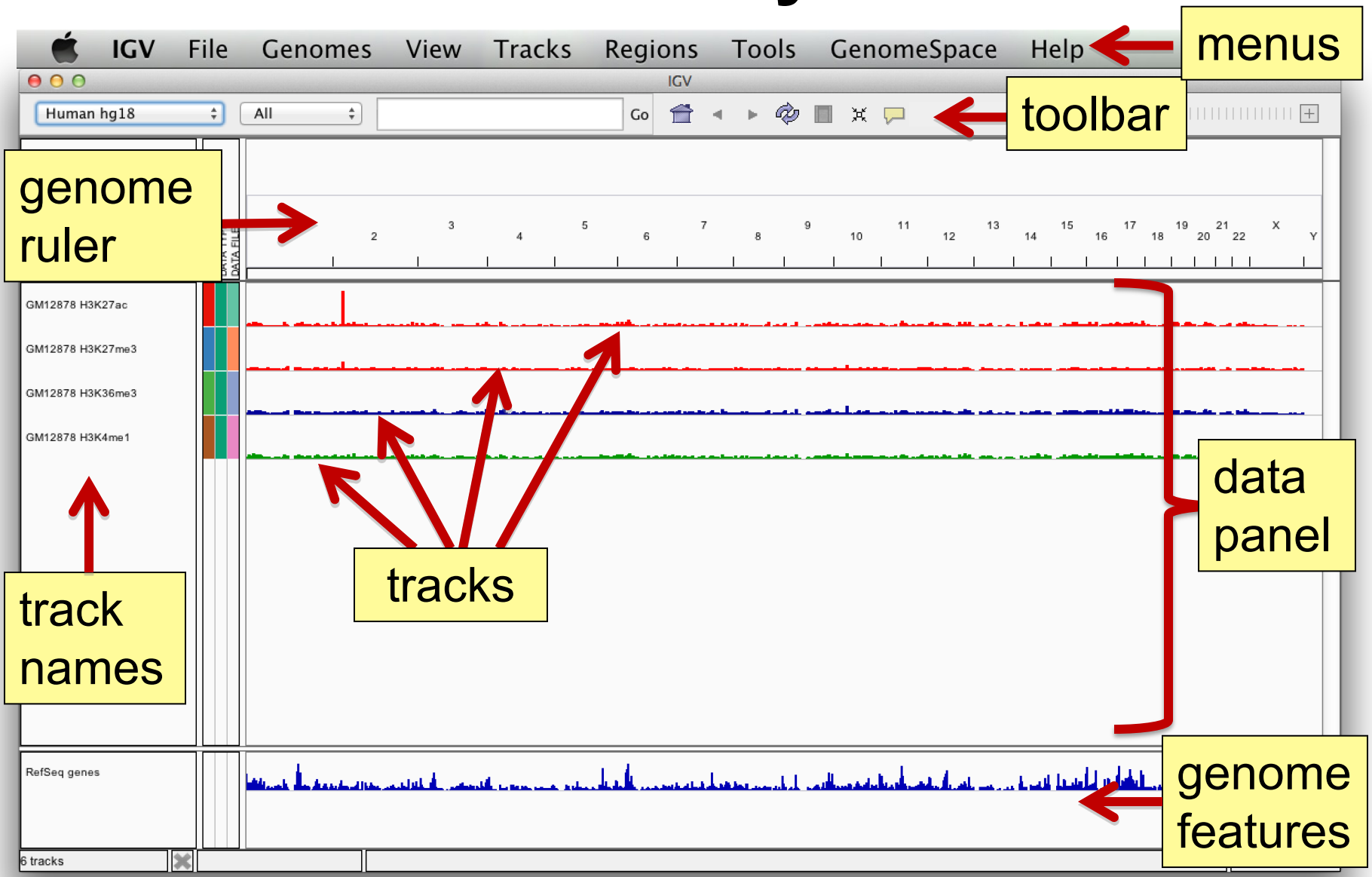




Open the **Tutorials** menu, select **UI Basics**

# Screen layout



# Screen layout



# File formats and track types

- The **file format** defines the track type.
- The **track type** determines the display options

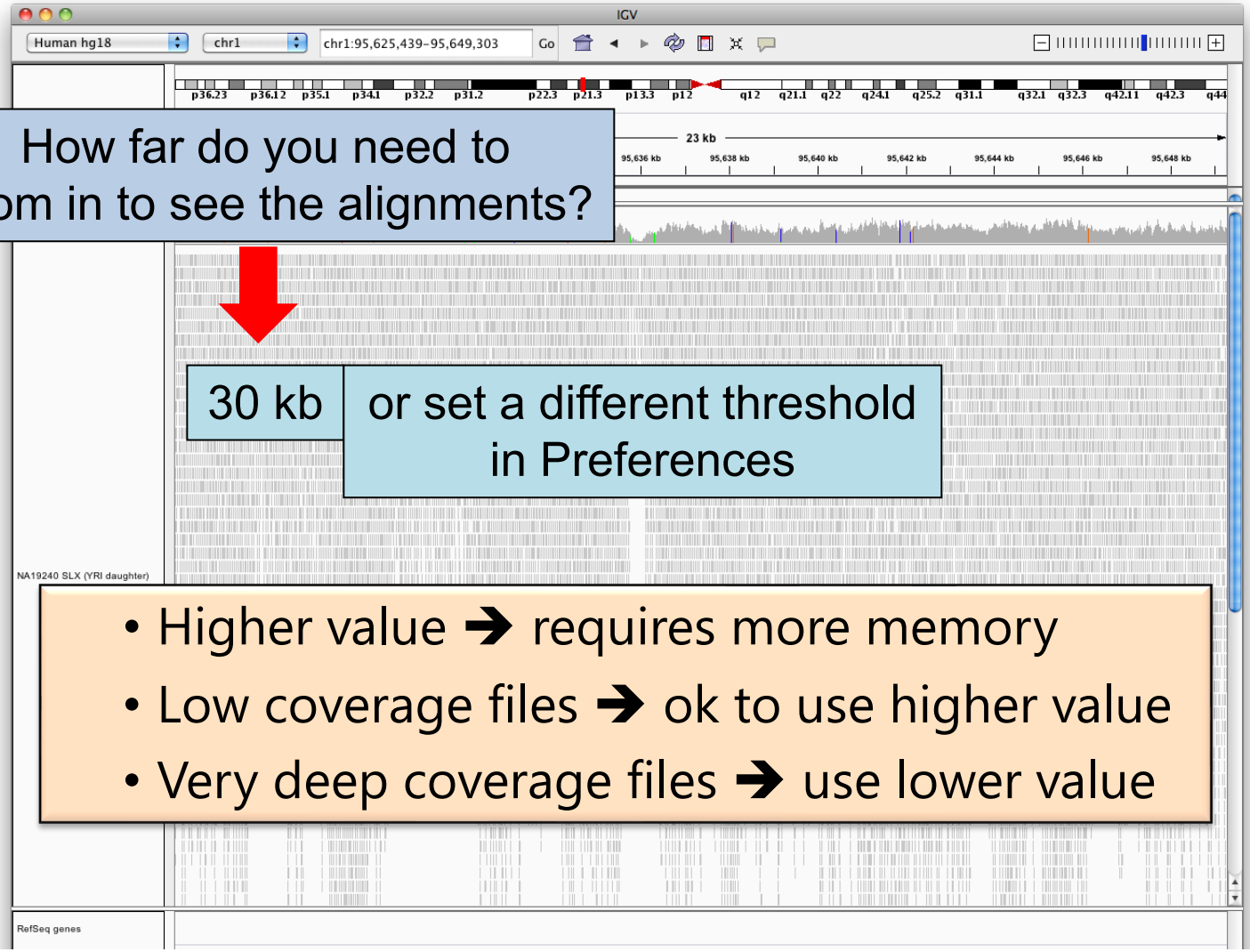
  - BAM
  - BED
  - BEDPE
  - BedGraph
  - bigBed
  - bigWig
  - Birdsuite Files
  - broadPeak
  - CBS
  - Chemical Reactivity Probing Profiles
  - chrom.sizes
  - CN
  - Custom File Formats
  - Cytoband
  - FASTA
  - GCT
  - CRAM
  - genePred
  - GFF/GTF
  - GISTIC
  - Goby
  - GWAS
  - IGV
  - LOH
  - 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 current list see: https://software.broadinstitute.org/software/igv/FileFormats

# Viewing alignments

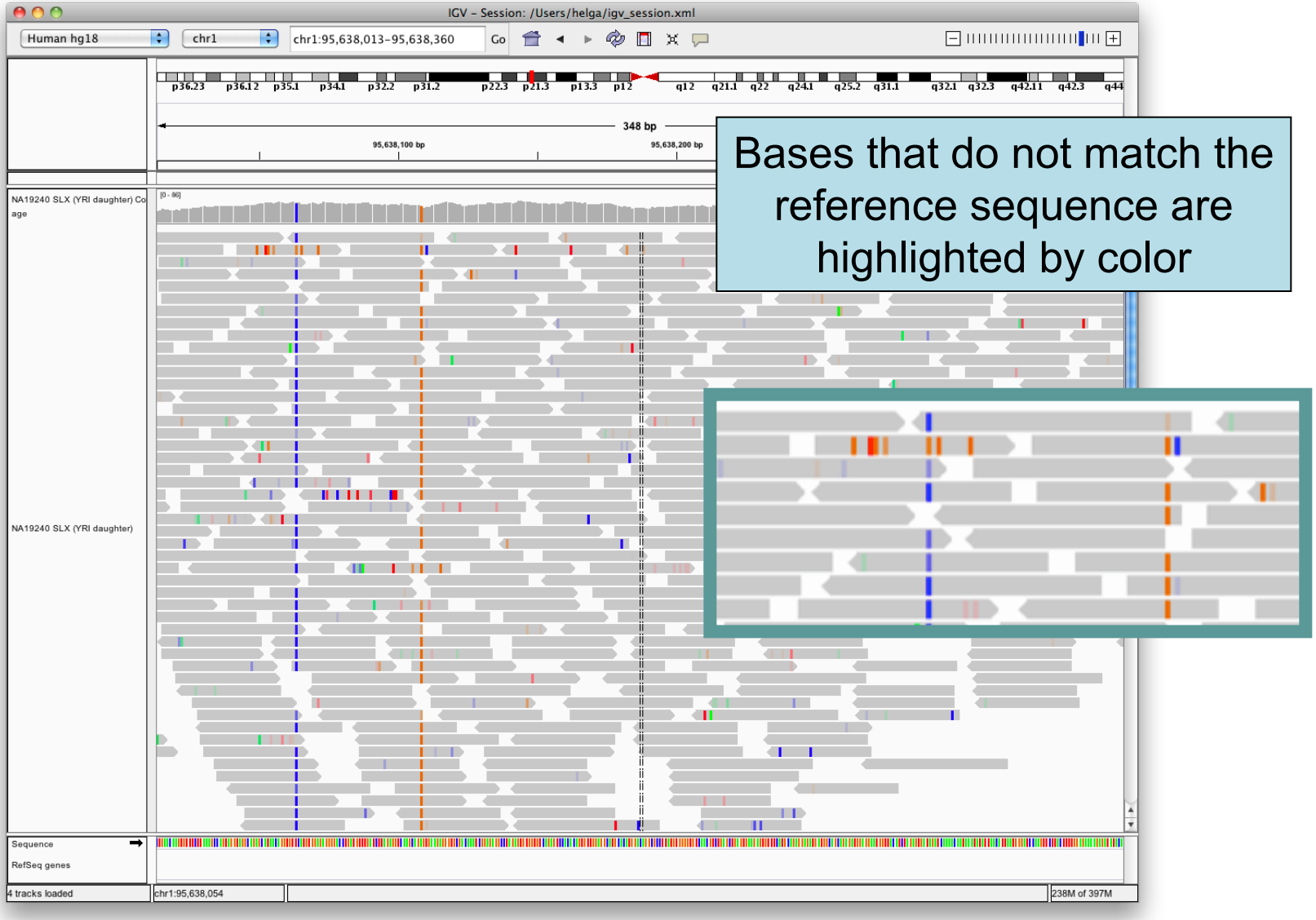
## Whole chromosome view

# Viewing alignments – Zoom in




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

# Viewing alignments – Zoom in

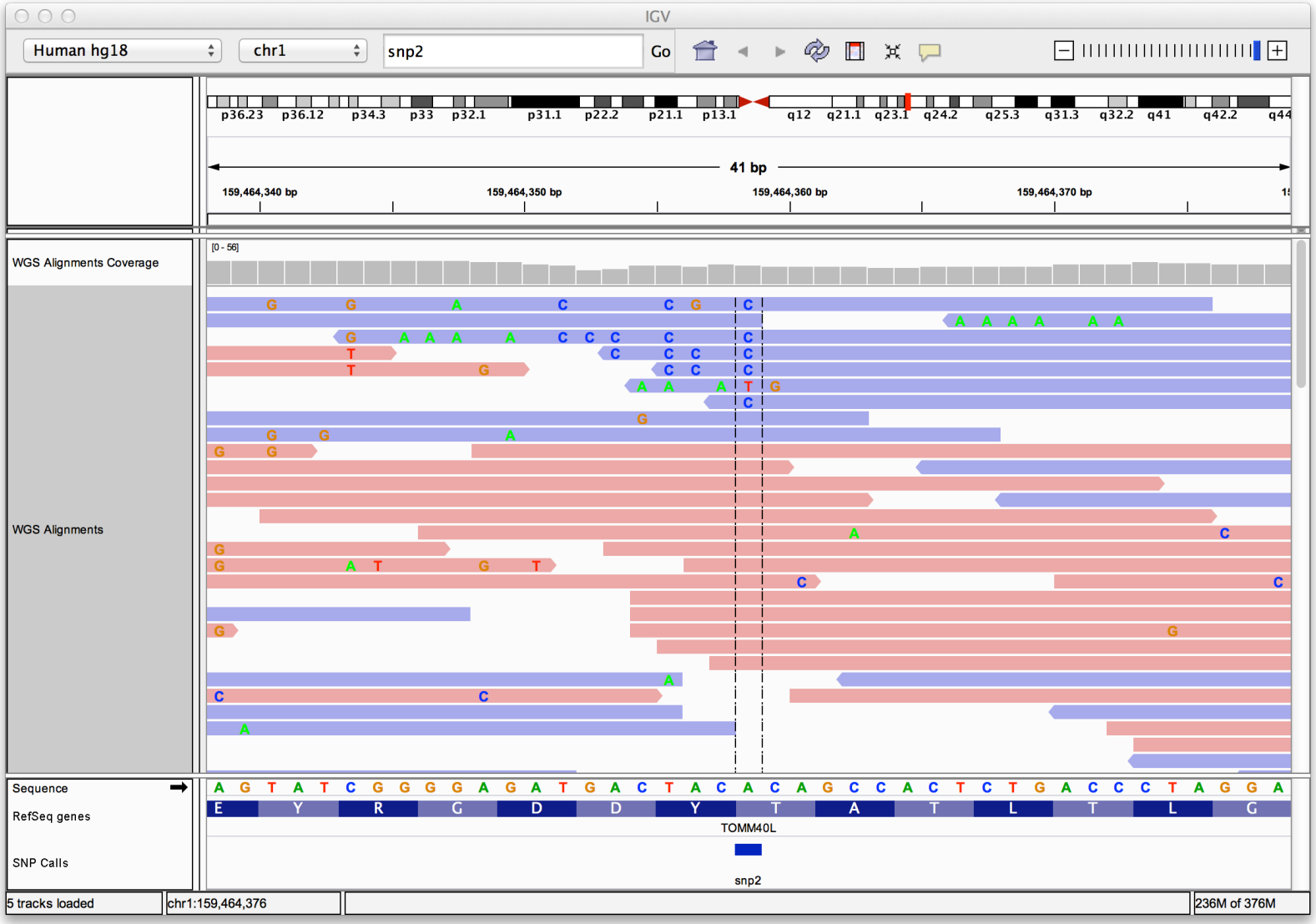


# SNVs and Structural variations

- 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



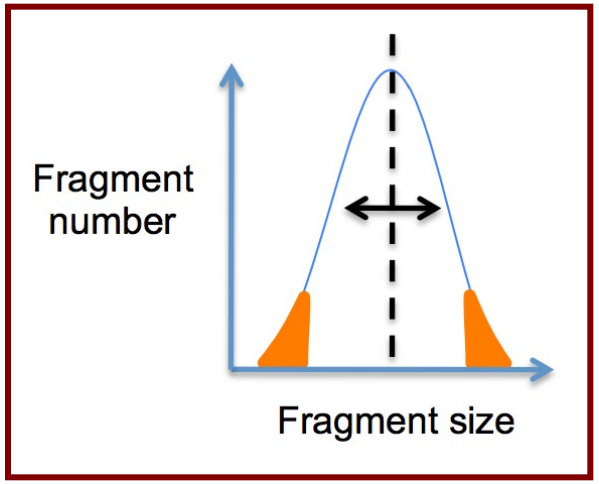
# Viewing Structural Events

- 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)

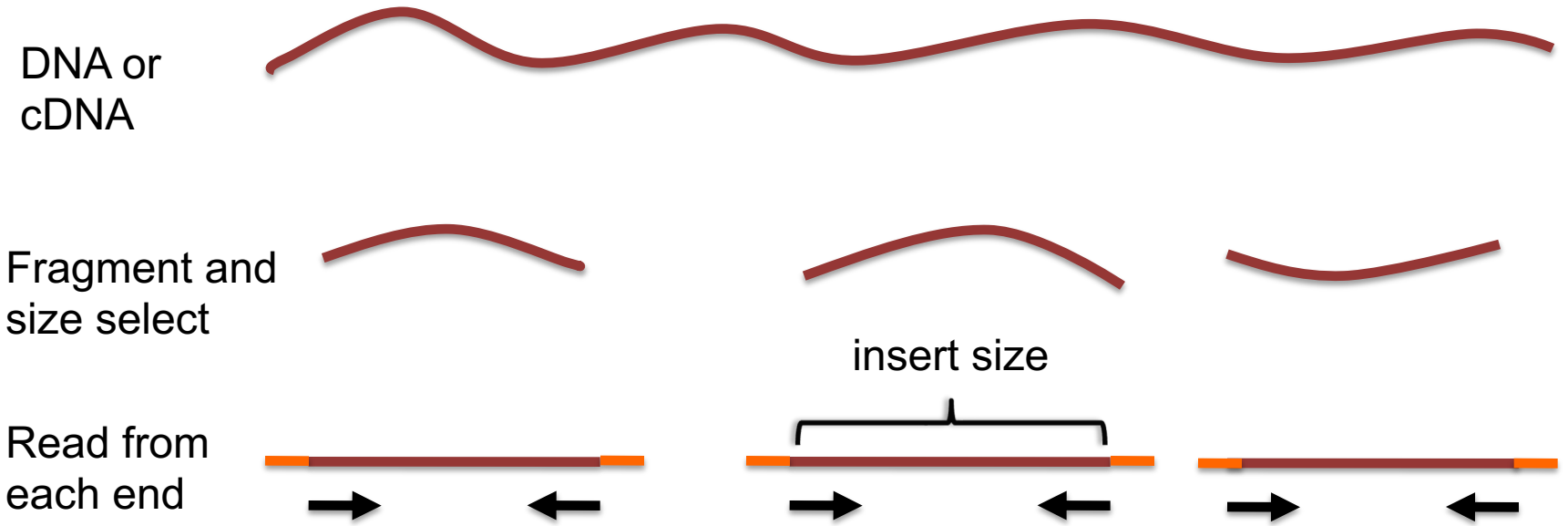
# Paired-end sequencing

DNA or cDNA

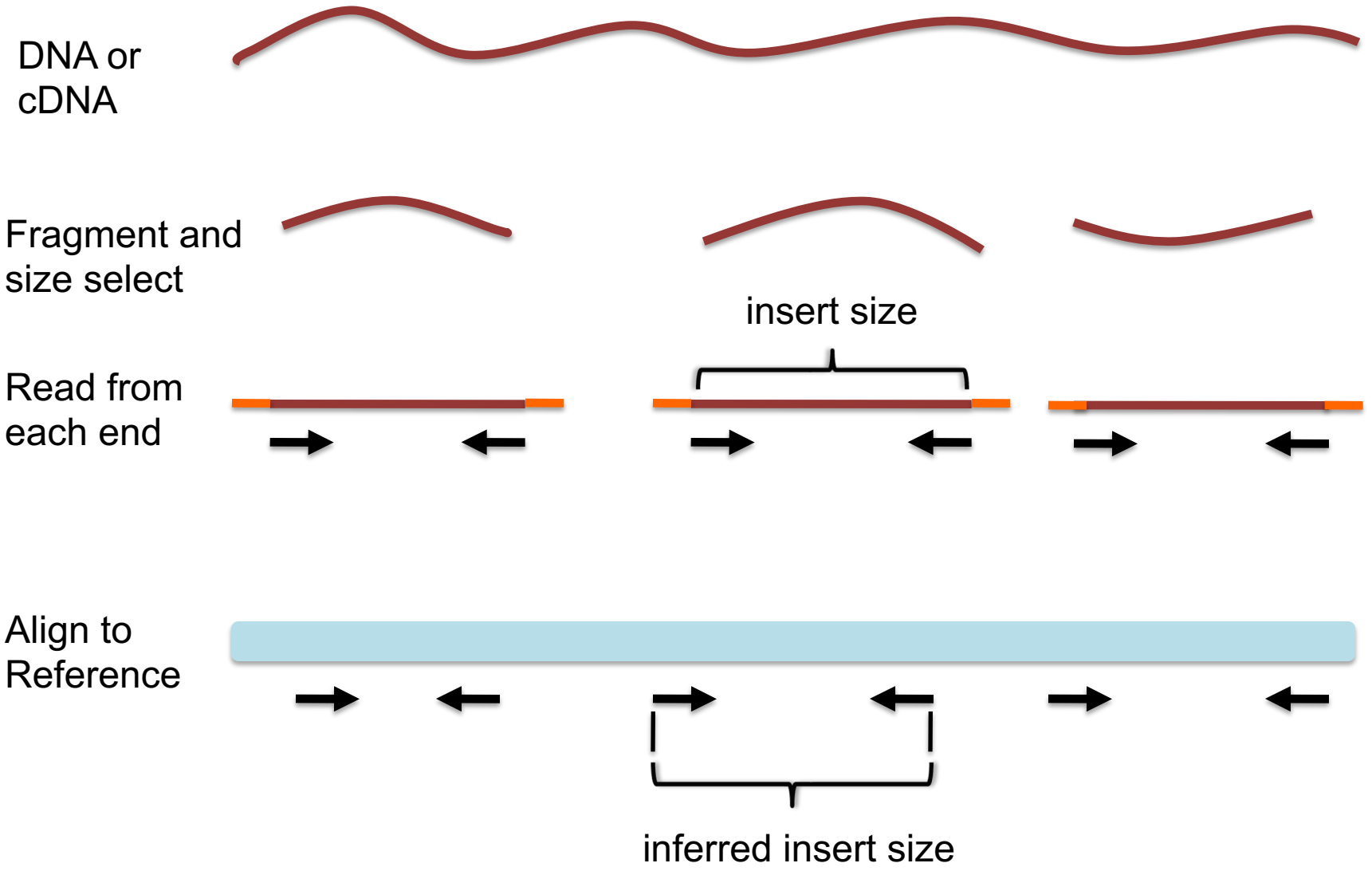
Fragment and size select



# Paired-end sequencing



# Paired-end sequencing



# Interpreting inferred insert size

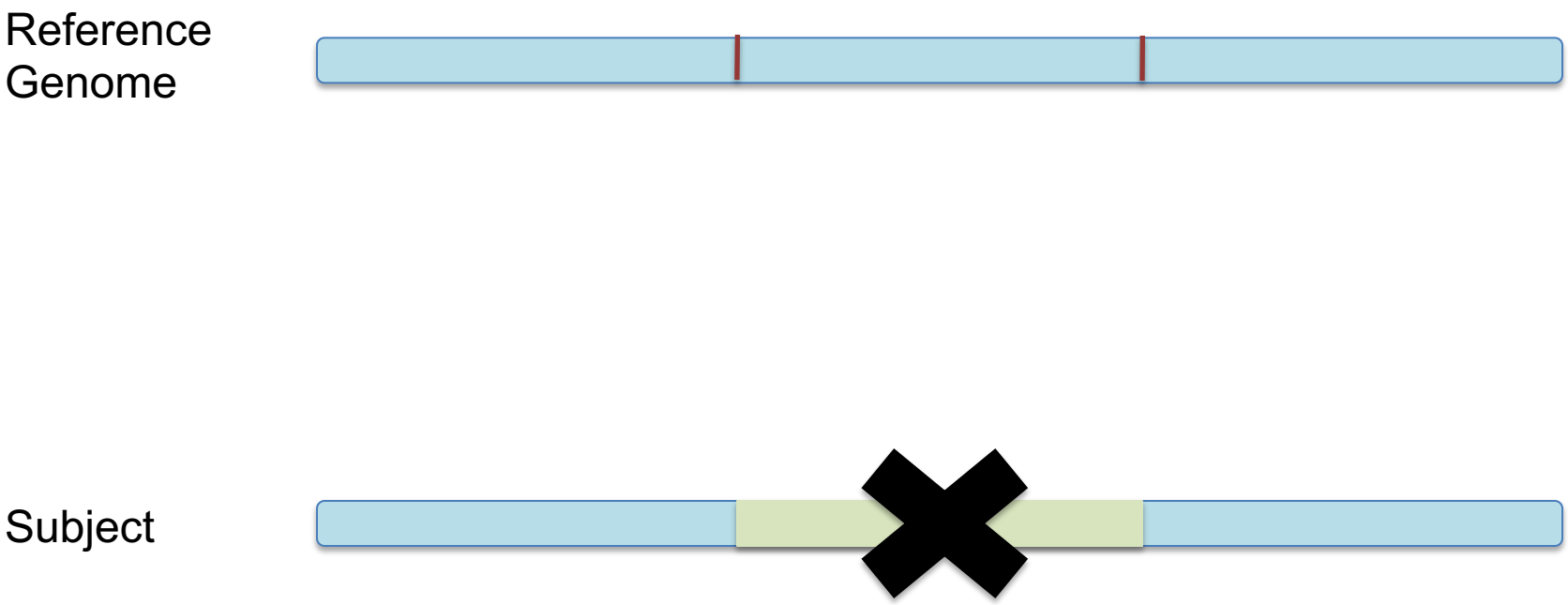
The “inferred insert size” can be used to detect structural variants including

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

# Deletion

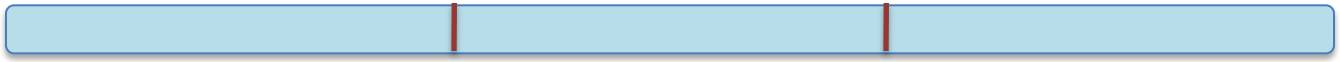
What is the effect of a deletion on inferred insert size?

# Deletion



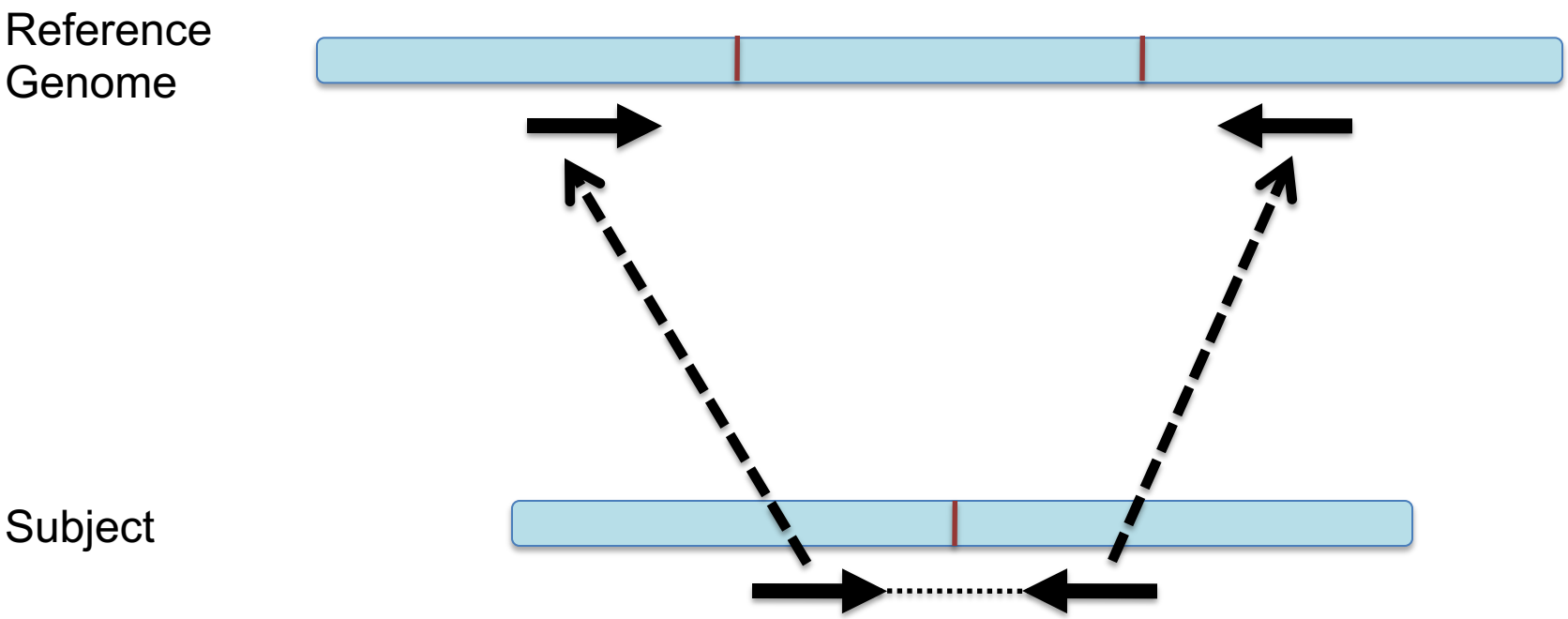
# Deletion

Reference Genome



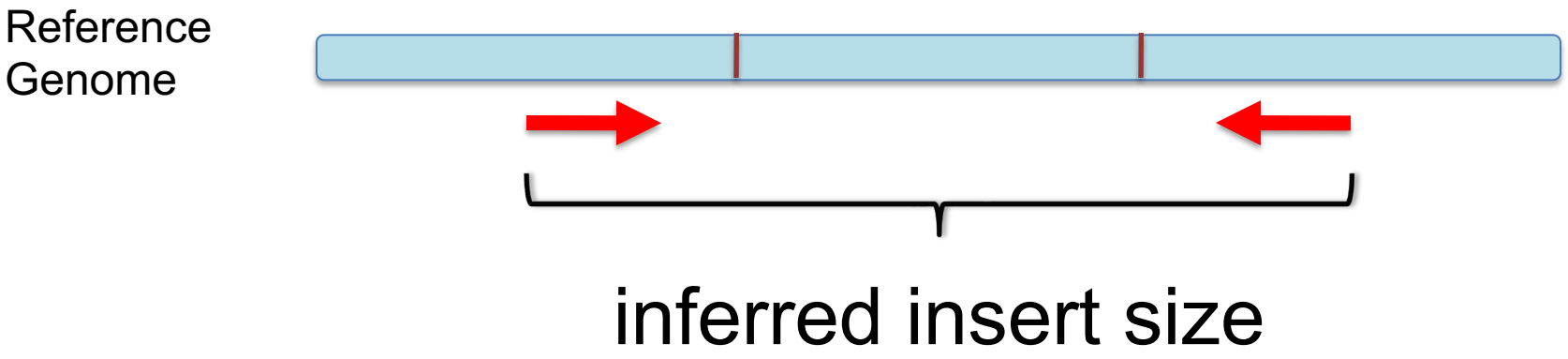
Subject

# Deletion



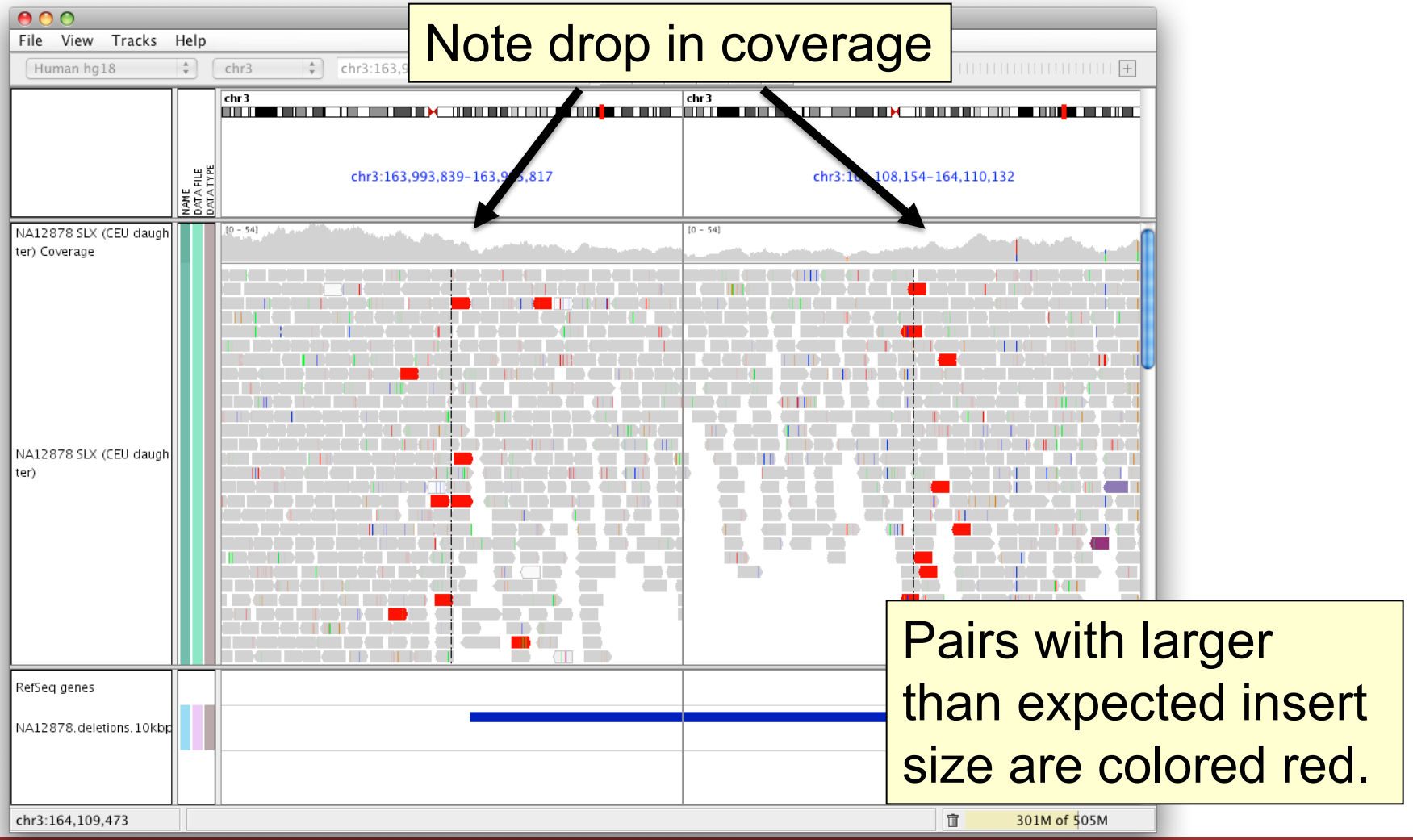
# Deletion

Inferred insert size is > expected value



# Color by insert size

# Deletion

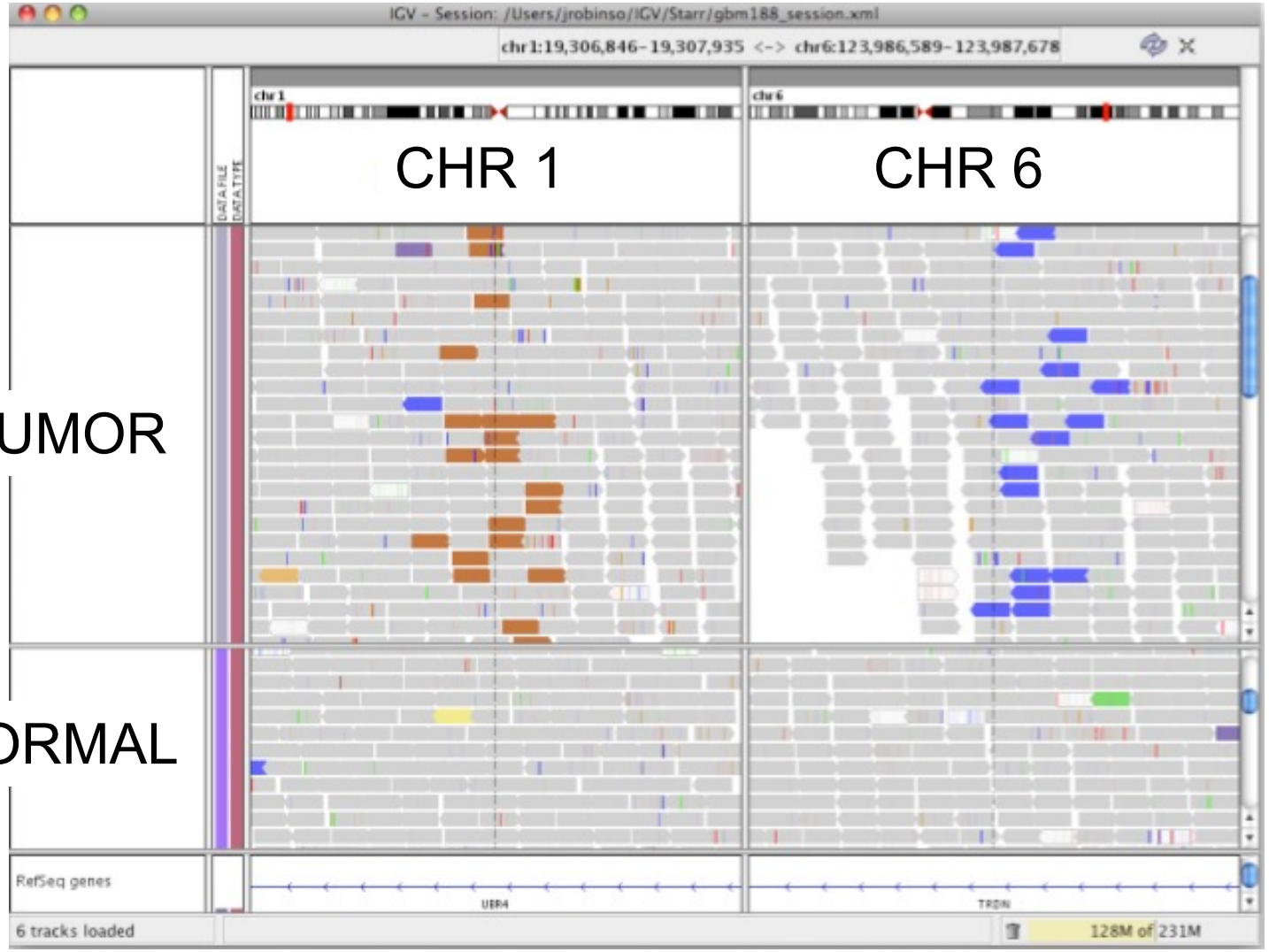

Note drop in coverage
File View Tracks Help
Human hg18
chr3
chr3:163,9
chr3
chr3
chr3:163,993,839-163,9
chr3:1 108,154-164,110,132
NAME
DATA FILE
DATA TYPE
NA12878 SLX (CEU daugh ter) Coverage
[0 - 54]
[0 - 54]
NA12878 SLX (CEU daugh ter)
RefSeq genes
NA12878.deletions.10kbp
chr3:164,109,473
301M of 505M
Pairs with larger than expected insert size are colored red.

# Insert size color scheme

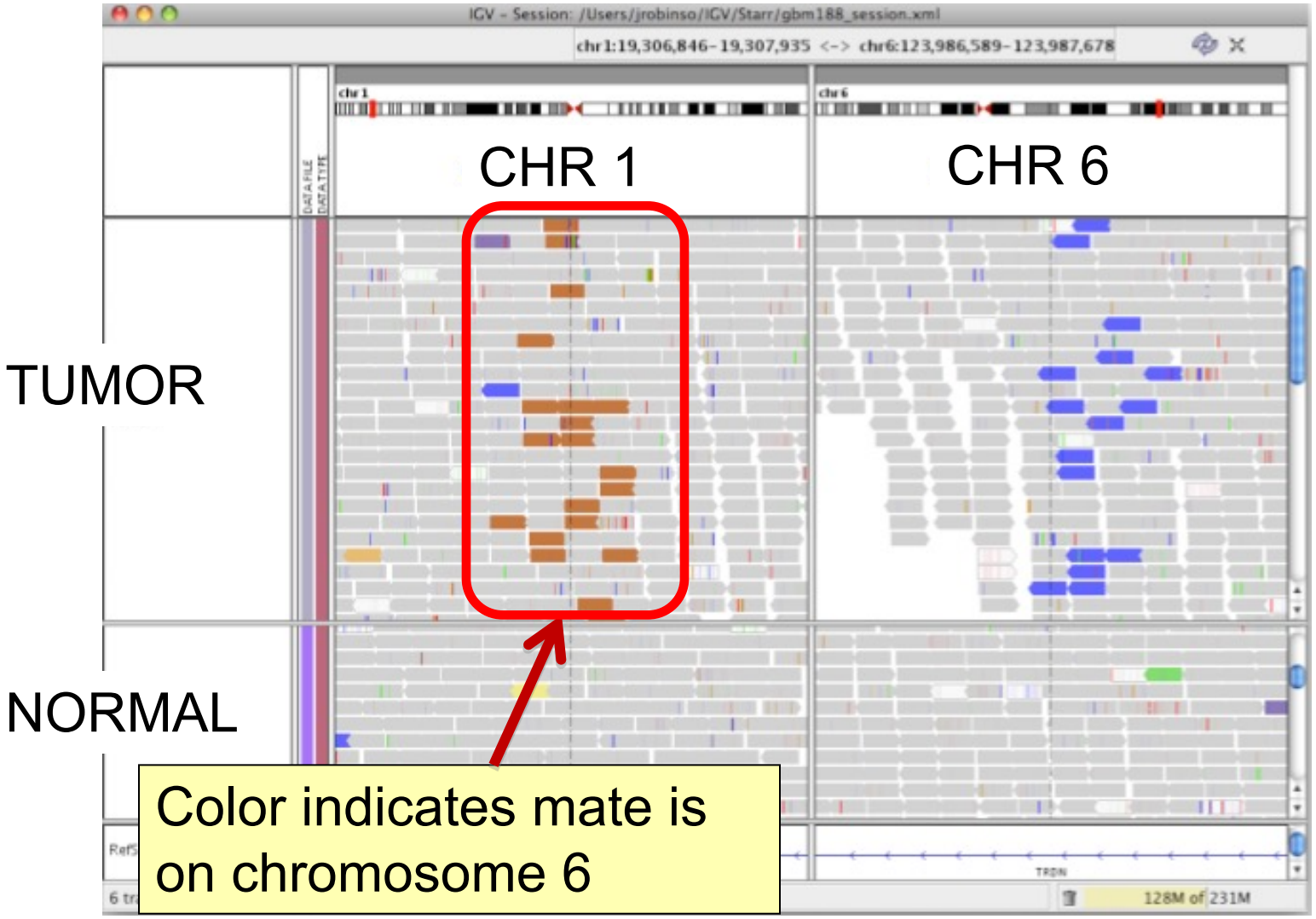
- Smaller than expected insert size: 
- Larger than expected insert size: 
- Pairs on different chromosomes

*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

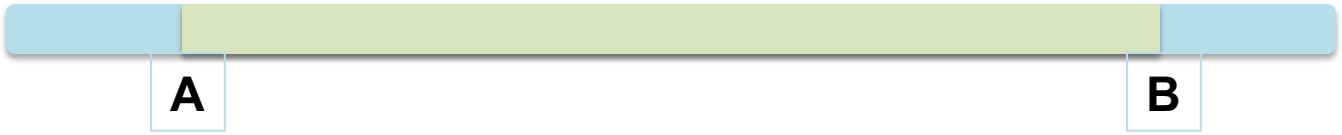
- read strand, left *vs* right, *and*
- read order, first *vs* second

# Inversion

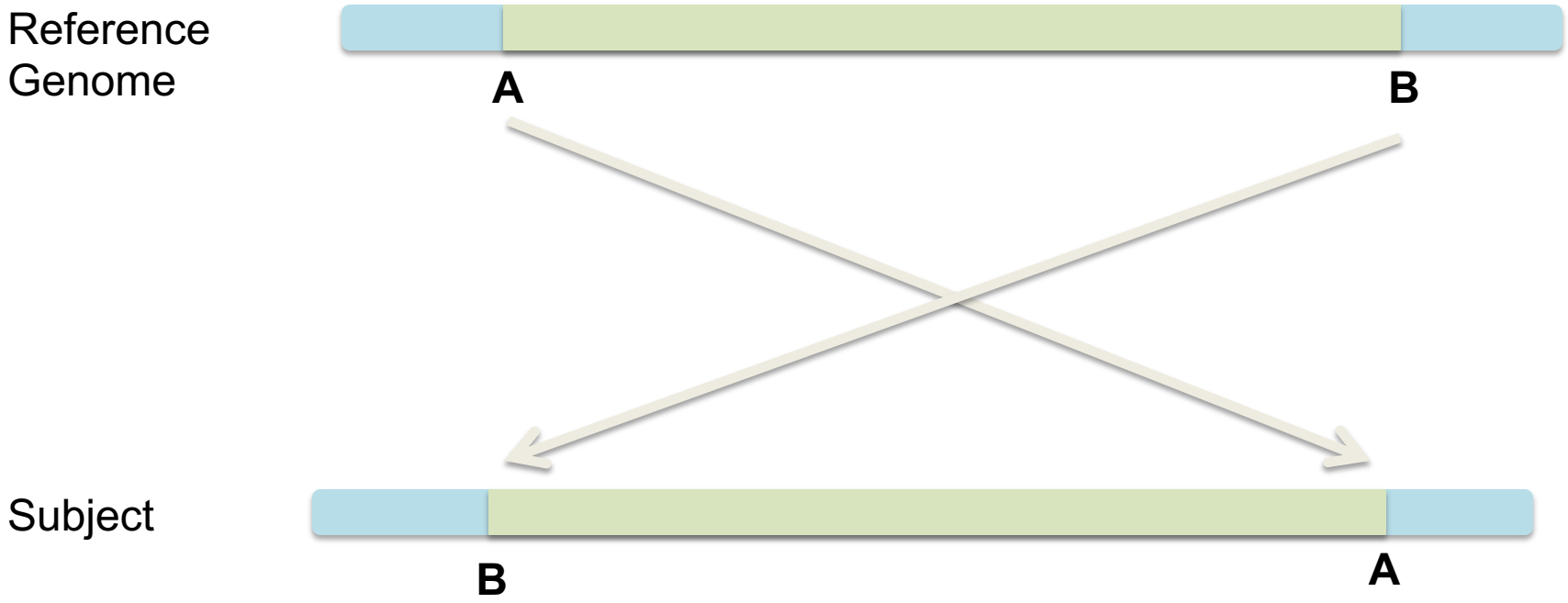
Reference genome

# Inversion

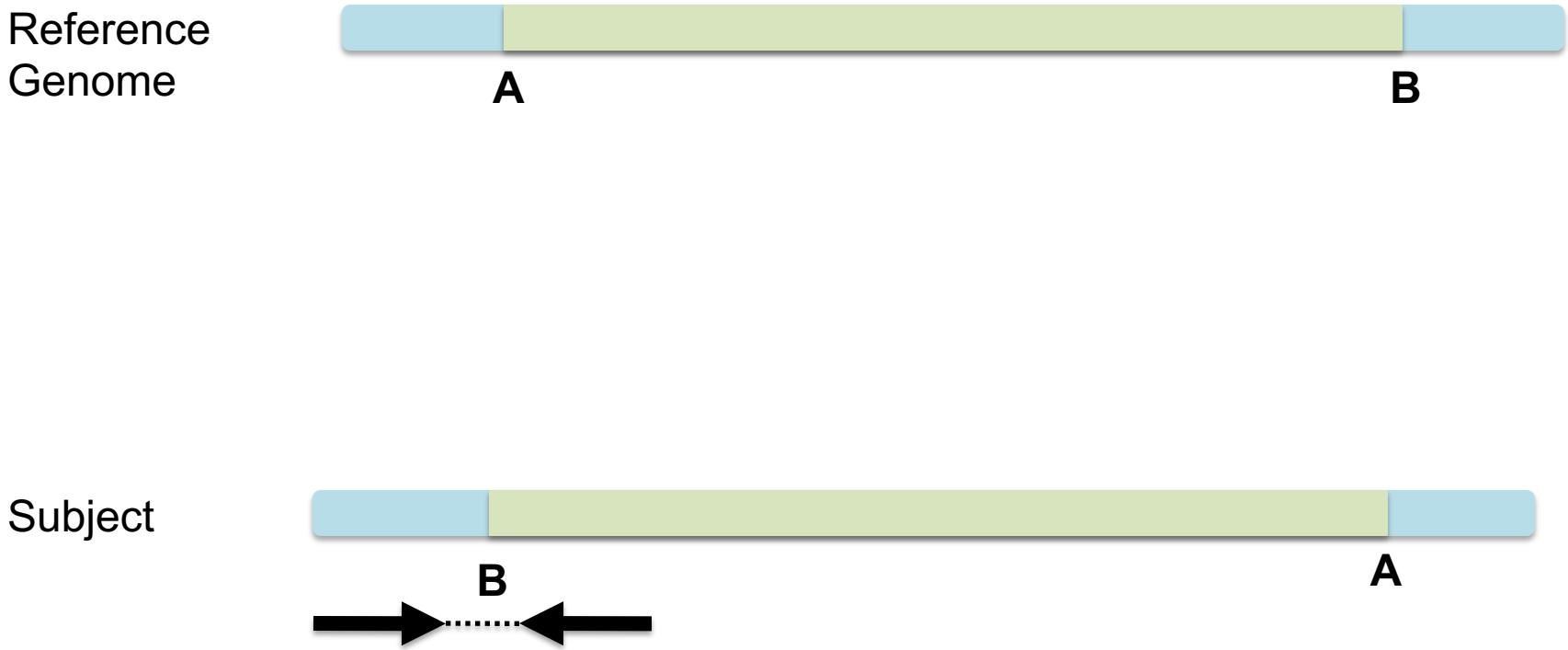
Reference genome



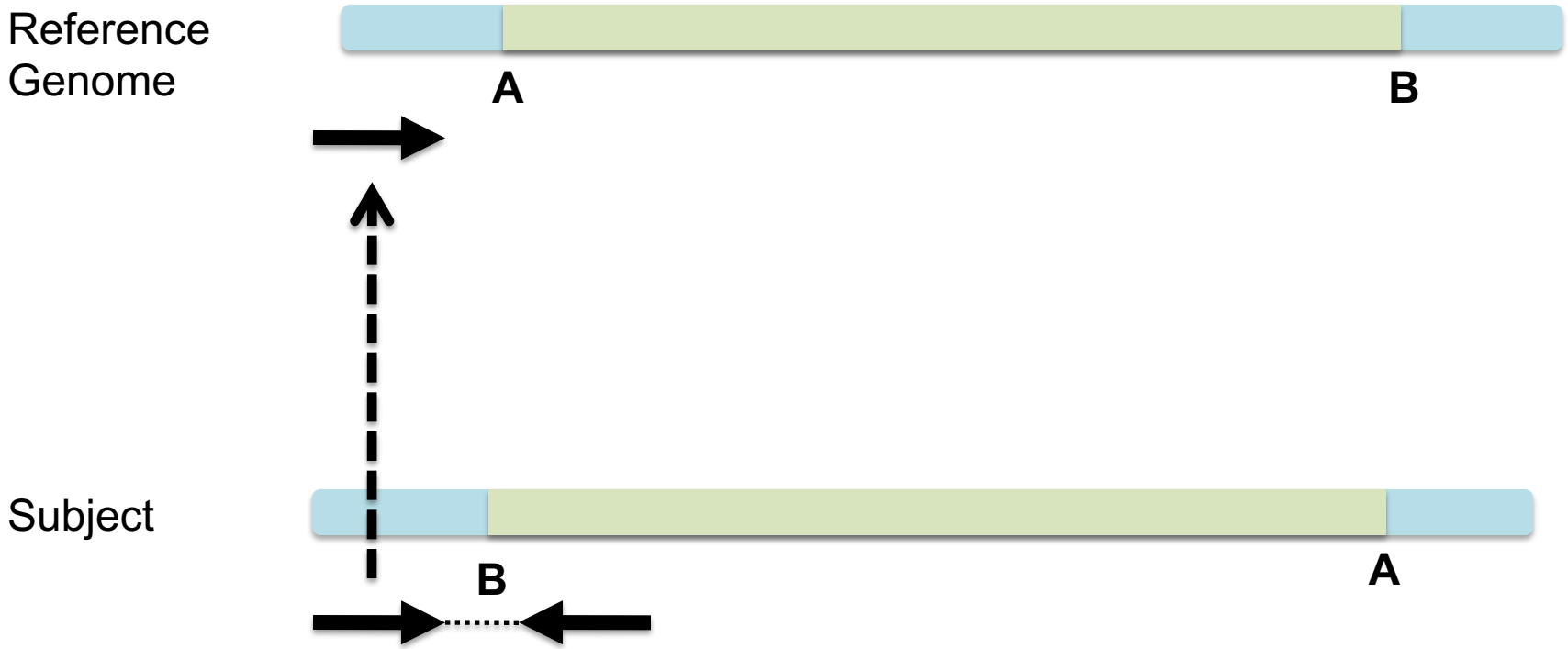
# Inversion



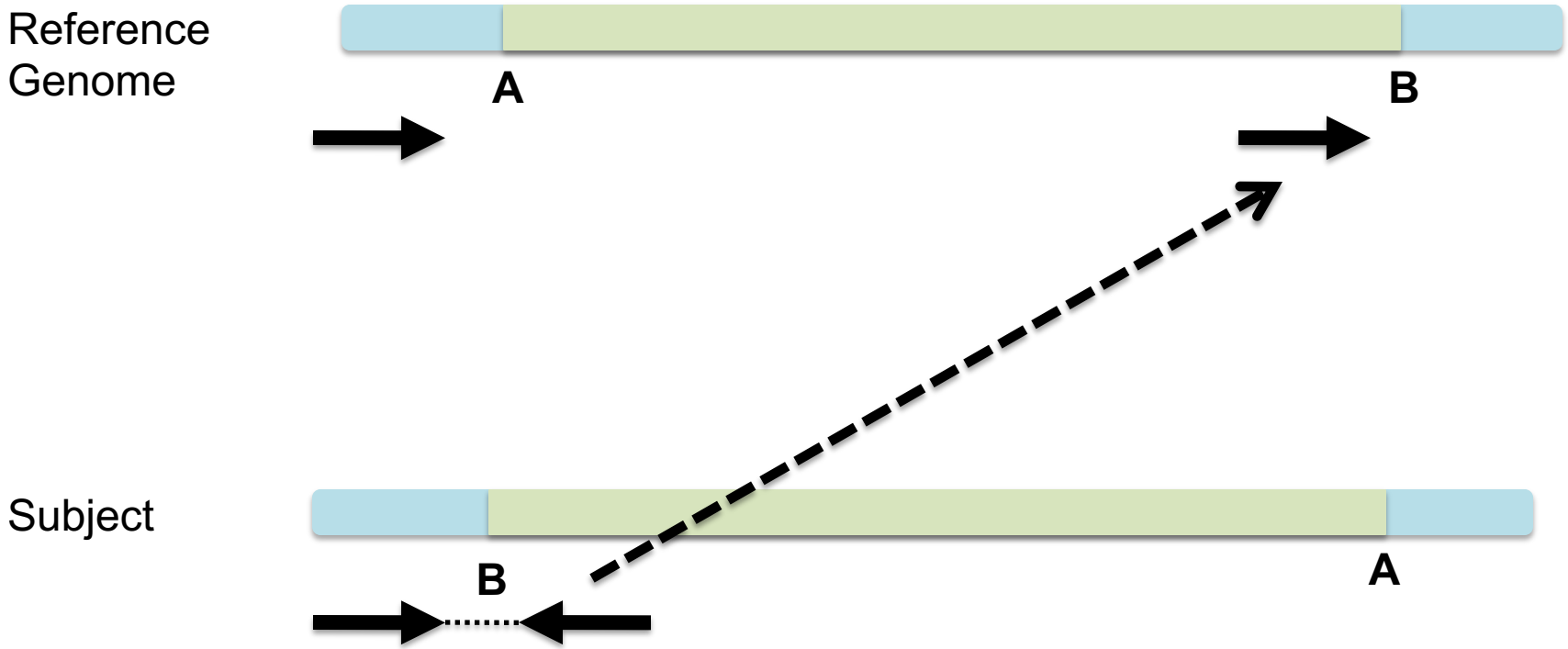
# Inversion



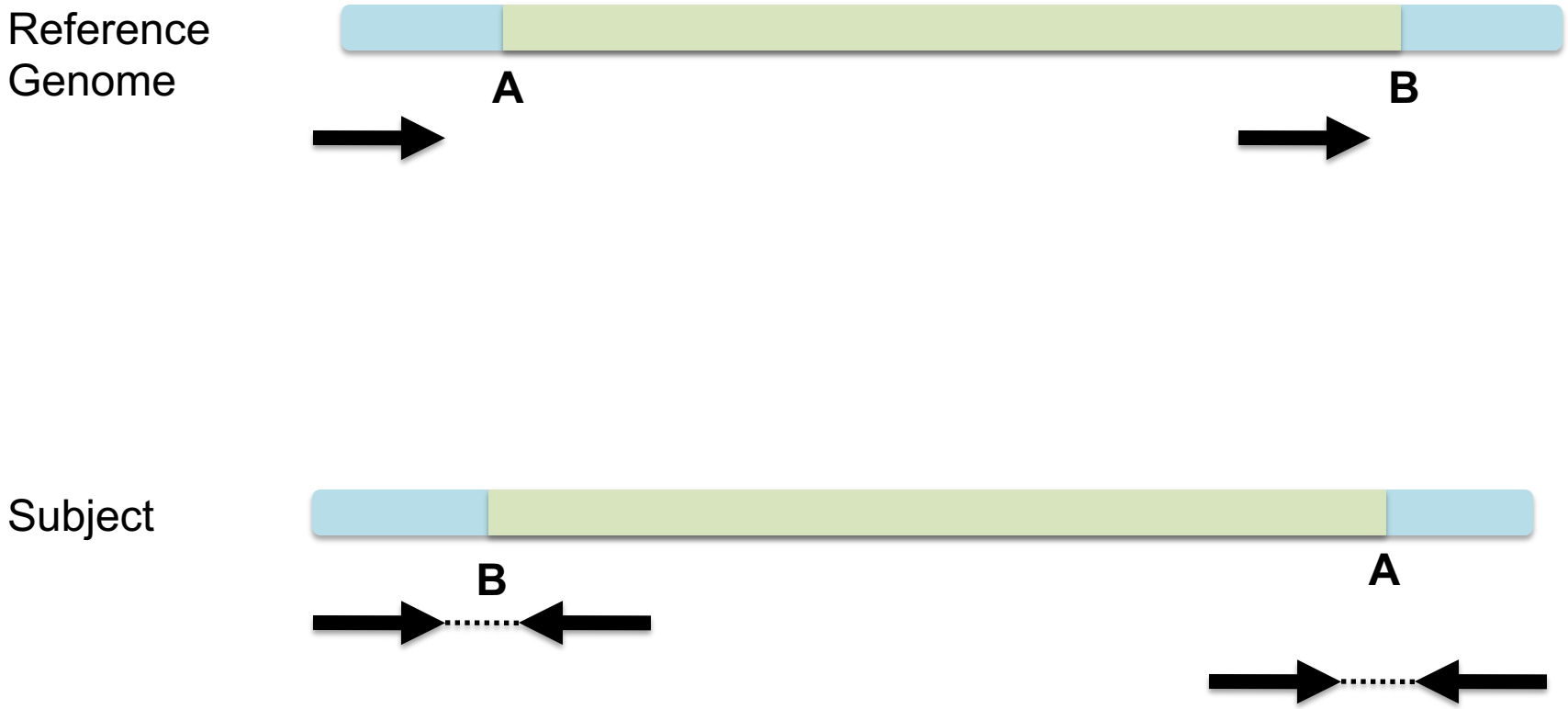
# Inversion



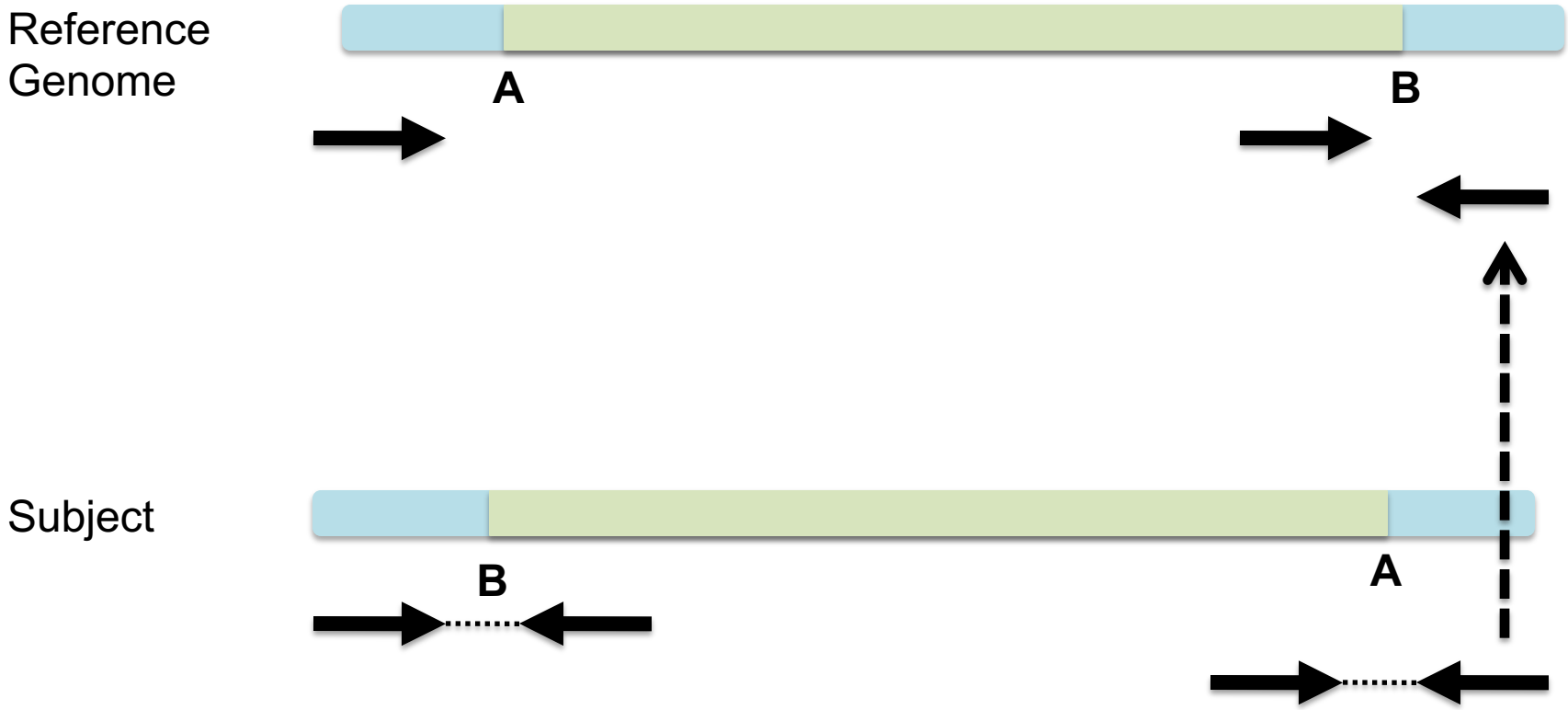
# Inversion



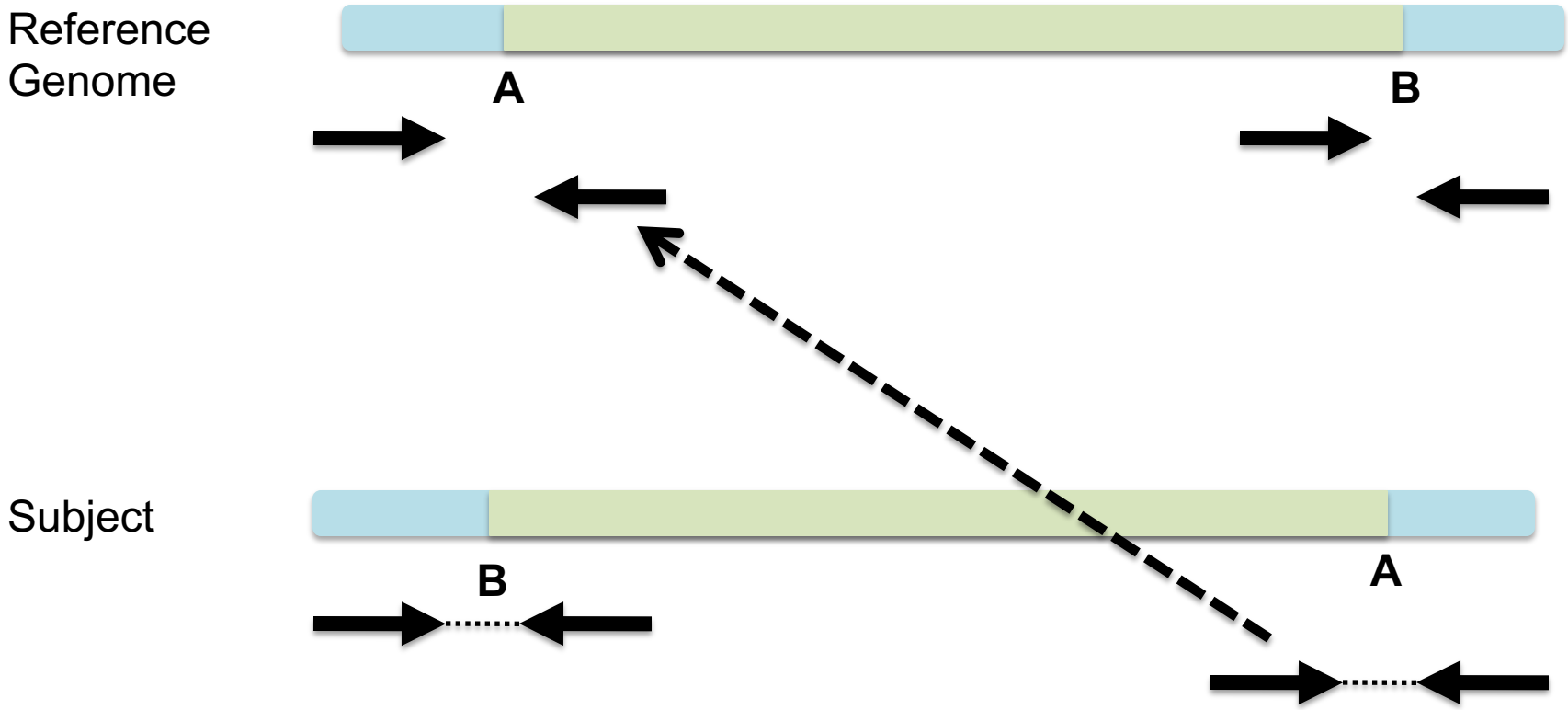
# Inversion



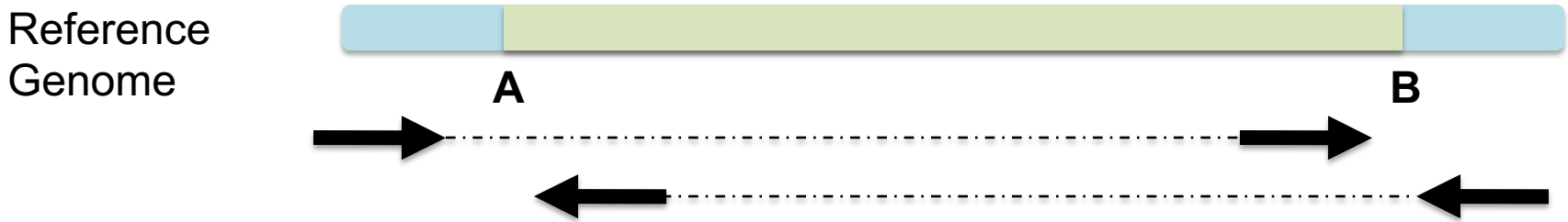
# Inversion



# Inversion



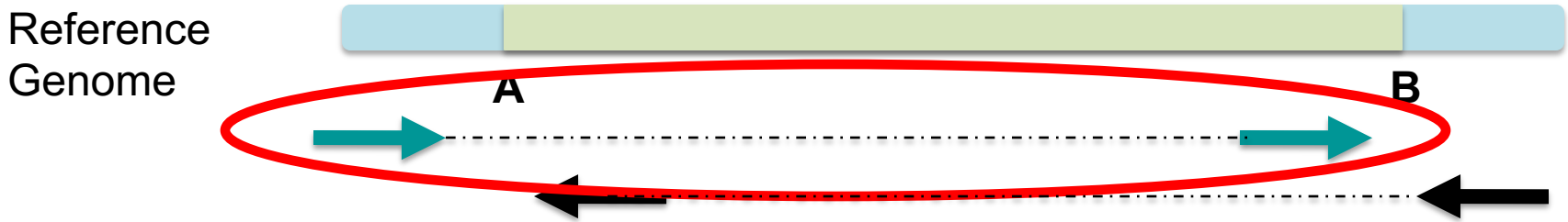
# Inversion



# Inversion

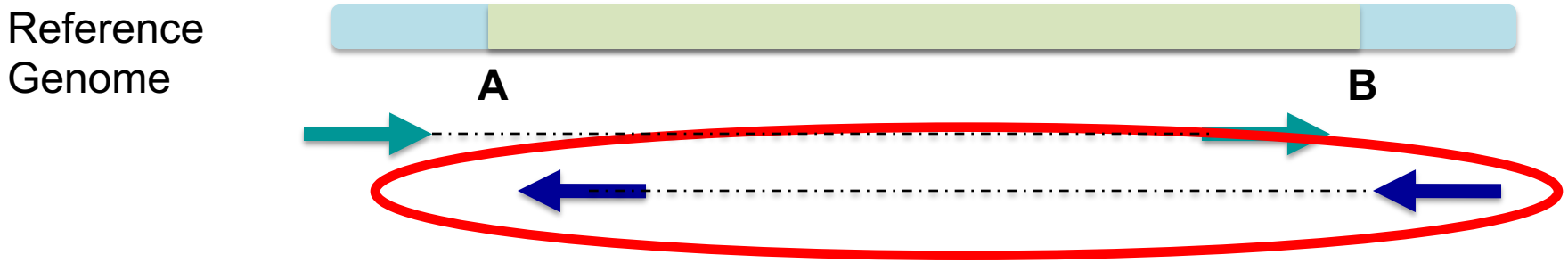

Anomaly: expected orientation of pair is inward facing ( → ← )

# Inversion



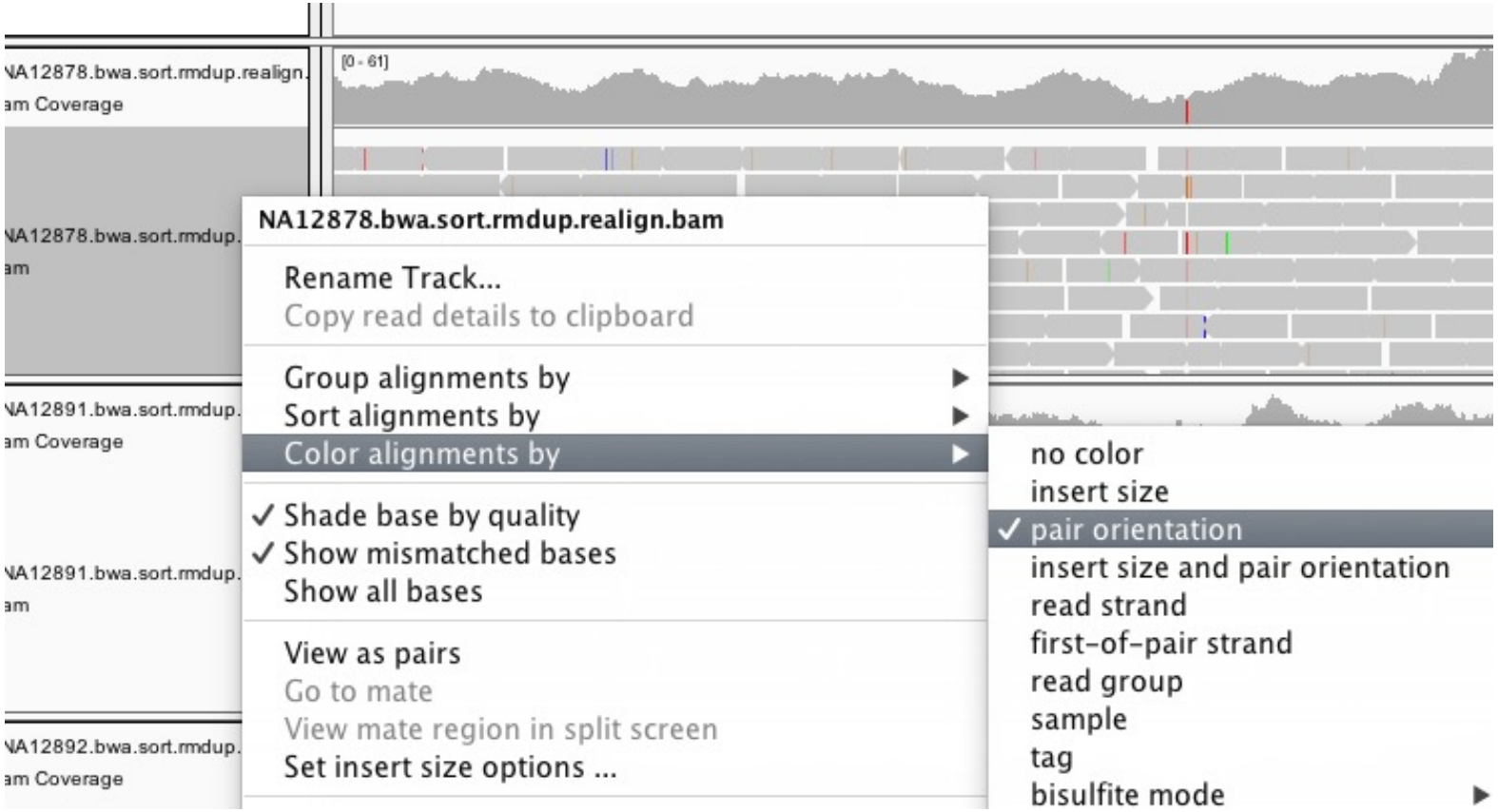

“Left” side pair

# Inversion

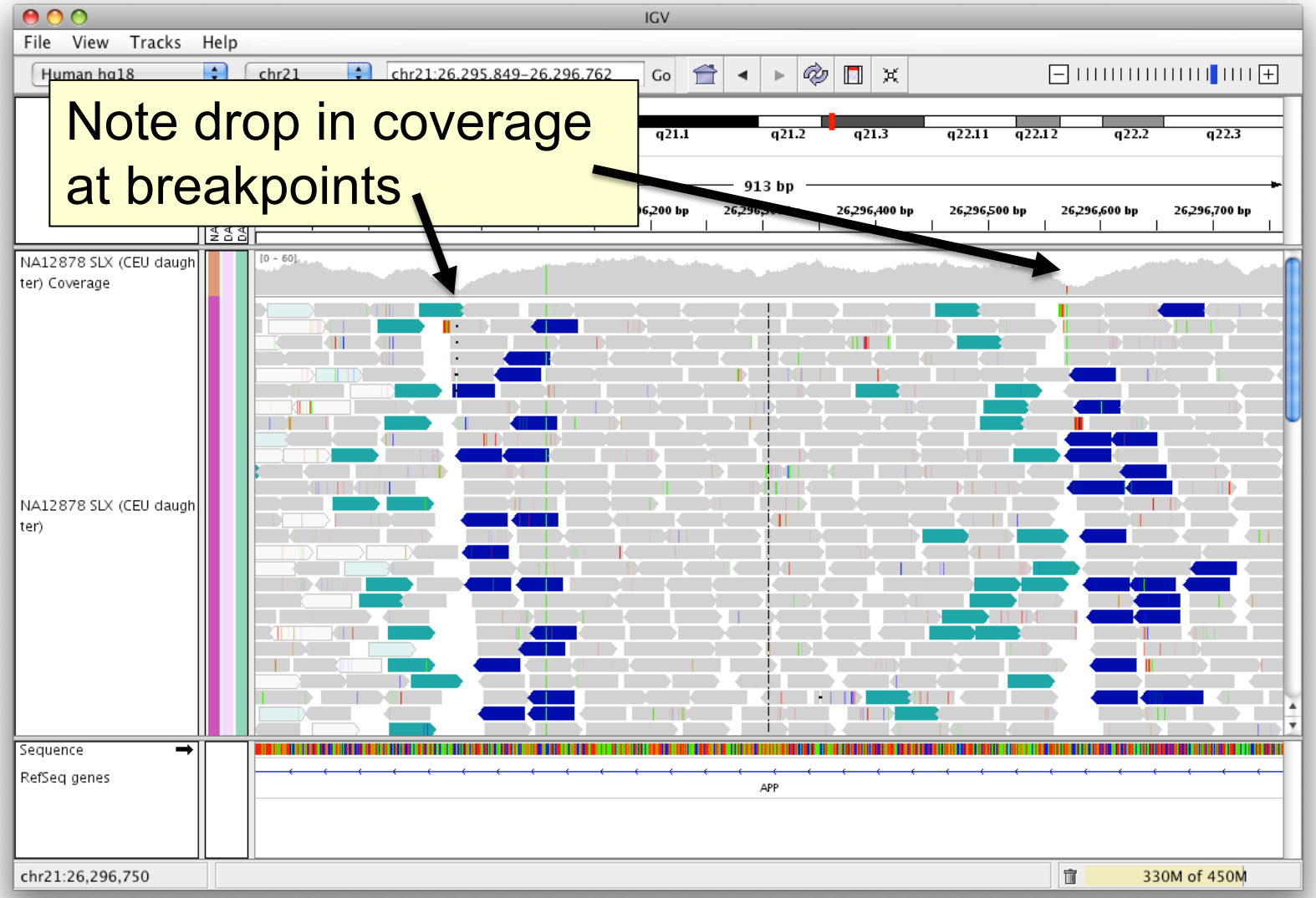



"Right" side pair

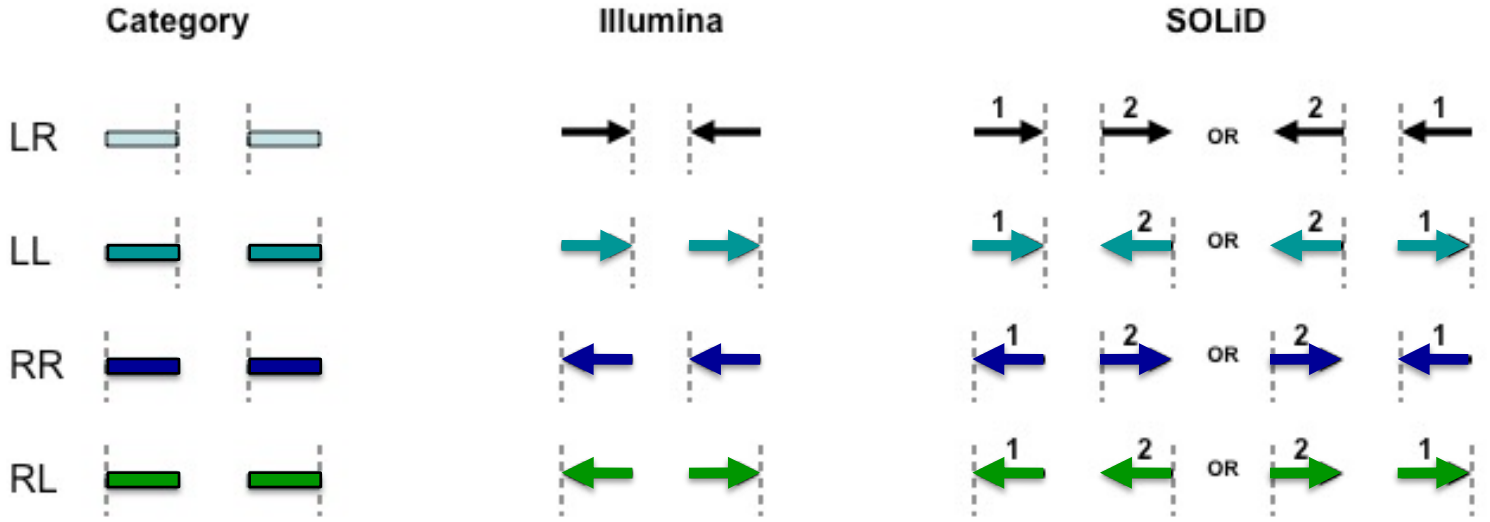
# Color by pair orientation



# Inversion


IGV
File View Tracks Help
Human hg18
chr21
chr21:26,295,849-26,296,762
Go
Note drop in coverage at breakpoints
913 bp
NA12878 SLX (CEU daughter) Coverage
NA12878 SLX (CEU daughter)
Sequence
RefSeq genes
APP
chr21:26,296,750
330M of 450M

## Interpretation of read pair orientations

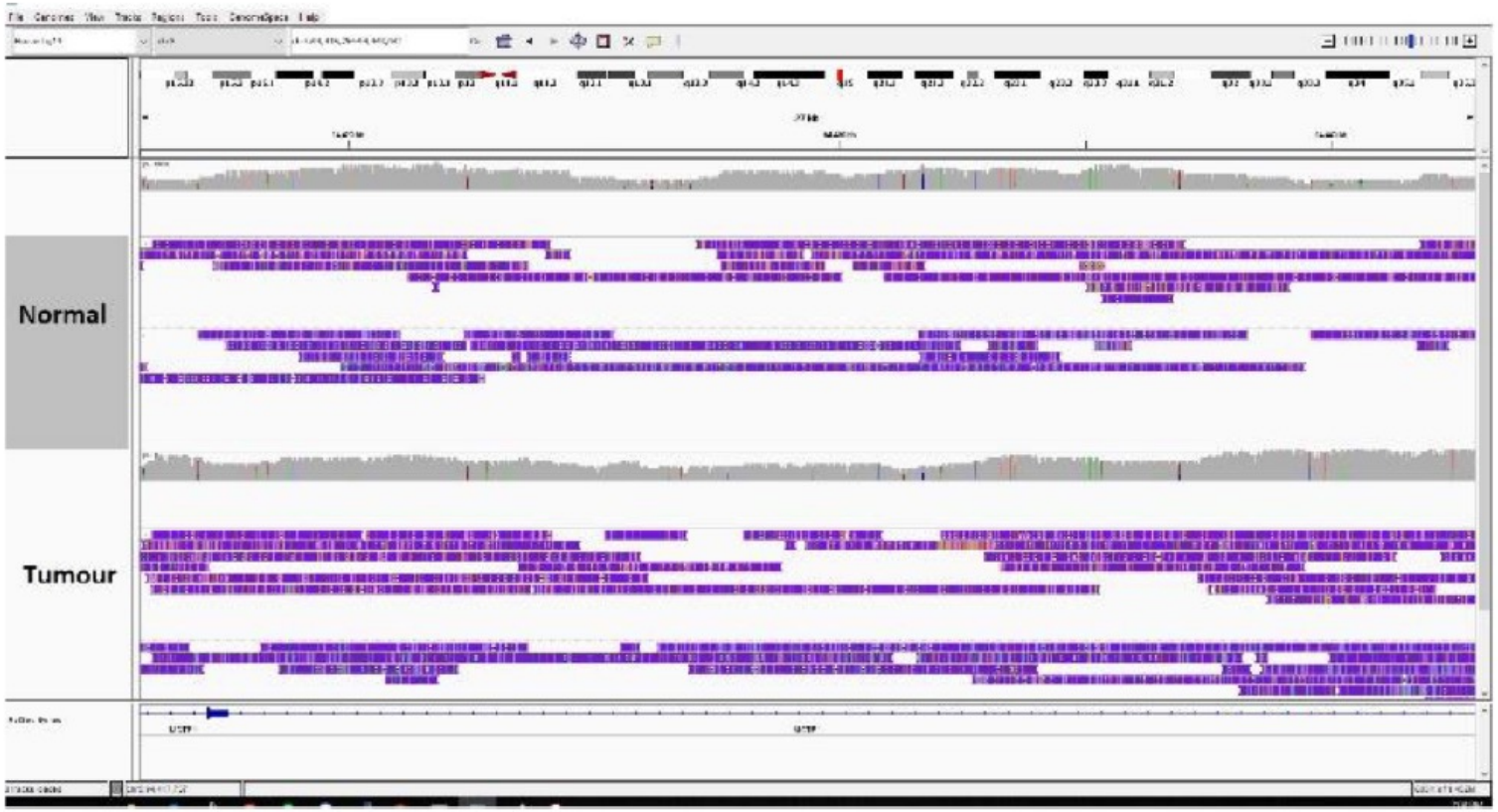



| | |
|---|---|
| LR | Normal reads.<br>The reads are left and right (respectively) of the unsequenced part of the sequenced DNA fragment when aligned back to the reference genome. |
| LL,RR | Implies inversion in sequenced DNA with respect to reference. |
| RL | Implies duplication or translocation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

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

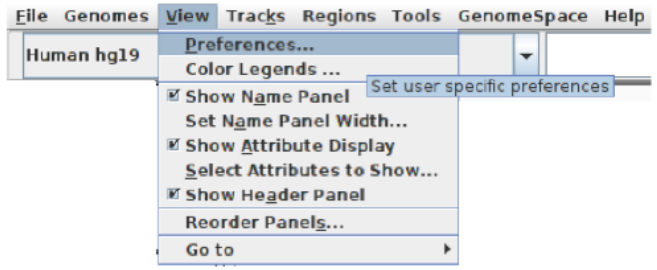
*Figure courtesy of Bob Handsaker*

# Long read considerations

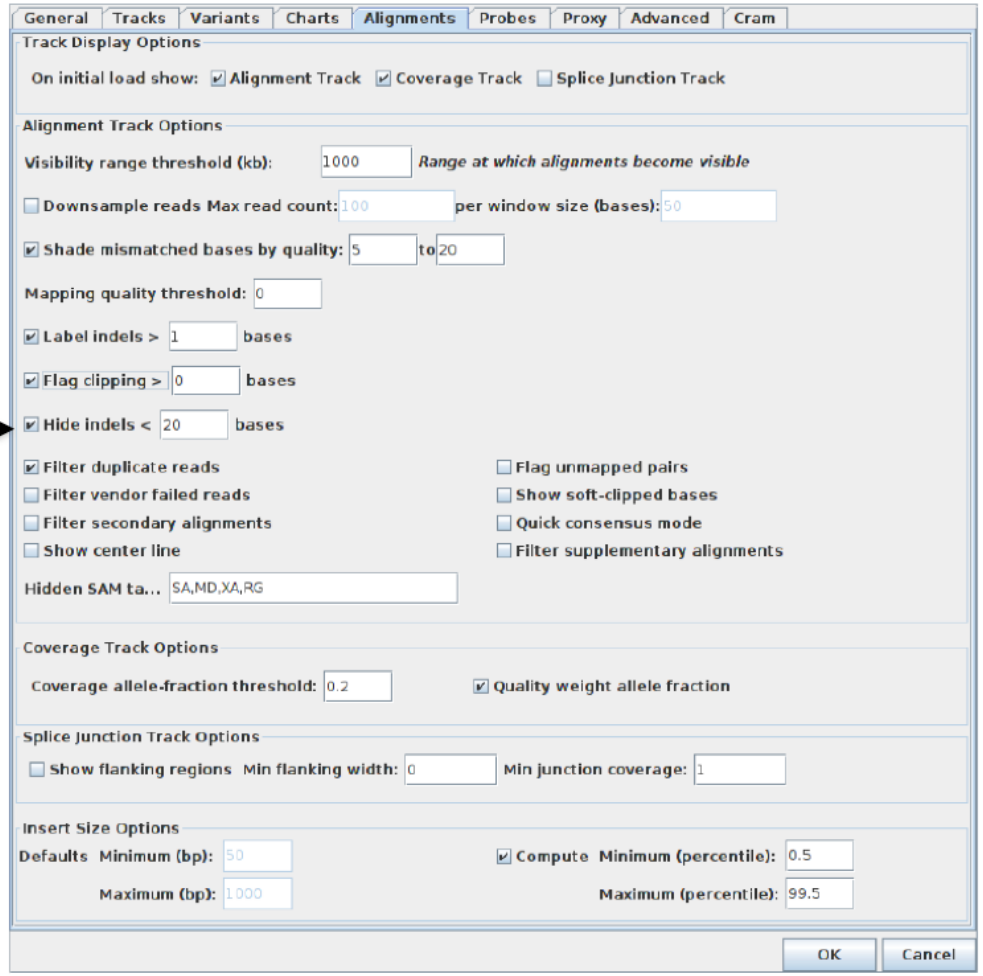



- Commonly see lots of small indels and single base errors that are simply noise
- Can be removed to be able to view the data more cleanly

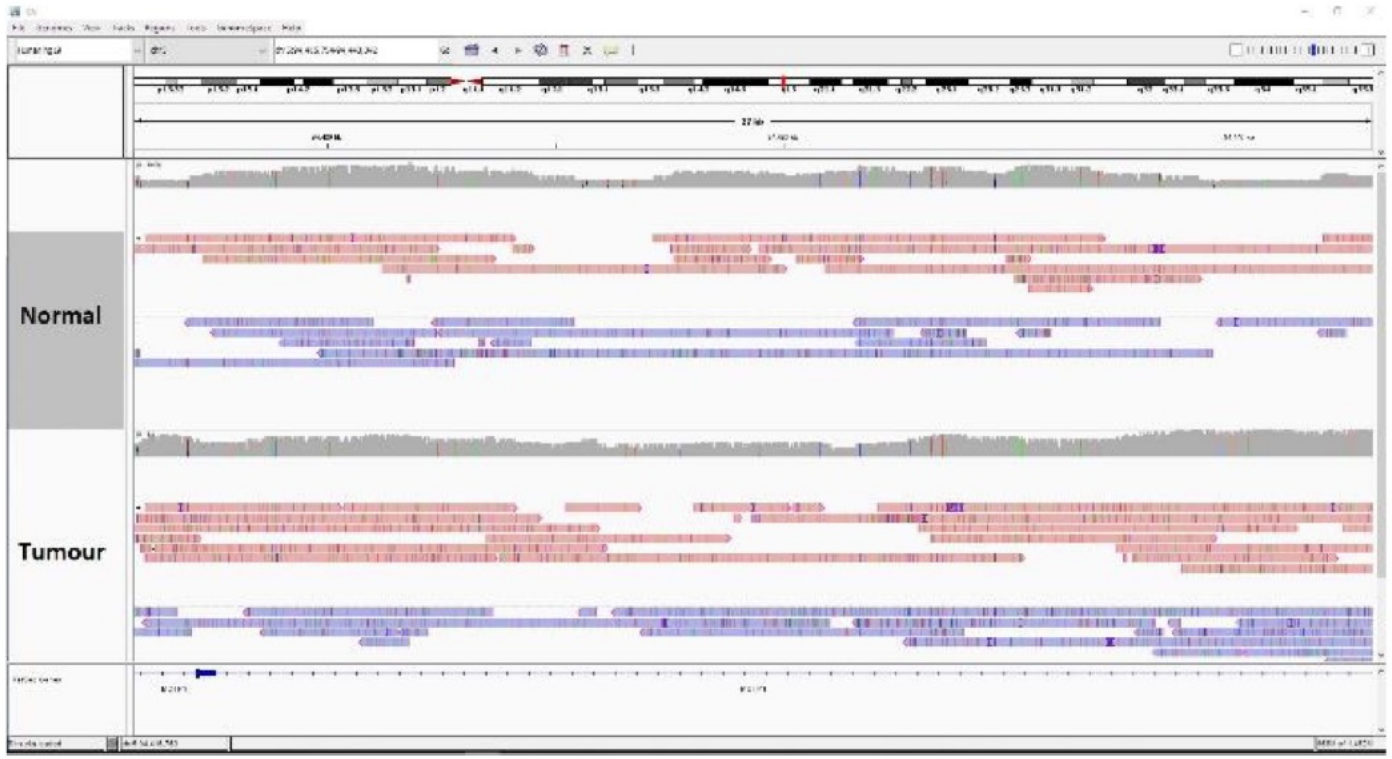
# Long read considerations




Setting an indel threshold hides noise from small indels

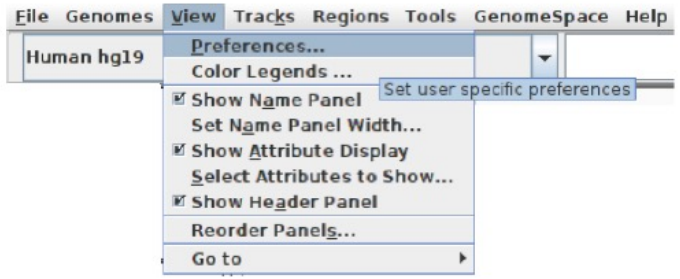


# Long read considerations

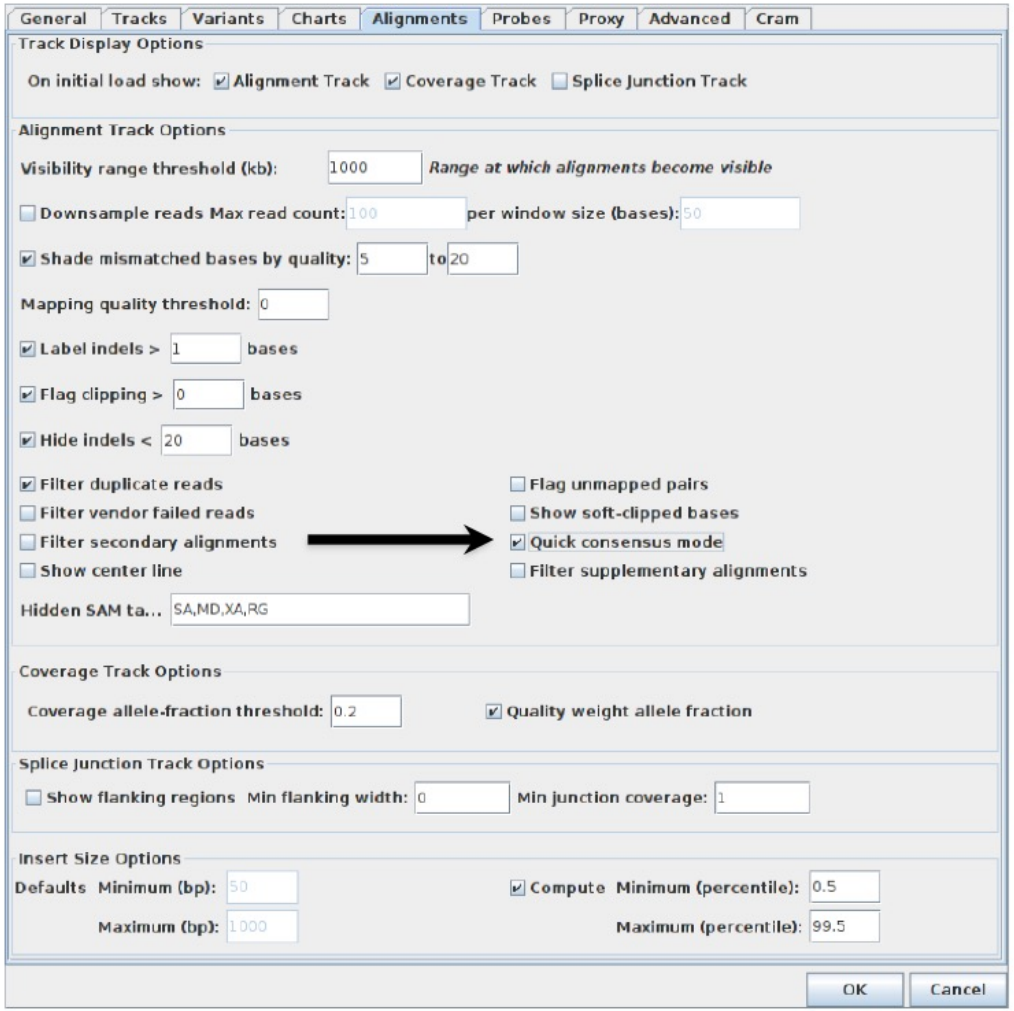



- Reads are not all purple dashes
- Next step would be to call a consensus at each position

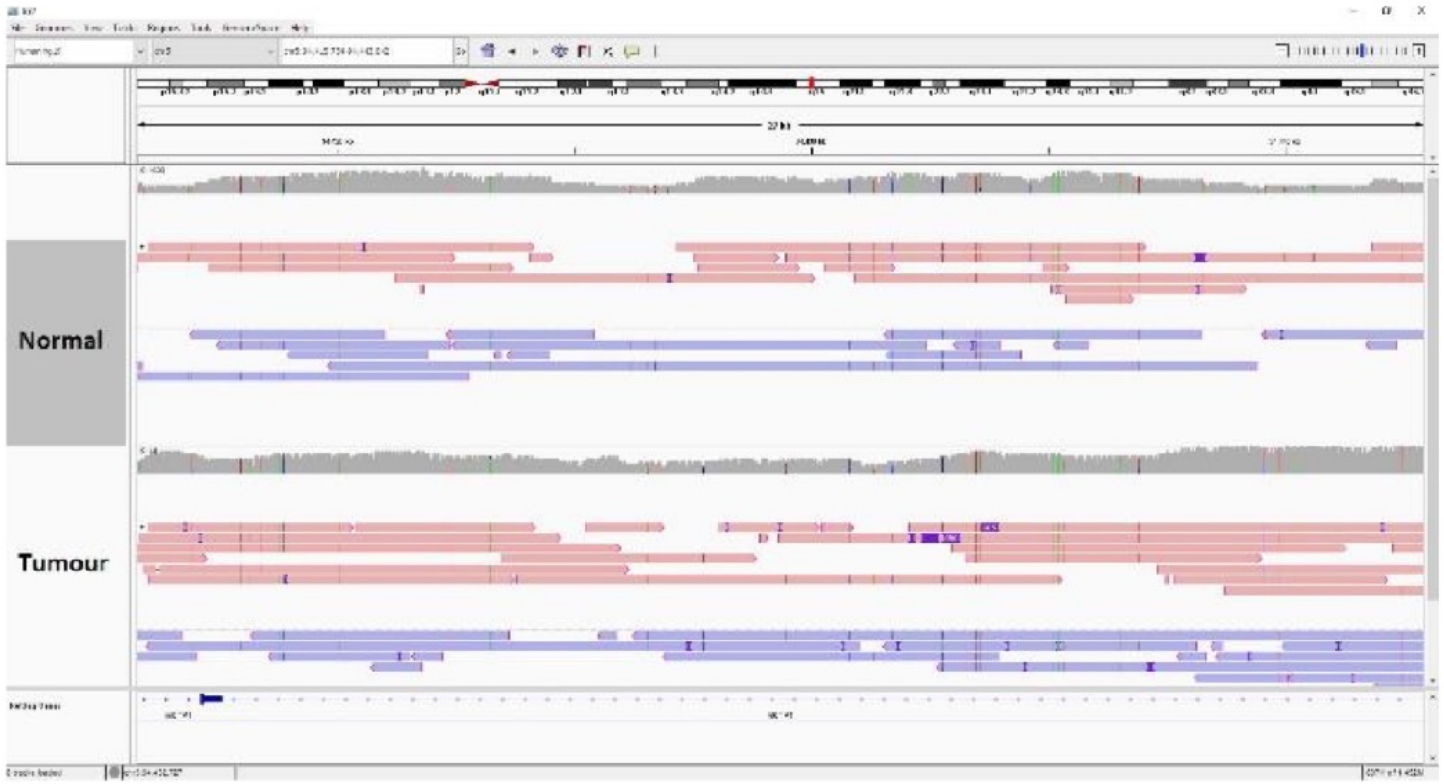
# Long read considerations




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



ARTICLE | Genetics inMedicine

Open

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

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

**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.

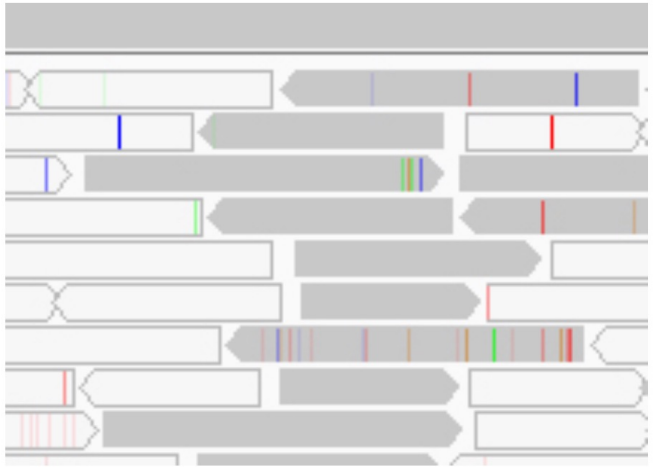
**Results:** After reading the SOP, average accuracy in somatic variant identification increased by 16.7% ($p$ value = 0.0298) and average interreviewer agreement increased by 12.7% ($p$ value < 0.001). Manual review conducted after reading the SOP did not significantly increase reviewer time.

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

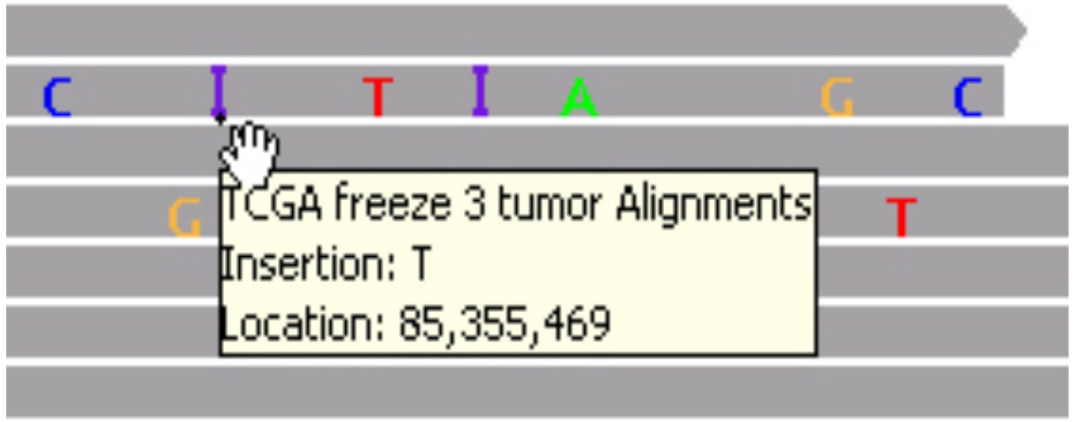
*Genetics in Medicine* (2018) https://doi.org/10.1038/s41436-018-0278-z

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

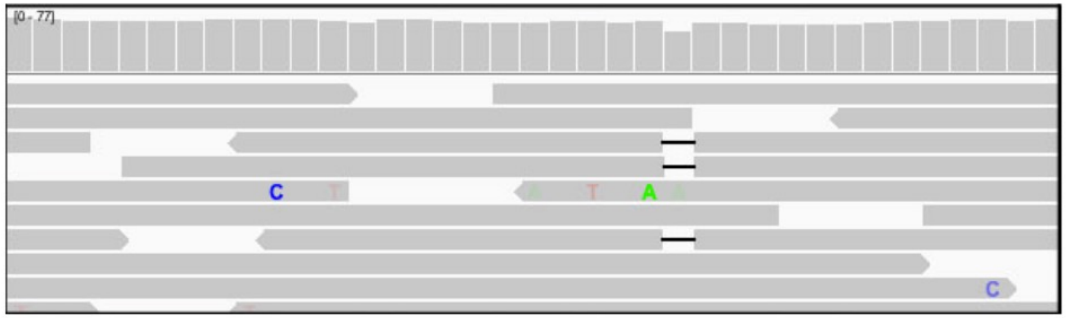
# Other notes



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

Workshop Sponsors:


Canadian Centre for Computational Genomics